Supplementary Information for:

Complementary symbiont contributions to plant decomposition in a fungus-farming termite

Michael Poulsen, Haofu Hu, Cai Li, Zhensheng Che u, Saria Otani, Sanne Nygaard, Tania Nobre, Sylvia Klaubauf, Philipp M. Schindler, Frank Hauser, Hailin Pan, Zhikai Yang, Anton Sonnenberg, Z. Wilhelm de Beer, Yong Zhang, Michael J. Wingfield, Cornelis J. P. Grimmelikhuijzen, Ronald P. de Vries, Judith Korb, Duur K. Aanen, Jun Wang, Jacobus J. Boomsma and Guojie Zhang

Table of contents

1 Collection and preparation of material

1.1 Collections

A *Macrotermes natalensis* colony (Mn117) was collected from a field population in South Africa in Mookgophong (S24°40’30.5” E28°47’50.4”, elevation 1045 m) in 2011. Workers and soldiers were collected from the exposed fungus comb using sterile forceps and were placed directly in RNAlater® on ice. The royal chamber was excavated and the queen and king were stored in RNAlater®. All samples were kept frozen until DNA extraction.

1.2 *Termitomyces* isolation and protoplasting

The homokaryotic strain of *Termitomyces* used for genome sequencing was obtained using a standard protoplasting procedure (1,2) of a heterokaryon isolated from a colony of *M. natalensis* (Mn105). Mycelium was regenerated from protoplasts and grown for two weeks on malt yeast extract agar medium (MYA). DNA was isolated from a small mycelium fragment using a CTAB extraction. To check the nuclear status of the regenerated mycelia, i.e. whether these were heterokaryotic or homokaryotic, we used standard PCR to amplify the highly variable intron of the Elongation Factor 1α gene (EF1-α), which we knew was polymorphic between the two nuclei of this strain. The PCR product was sequenced at MWG Biotech (http://www.mwg-biotech.com/). By analyzing the DNA electropherograms, we selected the homokaryon to be genome sequenced (labeled P5) from the collection of these EF-1α sequenced strains.

2 DNA and RNA extraction

2.1 Termite DNA and RNA extraction

Four parallel extractions of termite DNA were performed from the single *M. natalensis* queen of colony Mn117. Approximately 300 mg of queen ovary tissue per tube was transferred to a 15 ml Falcon tube, taking care not to contaminate the material with gut tissue and contents. The material was grinded with a pestle after addition of 5 ml of lysis buffer (10 mM Tris-Cl pH 7.5, 400 mM NaCl, 2 mM EDTA pH 8). 200 µl 20% SDS and 300 µl 5 mg/ml proteinase K was added and samples were incubated for several hours at 50°C until tissue was completely digested. To remove RNA contamination, 4ul RNAase/ml was added to each tube, and tubes were left on a slowly rotating wheel for 15min. Five ml phenol equilibrated with 500 mM Tris pH 8 was added and tubes were placed on a slowly rotating wheel for 15 minutes, after which they were centrifuged at 3500 rpm for 10 minutes. The aqueous phase was transferred to another Falcon tube. Five ml phenol/chloroform equilibrated with 500 mM Tris pH 8 was added and tubes were placed on slowly rotating wheel for 15 minutes, after which they were centrifuged at 3500 rpm for 10 minutes. The aqueous phase was transferred to another Falcon tube. Five ml chloroform equilibrated with TE was added, and tubes were placed on a slowly rotating wheel for 15 minutes, after which they were centrifuged at 3500 rpm for 10 minutes. The aqueous phase was transferred to another Falcon tube, after which 0.5 volume of isopropanol was added and DNA was allowed to precipitate by gently swirling the tube for a few minutes. The DNA was caught on a hook made from a heated tip of a Pasteur pipette and all excess solution was drained before the DNA was transferred to a 15ml Falcon tube containing 1 ml TE. Tubes were placed overnight on a gently rocking platform to allow for the DNA to be re-suspended.

To aid genome assembly and annotation, we extracted and pooled RNA from workers, soldiers, king and queen from four colonies of *M. natalensis* (Mn115, Mn116, Mn117, and Mn118), all collected in South Africa in 2011. RNA was extracted using a modified Qiagen RNeasy Plant mini Kit (cat.no. 74903) and a traditional phenol/chloroform protocol. Tissue samples were subjected to mini-prep treatment in RLC buffer from the kit. Beta-mercaptoethanol (10 µl/ml) and RLC buffer was added up to a total volume of 1 ml. Subsequently, two phenol/chloroform steps and a chloroform step were performed, using premixed phenol/chloroform pH 8 (Sigma, cat.no. P2069-100ML) and for each step, one sample volume phenol/chloroform or chloroform was added, tubes were vortexed for 30 seconds, and centrifuged for 5 min at 13000rpm. The upper phase was transferred to a new Eppendorf tube after each step. After the chloroform step, the sample was mixed with ethanol and further processed as described in the Qiagen RNeasy Plant mini Kit manual. After extraction, RNA concentrations and quality were evaluated using NanoDrop.

2.2 *Termitomyces* DNA extraction

Pure culture of the P5 homokaryon was transferred to generate lawns on Petri plates (9.5cm diameter) containing potato dextrose agar (PDA). DNA extractions were performed from each of 10 Petri plates using the DNA extraction protocol described above for *M. natalensis*.

2.3 Metagenome DNA extractions

Whole guts were dissected from thawed RNAlater®-stored major workers, minor soldiers and the queen from *M. natalensis* colony Mn117. For workers and soldiers, duplicate extractions were performed on 2*50 guts pulled out of individuals aseptically, which were pooled after extraction. The entire queen gut was dissected out aseptically and suspended in RNAlater® and multiple extractions were done, each on ca. 300mg gut material and content, and extracts were subsequently pooled. For all gut samples, DNA extraction was done with a modified Qiagen Blood and Cell culture mini-kit protocol (Qiagen, cat. No. 13323) with a chloroform extraction step following protease K incubation. In the final step, one volume chloroform/isoamyl alcohol (24/1) was added, tubes were placed on a slowly rotating wheel for 15min, and subsequently spun at 3000g for 10 min. The supernatant was transferred to the spin columns included in the Qiagen kit, and the remaining of the manufacturer's protocol was subsequently followed.

3 Genome sequencing and assembly

3.1 Assembly of the *Macrotermes natalensis* genome

We employed a whole genome shotgun (WGS) sequencing approach and used Illumina HiSeq2000 to produce the genome sequences. The whole genome DNA samples of *M. natalensis* were used to construct nine paired-end sequence libraries with different insert-size at 200bp, 250bp, 500bp, 800bp, 2kb, 5kb, 10kb and 20kb. To construct small (≤ 800bp) insert-size libraries, ca. 5μg of DNA was sheared to fragments, end-repaired, A-tailed, and ligated to Illumina paired-end adapters (Illumina). The ligated fragments were then selected at desire size on agarose gels and amplified by LM-PCR. For large insert-size libraries, around 20-40μg of DNA was sheared into desire insert size using nebulization for 2 kb or HydroShear (Covaris) for 5 kb, 10 kb and 20 kb. DNA fragments were then end-repaired with biotinylated nucleotide analogues, size selected at 2, 5, 10 and 20kb, and finally circularized by intra-molecular ligation. Circular DNA molecules were sheared with Adaptive Focused Acoustic (Covaris) at an average size of 500bp, and the biotinylated fragments were purified with magnetic beads (Invitrogen). These fragments were end-repaired, A-tailed and ligated to Illumina paired-end adapters, size-selected again and amplified by LM-PCR. All libraries were sequenced on the Illumina HiSeq 2000 and a total of more than 130Gb of sequence data was generated (Table S1).

Before assembly, several filtering steps were done to exclude low quality reads. The following filtering criteria for raw reads were used: 1) filter reads with ≥ 5% of Ns or polyA; 2) filter reads with ≥ 50 low-quality bases (Phred score ≤ 7); 3) filter reads with adapter contamination; 4) filter paired reads overlapping each other with ≥ 10bp (allowing 10% mismatch); 5) filter PCR duplications (reads were considered duplications when read 1 and read 2 of the same paired-end reads were identical). After filtering, 89.8Gb high quality reads were retained (Table S1).

The sequencing data from libraries of 200bp and 500bp insert-sizes were used to estimate the genome size of *M. natalensis* by K-mer analysis. According to a 17-mer distribution (Figure S1a), the genome size of *M. natalensis* was estimated to be 1.309Gbp (Table S2). We applied SOAPdenovo (v2.03, 3) to assemble the genome with optimized parameters (parameter“-K 39-d 0 -M 3–D 1 –F”). Contigs were first constructed with the data from small (≤ 800bp) insert-size libraries. Scaffolds were then joined by the contigs using paired-end information from small and large insert-size libraries. Gapcloser (v1.10, released with SOAPdenovo, default parameters) and kgf (v1.18, released with SOAPdenovo, default parameters) were introduced to do the local reassembly for the unresolved gap regions. To cover gaps as much as possible at the gap-filling step, we utilized 83.5Gb sequencing data from small insert-size libraries and reads from termite gut samples that could be mapped to the termite genome (see also Section 3.3). Only 6.8GB of the termite sequence data from the gut metagenome sequencing could be mapped to the assembly. After the gap filling process, the total size of final assembly was 1.17Gb, with a contig N50 of 15Kb and a scaffold N50 of almost 2Mb (Table S3).

After assembly, all original reads were aligned to the scaffold sequences with SOAPaligner (v2.21, 4, default parameters) to evaluate sequencing coverage. Based on the short-read alignment results, we calculated the depth of each base, producing an average coverage depth for the entire genome of 69X (Figure S1c). The overall GC content of the genome was 39.9%. GC content for non-overlapping 5kb sliding windows on the genome was determined, and the GC versus depth scatter plot indicates no obvious GC bias during sequencing (Figure S1e).

To identify potential presence of contaminated sequences in the assembly, we used BLASTn (5) to search against the NCBI nucleotide collection database of bacteria and fungi. Scaffold sequences were considered as candidate contaminated sequences if the BLASTn hit e-value was smaller than 1e-5 and the alignment length was larger than 50% of the entire length. BLASTn found no long and/or high-score alignments, so this assembly is unlikely to contain contaminated sequences.

In order to evaluate the completeness of the termite genome assembly, we compared a set of core eukaryotic genes to our assembly sequence using the Core Eukaryotic Genes Mapping Approach (CEGMA) (v2.4, 6). Out of 248 ultra-conserved core eukaryotic genes (CEGs), 246 could be aligned to the *M. natalensis* genome and 233 these with alignment lengths of more than 70% of the length of the protein coded for (Table S3).

3.2 Assembly of the *Termitomyces* genome

Termitomyces was sequenced and assembled with similar methods as *M. natalensis*. 12.5Gb of raw reads were generated from 5 libraries with different insert-size in HiSeq 2000 (Table S10). After filtering, 6.82Gb of high-quality reads were used in SOAPdevono for assembly of the genome, which has an estimated size in 83.7Mbp based on K-mer analyses (Table S11). The final assembly of *Termitomyces* has a contig N50 of 22Kb, a scaffold N50 of 262Kb, and total length of 68.5Mb (Table S11). After assembly, reads were aligned to the genome using SOAPaligner, providing an average depth estimate of about 81X (Figure S1d), with the GC versus depth scatter plot indicating no obvious GC bias during sequencing (Figure S1f). We also checked for contaminated sequences in this assembly using the same method as for the termite genome and found no indications of contamination. CEGMA (6) was also here used to evaluate the completeness of the *Termitomyces* genome assembly, and 244 of the 248 CEGs could be aligned to the genome (240 of which had alignment lengths >70% of the length of the coded protein length, Table S11).

3.3 Metagenome assemblies

A paired-end DNA library was constructed from DNA from a single queen gut, 100 worker guts, and 100 soldier guts from *M. natalensis*. For each of these samples, ca. 5µg of DNA was first sheared to ~350bp fragments, and then end-repaired, A-tailed, and ligated with Illumina paired-end adaptors (Illumina). The ligated fragments were selected at the desire size on agarose gels and amplified by LM-PCR. The libraries were then sequenced on the Illumina HiSeq 2000, with read lengths of typically 90-100bp (Table S17).

Before assembly, low quality reads were filtered, so that reads with more than 3 Ns or more than 40 low quality bases or adaptor sequence were removed. Next, all reads were mapped on to the termite and *Termitomyces* assemblies by allowing 10 mismatches to filter out host-contaminated reads (Table S18). A total of 35.7Gb and 14Mb sequence reads from the metagenome sequencing mapped to the termite and *Termitomyces* genome assembly, respectively, and these reads were consequently removed.

High-quality clean reads of each gut sample were assembled using SOAPdenovo (v2.03, parameter “-L 100 -R -D 1”), optimized by testing different K-mer lengths. Assemblies with the longest contig N50 were chosen as the final result. A K-mer size of 29bp was chosen for the queen gut metagenome and 49bp for soldier and worker gut metagenomes, producing an assembly of ca. 33Mb for the queen gut metagenome, ca. 446Mb for the worker, and ca. 337Mb for the soldier gut metagenome (Table S19). To measure the read usage and the depth of each assembly, we mapped the clean reads from each gut to their corresponding assemblies using SOAPaligner with default parameters (Table S20). 11% of the clean reads for the queen gut, 31% for the worker guts and 26% for the soldier guts could be mapped back to their corresponding assemblies. Thus, the average depth we estimated for queen gut, worker gut and soldier gut assemblies was 7.69X, 5.46X and 4.38X respectively.

4 Repeat annotations

We identified known transposable elements (TEs) in the genome using RepeatMasker (v3.2.9, Smit AFA http://www.RepeatMasker.org.) and RepeatProteinMask against the Repbase for the genomes of both *M. natalensis* and *Termitomyces*. Furthermore, RepeatScout (7) and PILER (8) were used to do *de novo* repeat finding in the *Termitomyces* genome, while only PILER was used for the *M. natalensis* genome. Based on the *de novo* annotations, we constructed a repeat library using RepeatScout with default parameters. This library was then used as a reference by RepeatMasker to identify additional high and medium copy repeats (>10 copies) in the assemblies. We also predicted tandem repeats using TRF (v4.4, 9). Repetitive elements constituted ca. 67% of the assembly of the *M. natalensis* genome, which is a very high score among the insect genomes sequenced so far. In *Termitomyces*, repetitive elements constituted ca. 65% of the assembly, implying that the unassembled regions of both assemblies likely contained many repetitive elements (Table S4).

5 Protein-coding gene annotations

5.1 Annotation of the *M. natalensis* genome

We used homology-based annotation, *de novo* annotation and transcriptome data to predict protein-coding genes in the assembly, and the results of these three methods were integrated for obtaining a final gene set.

Protein sequences from 10 species were used in homology-based predictions: *Apis mellifera*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* and six ants (*Atta cephalotes*, *Acromyrmex echinatior*, *Camponotus floridanus*, *Harpegnathos saltator*, *Linepithema humile* and *Pogonomyrmex barbatus*). Protein sequences of each gene set were aligned to the termite genome by TBLASTN. For each aligned region on the termite genome, the most similar homolog, which at the same time was longer than 50% of the query protein length, was selected. Gene structures were predicted by GeneWise (10) based on homology information. To reduce false positives, only predictions with CDS length of >150bp and a GeneWise score of >50 were kept.

AUGUSTUS (11) and SNAP (12) were used for *de novo* annotation. The training genes for the two programs were 500 randomly selected genes with complete ORFs from the homology-based annotation of *Apis mellifera*. With these training genes, SNAP and AUGUSTUS estimated the parameters and predicted gene models. To reduce false positives, we only kept *de novo* predictions that were supported by both methods for subsequent analyses.

We generated 4.17 Gb of RNA-seq reads for annotation purposes (Table S5). Tophat (13) was used to align raw reads to the genome in order to identify exon-exon splice junctions, and Cufflinks (14) was used to reconstruct transcripts from the spliced alignments. 38,774 transcripts were assembled. Subsequently, the transcripts assembled from the RNA-seq data were used to improve the gene set. First, a Markov model was estimated from the 500 training gene set used in de novo annotation by two awk scripts which are included with Geneid gene annotation tools (v1.3) (15). For the exon sequences, we estimated the transition probability distribution of each nucleotide given the penta-nucleotide preceding it for each of the three possible frames and an initial probability matrix from the pentamers observed at each codon position using the awk program MarkovMatrices.awk. For the intron sequences, a single transition matrix was computed as well as a single initial probability matrix using the awk script MarkovMatrices-noframe.awk. Then the coding potential of each reading frames in the transcript were computed based on the Markov model. Transcripts with complete ORFs were picked out and the redundant isoforms were removed by keeping the longest ORF for each locus. At last, we identified 7062 transcripts with a complete ORF. Then these ORFs were integrated with GLEAN annotation by replacing the incomplete GLEAN gene models. After integration, several steps were done to refine the gene set. Gene models with good evidence that were not part of the integrated gene set were added to the final gene set: (1) genes in homology predictions that have complete ORFs and GeneWise (10) scores larger than 80; (2) complete ORFs inferred from transcripts; and (3) *de novo* predictions with a putative SwissProt function (see section 5.3). According to Interpro and SwissProt annotations, we filtered out genes related to transposable elements. Some genes were manually checked and problematic genes were corrected before downstream analyses. The final gene set that we obtained had 16,310 genes (Table S6).

5.2 Annotation of the *Termitomyces* genome

The methods for annotation of *Termitomyces* were similar as for *M. natalensis*. We performed *de novo* and homology-based annotation and then integrated these into a final gene set.

Protein sequences of seven fungal species (*Saccharomyces cerevisiae, Aspergillus fumigatus, Agaricus bisporus, Coprinopsis cinerea, Laccaria bicolor, Pleurotus ostreatus*, and *Schizophyllum commune*) were used to perform homology-based annotation. AUGUSTUS and SNAP were introduced to generate the de novo predictions. The training set was 500 randomly selected genes with complete ORFs from homology annotation of *Saccharomyces cerevisiae*. Only de novo predictions supported by both programs were kept.

We merged the *de novo* and homologous annotations by GLEAN. Some homologous annotations with good evidences were not included in the integrated gene set. We added those homologous annotations with complete ORFs and we manually curated some gene models during analyses. The final gene set contained 11,556 genes (Table S12).

We assessed the quality of assembly and gene annotation by aligning 1382 ESTs of *Termitomyces* from *Macrotermes gilvus* downloaded from GenBank (16). Of these, 1165 aligned to the *Termitomyces* genome, and 86% of them were annotated as genes, suggesting that most gene regions were assembled and annotated (Table S13).

5.3 Metagenome annotations

Gene predictions for bacteria and archaea in the three gut microbiomes were done using the combined GeneMark-P* and GeneMark.hmm-P software with pre-computed models based on 265 sequenced genomes from NCBI (17).

6 Functional annotation

Several functional databases were searched to assign putative functions for the predicted genes. SwissProt (18) annotations were assigned according to the best match of the alignments generated by BLASTP, requiring aligned ratios > 0.5 for both query and target sequences. InterproScan (19) was used to annotate motifs and domains of translated proteins. Gene sequences were searched against SUPERFAMILY, Pfam, PRINTS, PROSITE, ProDom, Gene3D, PANTHER and SMART databases in Interpro. GO terms for each gene were obtained from the Interpro database. KEGG annotation (20) was performed using the KAAS online server (21), and the SBH method against the eukaryotic species set (Tables S7, S14).

For each metagenome, KEGG annotation was done by the KAAS online server using the SBH method against the prokaryotic species set, while COG annotation for each gene was determined by BLASTp (5) to the COG database with an e-value cut-off of 1e-5. The statistics of the results of functional annotations are given in Table S21.

7 ncRNA annotation

Four types of ncRNAs were annotated in the *M. natalensis* and *Termitomyces* genomes: microRNA (miRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and small nuclear RNA (snRNA). We used tRNAscan-SE (22) and INFERNAL (23) to predict ncRNAs in the genome. The tRNA genes were predicted by tRNAscan-SE with eukaryote parameters. The rRNA fragments were identified by aligning the rRNA template sequences from invertebrate animals in Rfam (release 9.1, 24) using BLASTn (5) with an E-value cut-off 1E-5. The miRNA and snRNA genes were predicted by searching the Rfam database with INFERNAL. The results of ncRNA annotation are shown in Tables S8 and S15.

8 Gene family evolution

8.1 Construction of gene family trees

To gain insight into the evolution of termite gene families, we clustered the genes of the following 14 genome-sequenced species: *Acromyrmex echinatior, Acyrthosiphon pisum, Apis mellifera*, *Caenorhabditis elegans, Camponotus floridanus, Drosophila melanogaster, Daphnia pulex, Harpegnathos saltator, Nasonia vitripennis, Pediculus humanus, Tribolium castaneum, Atta cephalotes,* and *M. natalensis*. The Treefam method (25) was used for constructing gene families. The statistics of the clustering result are given in Table S9.

Similarly, we chose seven genome-sequenced fungal species in addition to *Termitomyces* to construct gene families: *Saccharomyces cerevisiae, Aspergillus fumigatus and five Agaricales: Agaricus bisporus, Coprinopsis cinerea, Laccaria bicolor, Pleurotus ostreatus,* and *Schizophyllum commune*. The statistics of the clustering result are given in Table S16.

8.2 Expanded and contracted gene families in *M. natalensis*

CAFE (26) was used to detect gene families that have undergone expansion or contraction in *M. natalensis* compared to other species. CAFE uses a stochastic model of gene birth and death to infer statistically significant gains and losses in gene families, with a phylogenetic tree and a table of gene copy numbers in each organism. A family-wide significance threshold of 0.05 was used. We checked the candidate families detected by CAFE to filter out artifacts (i.e., clustering biases, incorrect annotations). We found three reliable contracted families in the *M. natalensis* lineage (Table S22), but no reliable expanded families. Phylogenies of these families are shown in Figure S2.

Two of the contracted families (esterase FE4 and trypsin) are associated with digestive systems. Esterase FE4 (Figure S2a) had best hits to ESTF_MYZPE in the SwissProt database, producing 13 members in *M. natalensis* but 16-34 members in the other genome sequenced insects (16 in the fungus-growing leafcutter ant *A. echinatior*, at least 20 in other species). Previous studies in aphids found that esterase FE4 is involved in resistance to insecticides (27-28). The contracted family trypsin (Figure S2c) has only one member in *M. natalensis*, but at least 3 members in other genome-sequenced species. Trypsin (EC 3.4.21.4) is a serine protease found in the digestive system of many organisms (29). A previous study in mosquitoes showed that members of the trypsin gene family play a crucial role during the digestion of the blood meal in the gut (30). Contraction of these two families may thus be attributed to the relatively uniform diet of *M. natalensis*.

Another contracted gene family belonged to the short-chain dehydrogenase/reductase (SDR) superfamily (Figure S2b), which is a very large protein family whose members play critical roles in lipid, amino acid, carbohydrate, cofactor, hormone and xenobiotic metabolism as well as in redox sensor mechanisms (31). In this contracted family, there are 9 members in *M. natalensis* in comparison to 11-25 present in other genome-sequenced insects. Interestingly, the fungus-growing ant *Acromyrmex echinatior* and the honeybee (*A. mellifera*), both with relatively uniform diets, also appear to have reduced members of this family: 11 in *A. mellifera* and 13 in *A. echinatior*, while other species have at least 16. This contraction could therefore also be associated with diet. No candidate gene family losses were identified in *M. natalensis* after gene clustering and careful checking.

8.3 Expanded and contracted gene families in *Termitomyces*

We identified gene family expansions in *Termitomyces* by comparing to other *Agaricales*, and families that have twice the number of genes in *Termitomyces* compared to the average number in other fungi were also considered as expanded families. To reduce false positives, we subsequently checked the annotation and function of the genes in these families. We found 10 reliable gene family expansions (Table S23), among which some have hydrolase activity and are involved in carbohydrate metabolism (e.g., chitinase and acid phosphatase), implying a possible role in plant biomass degradation (Figures S3-S5). We also investigated possible gene family contractions in *Termitomyces* based on gene clustering, but did not find any good candidates.

8.4 Gene losses in *Termitomyces*

We investigated putative gene losses in *Termitomyces* based on gene clustering with *Saccharomyces cerevisiae*, *Aspergillus fumigatus* and other *Agaricales*, defining gene loss as families with no homologs in *Termitomyces*, but with members being present in other fungi. To reduce the risk of false positives, several screening steps were applied. All homologous genes in these families were aligned to the *Termitomyces* gene set by BLASTp (5). If no homologous genes were found in *Termitomyces*, we re-ran GeneWise to make sure the lost genes were not due to mistakes in the annotation. If no good gene structure could be found, genes were considered lost in *Termitomyces*, resulting in four reliable families with no members in *Termitomyces* (Table S24), among which two (α-glucosidase and oxysterol-binding protein) were found in *M. natalensis* and one (α-glucosidase) was also found in the three metagenomes (Table S25). α-glucosidase belongs to glycoside hydrolase family 13 and is required for galactose metabolism (KEGG 'map00052') and starch and sucrose metabolism (KEGG 'map00500'). Thus, the loss of α-glucosidase in *Termitomyces* is likely to be a significant marker of functional complementarity in the symbiosis with the termite gut microbiota.

9 Phylogenetic placement of *M. natalensis* and *Termitomyces*

Phylogenetic trees of insects and fungi were constructed based on single-copy gene families after obtaining the clustering results of the 14 insects and 8 fungi described above (Tables S9 and S16). Peptide sequences were aligned by MUSCLE (32) and transformed into codon based CDS alignments. A supergene was constructed by concatenating these alignments. The first and second sites of each codon were used for tree construction. We chose a maximum likelihood method implemented in PhyML (v3.0, 33) to build phylogenetic trees, using the HKY85 model and 100 bootstrap runs. The unrooted tree was converted to a rooted tree by minimizing height, using TreeBeST (v1.9.2, http://treesoft.sourceforge.net/treebest.shtml).

We estimated the *M. natalensis* divergence time from the phylogenic tree of insects with three calibration times (*T. castaneum*-*D. melanogaster*: 355.0 MYA, *P. humanus*-*A. pisum*: 172.6 MYA and *C. elegans*-*Arthropoda*: 970.0 MYA) obtained from the Time Tree database (34). For the fungal phylogeny, only one calibration time (*Ascomycota*-*Basidiomycota* 908.0MYA ~ 1208.0MYA) was used. For the insect phylogeny, we used r8s (v1.8) and for the fungus phylogeny mcmctree in the PAML package (v4.4d) to perform estimations. The divergence time in fungi is less accurate because there is no suitable reference for divergence time among basidiomycetes and the only evidence for calibration is in the root of the tree. The phylogenetic trees with our divergence time estimations are given in Figure S6.

10 Comparative analyses of gut metagenomes

10.1 Phylogenetic classification of metagenomic reads

To obtain assignment of metagenome reads, we performed two separate classification analyses, PhymmBL (v3.2, 35) and BLASTn (5). Classification using PhymmBL was initially performed on the assembled reads, so that assignment could be done on the longest possible sequences. PhymmBL uses both interpolated Markov models (IMMs) and BLAST to taxonomically classify sequences, providing confidence scores at each taxonomic level. The database for running PhymmBL was downloaded from the Entrez Genome database of NCBI and included complete and draft genomes of bacteria, archaea, fungi and protozoa, as well as 12 bacteria draft genomes: 3 *Bacillus*, 2 *Pantoae*, 2 *Trabulsiella* and 5 *Enterobacter*. Subsequently, we used the paired reads from the original data set to blast sequence reads against the PhymmBL database of assembled genus-assigned contigs and about 50% of the reads for each sample could be mapped to the resulting database.

For BLASTn searches, we filtered using an e-value cut off of 0.05. Only the best hit per sequence was kept, and the non-redundant alignment length for a read aligned to the sequences of a genus had to exceed 30bp.

Using the combined PhymmBL and BLASTn results, we estimated the relative abundance of different genera present in the three metagenome samples, by counting the number of reads assigned to individual genera (Table S26). Only paired reads that obtained the same genus-level assignment with both methods were kept. In order to compare the genus-level diversity and skew in prokaryote genus-level-abundance profiles between workers, soldiers and the queen gut, we calculated Shannon-Weaver indices for all three (Fig. 4, main text) (36).

Lastly, we plotted rarefaction curves to check whether the sequencing depth of each metagenome was sufficient to cover the expected number of genera present, which the results (Figure S7) suggest is the case.

11 Survey for genes of particular interest

11.1 Carbohydrate active enzymes (CAZymes)

We used a combination of BLASTp (5) and HMM (37) to identify putative CAZymes in *Macrotermes natalensis* and *Termitomyces*. All predicted genes were subject to a full-length BLASTp (Blast v2.2.28+) (5) search against the CAZyme database. Sequences with a positive hit were thereafter checked using two different approaches. First, we used BLASTp (5) of all hits to a library built from GH, PL, CE, GT and CBM domains. This library was built by downloading the CAZyme sequences from NCBI according to the accession number provided by the CAZy database. Then, according to the GenBank annotation of the protein domain, we cut the domain sequences out and created a blast database of these domains. For example, we cut the region of Glyco_hydro_1 in http://www.ncbi.nlm.nih.gov/protein/ADD08340.1 as our domain sequence for GH1. The positive hits were also subjected to a HMMer (v3.1b1) (37) search using hidden Markov models built by aligning domain sequences for each CAZyme using MUSCLE, after which the alignment was used to build the HMM model in HMMer. If both methods placed a putative CAZyme in the same CAZy family, it was considered a reliable annotation. Tables S27 and S28 provide overviews of the proteins assigned to CAZymes for *Termitomyces* and *M. natalensis*, respectively. We performed a similar analysis for the recently genome-sequenced termite *Zootermopsis nevadensis* genome (38) (Results in Table S28). A comparison of *Termitomyces* CAZy profiles to those of 99 other fungi (39) is given in Table S30.

We used FASTY (40), part of the FASTA package v36.3.6d, with an e-value cut-off of 10^-6 to identify CAZymes in the metagenomes. The sequence libraries for the CAZy database used were obtained on 2013-03-22 and the results are given in Table S29. The absolute number of glycoside hydrolases identified in each symbiotic colony component (*Termitomyces*, *M. natalensis*, and gut microbiomes), along with known family activities based on www.cazy.org, is given in Table S31.

To assign putative bacterial genera to CAZymes in gut microbiomes, we combined the results of PhymmBL analyses on assembled metagenome reads and the CAZy analyses of identified genes. To validate the PhymmBL taxonomy assignments, we aligned the assembled metagenome sequences to the complete and draft genomes of bacteria, archaea, fungi and protozoa in the Entrez Genome database of NCBI using BLASTn (5). The results were filtered by the e-value cutoff 0.05 and for CAZymes where the total non-redundant alignment length was >150bp. The genus of the genome with the best hit was taken as the BLASTn validation result. The genera assignments to CAZymes were only included in the comparisons they were the same using both PhymmBL and BLASTn. The results of these analyses are given in Table S32.

9 To explore the extent to which the fungus-growing termite gut microbiome has changed functionally compared to wood-feeding higher termites, we did a comparison of the relative abundance of glycoside hydrolases (GH) in families identified in *M. natalensis* worker guts to those identified in *Nasutitermes* sp. (41), *Amitermes wheeleri* and *Nasutitermes corniger* (42), and in *Odontotermes yunnanensis* (43). Euclidean GH similarity distances calculated in R (package vegetarian) were recalculated after 10,000 Monte Carlo permutations of the distance matrix to obtain a non-parametric p-value, which showed that GH profiles of the fungus-growing termite microbiomes are significantly more similar to each other than to other termite guts (p=0.04). More specifically, compositions have shifted towards fewer GHs in families targeting more complex polysaccharides, while families targeting simpler plant components are enriched (Table S33; Fig. 3A). This analysis thus provides an independent test of our main conclusion that the fungus-growing termite gut has adjusted its functional role after the domestication of *Termitomyces*.

11.2 Mating type related genes in *Termitomyces*

We identified mating-type related genes (44) in *Termitomyces* and the results are listed in Table S34.

11.3 Immune genes and antimicrobial peptides in *M. natalensis*

We checked the *M. natalensis* genome for genes involved in immune defense. Amino acid sequences of *D. melanogaster* immune defense genes were used with BLASTp (BLAST, NCBI) to search for orthologous genes in *M. natalensis*. If the e-values of hits following the best match did not differ significantly from the e-value of the best match, up to five sequences were chosen. To validate these BLAST results, they were re-blasted against the NCBI database. We also checked the antimicrobial peptides in *M. natalensis*. Genes of *M. natalensis* were also used to search against the CAMP (Comprehensive Antimicrobial Peptide Database) database by BLASTp (5). The final result filtered out any BLAST hits with e-values of less than 1e^{-6}. These analyses revealed the presence of all orthologous genes and major immune-defense pathways of *Drosophila*, but we recovered only two antimicrobial peptides: a defensin and an ortholog of termicin (45) (Table S35).

11.4 Insect neuropeptides and protein hormones

Neuropeptides, protein hormones, and biogenic amines and their receptors steer central physiological processes such as reproduction, development, growth, feeding, and behavior (46-48). We found 39 neuropeptide genes in the genome of *M. natalensis*, while some neuropeptide genes present in other arthropods were absent. This neuropeptide "barcode" of presence-absence is characteristic for *M. natalensis* and not found in any other arthropod with a sequenced genome. We expect that this barcode must reflect the physiology of the termite but, at present, we do not understand the functional significance of the *M. natalensis* neuropeptide spectrum (Table S36).

9 A remarkable feature in *M. natalensis* is the very high copy number (fourteen) of allatostatin-A peptides present in the preprohormone. A similar high copy number of allatostatins has been found in the allatostatin preprohormone of the termite, *Reticulitermes flavipes* and in several cockroaches, which are known to be close relatives of termites (49). In termites and cockroaches, allatostatin-A peptides inhibit juvenile hormone (JH) production from the *corpora allata*, a pair of endocrine organs located near the brain (50). JH plays an important regulatory role in caste determination in termites and other social insects: Elevated JH titers in the termite *R. flavipes* cause pre-soldier termites to differentiate into soldiers (51) and, the other way around, soldiers suppress pre-soldier differentiation through a rapid decrease of JH titers in pre-soldiers (52). These findings therefore suggest that the high number of allatostatin-A peptides in the termite preprohormone might be involved in the regulation of caste determination.

1 We identified 39 neuropeptides, protein hormones, and biogenic amines and their receptors involved in central physiological processes such as reproduction, development, growth, feeding, and behavior (48), which formed a unique presence/absence spectrum in comparison to all other known arthropod genomes (Tables S36). Although *M. natalensis* had a high copy number of allatostatin-A, similar to cockroaches and the termite *Reticulitermes flavipes* (52), where these peptides have been hypothesized to inhibit juvenile hormone production and to mediate caste differentiation, respectively (53-54). This suggests that allatostatin-A peptides present in the termite pre-prohormone must be involved in the regulation of caste determination in the species, which has the most complex societies with several morphologically distinct altruistic castes of any termite.

12 *Termitomyces* growth on different media

For determining growth profiles on different carbohydrate substrates, we used Serpula minimal medium (55) adjusted to pH 6.0 and containing 1.5% agar (Invitrogen). Carbon sources were added at concentrations as indicated at www.fung-growth.org. Plates were inoculated in duplicate with 4µl of a solution containing 500 spores/µl and mycelium fragments, and cultures were grown at 25°C for 10 days. Figure S8a shows the spectrum of growth profiles of *Termitomyces* P5 on a range of minimal and complex media, and Figure S8b compares the growth profiles between *Termitomyces* and *Coprinopsis cinerea* Okayama7, *Pleurotus ostreatus* PC9, *Schizophyllum commune* H4-8, and *Aspergillus fumigatus* Af293 (see also www.fung-growth.org for complete growth profiles).

13 Supporting references

1. Sonnenberg AS *et al.* (2010) Protoplast transformation of filamentous fungi. *Methods Mol Biol* 638:3-19.
3. Li R *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20:265-272.
4. Li R *et al.* (2009*)* SOAP2: an improved ultrafast tool for short read alignment. *Bioinform* 25:1966-1967.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410.

6. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinform* 23:1061-1067.

7. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinform* 21:i351-i358.

8. Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinform* 21:i152-i158.

9. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucl Acids Res* 27:573-580.

10. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14:988-995.

11. Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinform* 19:ii215-ii225.

12. Korf I (2004) Gene finding in novel genomes. *BMC Bioinf* 5:59.

13. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinform* 25:1105-1111.

14. Trapnell C *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech* 28:511-515.

15. Parra G, Blanco E, Guigó R (2000). Geneid in Drosophila. *Genome Res* 10:511-515.

16. Johjima T, Taprab Y, Noparatnaraporn N, Kudo T, Ohkuma M (2006) Large-scale identification of transcripts expressed in a symbiotic fungus (*Termitomyces*) during plant biomass degradation. *Appl Microbiol Biotech* 73:195-203.

17. Lukashin A, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucl Acids Res* 26:1107-1115.

18. Boeckmann B *et al.* (2003) The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucl Acids Res* 31:365-370.

19. Zdobnov EM, Apweiler R (2001) InterProScan- an Integration Platform for the Signature-Recognition Methods in InterPro. *Bioinf* 17:847-848.

20. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl Acids Res* 28:27-30.

21. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucl Acids Res* 35:W182-W185.

22. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl Acids Res* 25:955-964.

23. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. *Bioinform* 25:1335-1337.

24. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. *Nucl Acids Res* 31:439-441.

25. Li H *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucl Acids Res* 34:D572-580.

26. De Bie T, Cristianini N, Demuth JP, Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinform* 22:1269-1271.

27. Field LM, Devonshire AL (1998) Evidence that the E4 and FE4 esterase genes responsible for insecticide resistance in the aphid *Myzus persicae* (Sulzer) are part of a gene family. *The Biochem J* 330:169-173.

28. Field LM, Williamson MS, Moores GD, Devonshire AL (1993) Cloning and analysis of the esterase genes conferring insecticide resistance in the peach-potato aphid, *Myzus persicae* (Sulzer). *The Biochem J* 294:569-574.

29. Rawlings ND, Barrett AJ (1994) Classification of peptidases. *Meth Enzymol* 244:1-15.

30. Müller HM, Crampton JM, della Torre A, Sinden R, Crisanti A (1993) Members of a trypsin gene family in *Anopheles gambiae* are induced in the gut by blood meal. *EMBO J* 12:2891-2900.

31. Kavanagh KL, Jörnvall H, Persson B, Oppermann U (2008) Medium- and short-chain dehydrogenase/reductase gene and protein families: the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. Cell Mol Life Sci 65:3895-906.

32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* 32:1792-1797.

33. Guindon S, Delsuc F, Dufayard JF, Gascuel O (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology* (Clifton, N.J.) 537:113-137.

34. Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. *Bioinform* 22:2971-2972.

35. Brady A, Salzberg, SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Meth* 6:673-676.

36. Shannon CE, Weaver W (1949) Mathematical theory of communication. Urbana, Ill.: Univ. Illinois Press. 117 pp.

37. Eddy SR (1998) Profile hidden Markov models. *Bioinform* 14:755-763.

38. Terrapon N et al. (2014) Molecular traces of alternative social organization in a termite genome. Nat Commun 5:3636 doi: 10.1038/ncomms4636.

39. Zhao et al. (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics 14:274.

40. Pearson WR, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. *Genomics* 46:24-36.

41. Warnecke F *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450:560-565.

42. He S *et al.* (2013) Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites. *PLoS ONE* 8:e61126.

43. Liu N *et al.* (2013) Metagenomic insights into metabolic capacities of the gut microbiota in a fungus-cultivating termite (*Odontotermes yunnanensis*). *PLoS ONE* 8:e69184.

44. Stajich JE *et al.* (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc Natl Acad Sci USA* 107:11889-11894.

45. Lamberty M et al. (2001) Insect immunity - Constitutive expression of a cysteine-rich antifungal and a linear antibacterial peptide in a termite insect. *J Biol Chem* 276:4085-4092.

46. Hauser F, Cazzamali G, Williamson M, Blenau W, Grimmelikhuijzen CJP (2006) A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Prog Neurobiol* 80:1-19.

47. Hauser F *et al.* (2008) A genome-wide inventory of Neurohormone GPCRs in the red flour beetle *Tribolium castaneum. Front Neuroendocrinol* 29:142-165.

48. Nässel DR, Winther AM (2010) *Drosophila* neuropeptides in regulation of physiology and behavior. *Prog Neurobiol* 92:42-104.

49. Elliot KL, Hehman GL, Stay B (2009) Isolation of the gene for the precursor of Phe-Gly-Leu-amide allatostatins in the termite *Reticulitermes flavipes*. *Peptides* 30:855-860.

50. Stay B, Tobe SS (2007) The role of allatostatins in juvenile hormone synthesis in insects and crustaceans. *Annu Rev Entomol* 52:277-299.

51. Scharf ME, Buckspan CE, Grzymala TL, Zhou X (2007) Regulation of polyphenic caste differentiation in the termite *Reticulitermes flavipes* by interaction of intrinsic and extrinsic factors. *J Exp Biol* 210:4390-4398.

52. Watanabe D, Gotoh H, Miura T, Maekawa K (2011) Soldier presence suppresses presoldier differentiation through a rapid decrease of JH in the termite *Reticulitermes speratus*. *J Insect Physiol* 57:791-795.

53. Yagi KJ, Elliott KL, Teesch L, Tobe SS, Stay B (2008) Isolation of cockroach Phe-Gly-Leu-amide allatostatins from the termite *Reticulitermes flavipes* and their effect on juvenile hormone synthesis. *J Insect Physiol* 54:939-938.

54. Lenkic L, Tiu HKS, Tobe SS (2009) Suppression of JH biosynthesis by JH analog treatment: mechanism of suppression and roles of allatostatins and nervous connections in the cockroach Diploptera punctata. *J Insect Physiol* 55:967-975.

55. Eastwood DC *et al.* (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. *Science* 333:762-765.

14 Supporting Tables

Table S1 Genome sequencing data for *Macrotermes natalensis*

Library ID	Read length/bp	Insert size/bp	Raw data/Mb	Raw Sequence depth/X	Filtered data/Mb	Filtered Sequence depth/X
MACgnnDAKDCAAPE	90	200	25,724.93	19.64	21690.06	16.56
MACgnnDAKDIAAPE	90	500	15,322.86	11.70	12188.90	9.31
MACgnnDAKDIAAPEI-11	100	500	8,782.45	6.7	7244.97	5.53
MACgnnDAKDMAAPE	90	800	26,705.6	20.39	19490.11	14.88
MACgmdDAADWAAPEI-3	49	2k	15,840.98	12.10	11712.62	8.94
MACgmdDAADLAAPEI-10	49	5k	12,633.33	9.65	7045.83	5.38
MACgmdDAADTAAPEI-9	49	10k	11,301.74	8.63	3567.68	2.72
MACgnnDAADUAAPEI-33	49	20k	6,878.69	5.25	694.36	0.53
SZAXPI000995-1	150	250	7,004.85	5.35	6195.99	4.73
Total			130,195.43	99.41	89830.52	68.58

Table S2. *M. natalensis* genome size estimation with 17-mer analyses

Species	Kmer numbers	Kmer depth	Estimated genome size	#Used bases	#Used reads	Cove rage

M. natalensis	27,501,748,278	21	1,309,607,060	33,878,965,270	398,576,062	25.86
Termitomyces	4,853,344,855	58	83,678,359	5,836,300,775	61,434,745	69.75

Table S3 Statistics and completeness assessment of the *M. natalensis* assembly.

	Contigs		Scaffold s	
	Size (bp)	Number	Size (bp)	Number
N90	3,178	78,470	331,359	653
N80	6,009	53,502	829,393	440
N70	8,851	38,287	1,198,304	323
N60	11,981	27,458	1,578,248	238
N50	15,640	19,314	1,997,143	173
Longest	243,672		10,840,804	
Total Size	1,115,012,471		1,172,292,920	
Total Number (>100bp)		282,004		145,794
Total Number (>2kb)		94,151		4,504
CEGMA analysis		Total CEGs	Aligned CEGs	Percent (%)
≥70 % aligned		248	233	93.95
All			246	99.19

Table S4 Statistics of repeats in *M. natalensis* and *Termitomyces* assemblies.

	Macrotermes natalensis		*Termitomyces*	
Type	Length (bp)	% in genome	Length (bp)	% in genome
DNA	34,170,043	2.9148%	1,247,470	1.8214%
LINE	63,196,604	5.3909%	921,120	1.3449%
SINE	432,439	0.0369%	5,551	0.0081%
LTR	3,963,034	0.3381%	11,857,172	17.3121%
Microsatellite	11,666,058	0.9951%	40,208	0.0587%
Minisatellite	69,190,396	5.9021%	808,332	1.1802%
Satellite	37,009,413	3.1570%	399,780	0.5837%
Other	19,331	0.0016%	146	0.0002%
Unknown	567,183,781	48.3824%	29,160,564	42.5759%
Total	786,831,099	67.1190%	44,440,343	64.8852%

Table S5 Summary of the RNA-seq data for *M. natalensis*.

Total reads Number	Number of reads mapped to genome	Mapped proportion (%)	#Assembled transcripts
46,349,928	38,625,090	83.33%	38,774

Table S6 Statistics of predicted protein-coding genes in *M. natalensis*

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	37673	10190.41	885.47	3.77	234.65	3356.88
	SNAP	72761	25556.31	907.50	5.45	166.54	5542.19

Homolog	*A.cephalotes*	13028	9452.49	914.62	4.23	216.27	2646.02
	A.echinatior	25443	6758.46	681.11	3.10	219.53	2892.47
	A.mellifera	10432	14467.12	1158.04	5.39	214.70	3031.00
	C.elegans	4995	10106.90	982.02	4.30	228.41	2767.70
	C.floridanus	32304	6011.55	712.85	2.89	246.59	2804.24
	D.melanogaster	7812	16127.08	1236.32	5.82	212.59	3094.31
	H. sapiens	12687	7945.18	923.98	3.83	241.51	2486.67
	H. saltator	27590	5865.65	727.83	2.92	248.88	2671.79
	L. humile	17368	8310.31	880.36	3.83	230.08	2630.87
	P. barbatus	14152	9106.24	899.53	4.08	220.71	2670.22
RNA-seq		38774	7187.29	1296.10	2.54	511.11	3837.79
Final gene set		16310	12471.71	1069.88	4.88	219.18	2937.69

Table S7 Functional annotation of the *M. natalensis* genome

		Gene Number	Percent (%)
Total		17362	-
Annotated	Swiss-Prot	11712	67.46
	KEGG	4068	23.43
	InterPro	10069	57.99
	GO	7839	45.15
Unannotated		4818	27.75

Table S8 Non-coding RNA genes in the assembly of *M. natalensis*

Type		Copy	Average length (bp)	Total length (bp)
miRNA		96	114.59	11,001
tRNA		92	76.28	7018
rRNA	rRNA	85	122.93	10449
	18S	23	198.35	4562
	28S	3	137.33	412
	5.8S	1	108.00	108
	5S	58	92.53	5367
snRNA	snRNA	184	114.57	21,080
	CD-box	9	95.56	860
	HACA-box	0	0.00	0
	splicing	37	141.62	5,240

Table S9 Summary of comparative gene clustering for *M. natalensis*

Species	Gene	#Clustered genes	#Cluster	#Unclustered genes
Acromyrmex echinatior	17278	13114	8426	4164
Acyrthosiphon pisum	33267	22710	8312	10557
Apis mellifera	10660	9726	7083	934
Atta cephalotes	18089	12027	8902	6062
Caenorhabditis elegans	20212	13356	4760	6856
Camponotus floridanus	16356	13750	8525	2606
Daphnia pulex	30899	21991	8096	8908
Drosophila melanogaster	13689	10334	5830	3355
Harpegnathos saltator	17191	14997	8465	2194

Macrotermes natalensis	17362	14695	6704	2667
Nasonia vitripennis	17084	14584	7533	2500
Pediculus humanus	10769	8607	6347	2162
Tribolium castaneum	16631	12246	7022	4385

Table S10 Genome sequencing data for *Termitomyces*

Lib ID	Read length (bp)	Insert size (bp)	Raw data (Mb)	Raw sequence depth (X)	Filtered data (Mb)	Filtered sequence depth (X)
MACheyDAMDCAAPEI-9	100	200	3980.18	47.57	3256.90	38.92
MACheyDAMDIAAPEI-1	100	500	2202.85	26.33	1869.70	22.34
MACheyDANDMAAPEI-2	100	800	1046.99	12.51	713.43	8.53
MACheyDAQDWAAPEI-19	49	2k	2502.66	29.91	376.96	4.50
MACgmdDAADWAAPEI-3	49	5k	2807.44	33.55	599.33	7.16
Total			12540.12	149.86	6816.32	81.46

Table S11 Statistics and completeness assessment of *Termitomyces* assembly.

	Contigs		Scaffolds	
	Size (bp)	Number	Size (bp)	Number
N90	5,589	3,088	55,182	288
N80	10,173	2,207	110,797	202
N70	14,448	1,649	159,423	151
N60	18,603	1,237	212,827	115
N50	22,750	908	262,000	86
Longest	125,259		932,855	
Total Size	67,802,156		68,490,755	
Total Number (>100bp)		15,693		11,244
Total Number (>2kb)		4,169		604
And complem analysis		Total CEGs	Aligned CEGs	Percent (%)
≥70 % aligned		248	240	96.77
All			244	98.39

Table S12 Statistics of predicted protein-coding genes in *Termitomyces*

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	9249	1759.55	1438.24	6.26814	229.452	62.9921
	SNAP	16017	1183.33	966.418	4.07517	237.148	72.5354
	Overlap	8844	1798.64	1470.61	6.37562	230.661	63.0232
Homolog	*A. fumigatus*	5221	1159.99	892.769	4.35913	204.804	81.551
	A. bisporus	9060	1531.7	1167.7	4.96744	235.07	93.7467
	B. cinerea	5642	2234.29	1558.35	5.93867	262.407	138.866
	L. bicolor	8236	1655.52	1288.86	5.43298	237.229	84.7118
	P. ostreatus	8408	1449.8	1144.81	5.30186	215.926	72.8974

	S. cerevisiae	4463	1758.59	929.874	3.46628	268.263	338.02
	S. commune	8620	1412.23	1090.96	4.92216	221.644	83.9115
Final gene set		11556	1729.26	1250.5	5.65914	220.971	104.757

Table S13 Assessment of *Termitomyces* genome assembly using EST

Dataset	Number	Total length (bp)	Covered by assembly	with >90% sequence in one scaffold		with >50% sequence in one scaffold	
				Number	Percent	Number	Percent
All	1165	563268	75.30%	274	23.52%	1021	87.64%
>100bp	1137	561330	76.40%	274	24.10%	1021	89.80%
>200bp	1071	551175	78.05%	272	25.40%	998	93.20%
>500bp	565	356203	86.67%	252	44.60%	564	99.82%

Table S14 Functional annotation of the *Termitomyces* genome

		Gene number	Percent (%)
Total		11556	-
Annotated	Swiss-Prot	6074	52.56
	KEGG	3262	28.23
	InterPro	6681	57.81
	GO	5104	44.17
Unannotated		4372	37.83

Table S15 Non-coding RNA genes in the assembly of *Termitomyces*

Type		Copy	Average length (bp)	Total length (bp)
miRNA		2	93	186
tRNA		382	86.6292	33179
rRNA	rRNA	165	84.0964	13960
	18S	135	87.0074	11746
	28S	26	74.1692	1944
	5.8S	4	67.5	270
snRNA	snRNA	23	135.043	3106
	CD-box	8	103.625	829
	HACA-box	0	0.00	0
	splicing	15	151.8	2277

Table S16 Summary of comparative gene clustering for *Termitomyces*

Species	Gene	#Clustered genes	#Cluster	#Unclustered genes
Aspergillus fumigatus	9630	6122	3695	3508
Agaricus bisporus	10438	8965	5521	1473
Coprinopsis cinerea	13544	10641	6136	2903
Laccaria bicolor	23132	18318	8023	4814
Pleurotus ostreatus	12330	10458	6109	1872
Saccharomyces cerevisiae	5882	3954	2540	1928
Schizophyllum commune	14652	11373	6249	3279
Termitomyces	11556	8483	5648	3073

Table S17 Statistics of raw data for gut metagenomic sequencing

Sample	Lib ID	Read length (bp)	Insert size (bp)	Data (Gb)

Queen gut	MACsnyMAEDFAAPEI-8	100	350	24.53
Worker gut	SZAXPI001321-1	100	350	15.61
Soldier gut	SZAXPI001322-2	100	350	20.42

Table S18 Statistics of gut metagenomic sequencing data after filtering

Sample	Raw reads	HQ reads	Clean reads	Clean reads (%)
Queen gut	245269798	213129834	27009470	11.01%
Worker gut	156142200	135125218	84862558	54.35%
Soldier gut	204170006	183801428	63067820	30.89%

Table S19 Gut metagenome contig and scaffold assemblies

Contigs

Sample Name	Number	Total length	n50	n90	Max length
Queen gut	213,887	30,991,251	128	104	13,310
Worker gut	1,364,893	386,222,446	288	152	84,184
Soldier gut	1,139,990	286,426,953	242	147	30,091

Scaffolds

Sample Name	Number	Total length	n50	n90	Max length
Queen gut	194,252	33,050,228	139	104	1,035,054
Worker gut	790,259	446,286,201	1,120	177	629,797
Soldier gut	663,398	337,369,669	1,002	167	113,561

Table S20 Reads mapped to the metagenome assemblies

Sample Name	Total Reads	Mapped to own contigs			Average Depth
		PE	SE	%	
Queen gut	19,189,874	1,573,588	577,991	11.21%	7.69
Worker gut	20,171,617	6,296,770	6,324,469	31.29%	5.46
Soldier gut	33,777,238	3,419,468	5,499,309	26.40%	4.38

Table S21 Functional annotation of the three gut metagenomes

		Queen gut		Worker gut		Soldier gut	
		Gene Number	Percent (%)	Gene Number	Percent (%)	Gene Number	Percent (%)
Total		52852	-	1231500	-	920601	-
Annotated	COG	7184	13.593%	342906	27.845%	249164	27.065%
	KEGG	2998	5.672%	128122	10.404%	94359	10.250%

Un-annotated	45556	86.195%	884873	71.853%	668689	72.636%

Table S22 Contracted gene families in *M. natalensis*

	Camponotus floridanus	*Acromyrmex echinatior*	*Apis mellifera*	*Nasonia vitripennis*	*Drosophila melanogaster*	*Acyrthosiphon pisum*	*Macrotermes natalensis*	P-value
Esterase EF4	27	16	20	34	28	24	13	<0.0001
Dehydrogenase/reductase SDR family	23	13	11	25	18	23	9	<0.0001
Trypsin	3	4	3	15	11	3	1	0.0002

Table S23 Expanded gene families in *Termitomyces*.

Protein name	*Termitomyces*	*Agaricus bisporus*	*Amanita thiersii,*	*Coprinopsis cinerea*	*Gloeophyllum trabeum*	*Hebeloma cylindrosporum*	*Laccaria bicolor*	*Piloderma croceum*	*Pleurotus ostreatus*	*Schizophyllum commune*	*Aspergillus fumigatus*	*Saccharomyces cerevisiae*
Acid phosphatase	5	0	0	0	0	3	1	3	0	2	2	0
Uncharacterized protein Mb0912	3	0	0	0	0	1	0	0	1	1	1	0
Polyamine oxidase	5	0	0	0	2	1	1	1	2	1	3	1
Pre-mRNA-splicing factor ISY1	4	1	1	1	1	1	1	2	1	1	1	1
Para-hydroxybenzoate--polyprenyltransferase, mitochondrial	8	3	4	4	6	3	2	2	3	5	4	2
Vacuolar protein sorting-associated protein 26B-like	4	1	1	1	1	1	1	1	1	1	1	1
Chitinase 1	5	4	6	1	2	3	1	1	3	2	5	1
Autophagy-related protein 13	4	1	1	0	1	1	1	0	1	1	1	0
Probable feruloyl esterase	7	0	0	0	2	0	0	0	0	3	5	0
Unsaturated rhamnogalacturonyl hydrolase	3	1	1	0	0	0	0	0	1	1	1	0

Table S24 Putative lost gene families in *Termitomyces.*

Protein Name	*Agaricus bisporus*	*Aspergillus fumigatus*	*Coprinopsis cinerea*	*Laccaria bicolor*	*Pleurotus ostreatus*	*Saccharomyces cerevisiae*	*Schizophyllum commune*
Smr domain-containing protein	4	1	1	1	1	1	1
Alpha-glucosidase MAL32	1	5	2	3	1	7	1
Oxysterol-binding protein homolog 2	1	1	1	2	1	2	1

54S ribosomal protein L51, mitochondrial	1	1	1	1	1	1	1

Table S25 Presence/absence in gut metagenomes of putatively lost *Termitomyces* and *M. natalensis* genes.

Protein Name	*M. natalensis*	Queen gut	Worker gut	Soldier gut
Smr domain-containing protein	0	0	0	0
Alpha-glucosidase MAL32	5	8	21	21
Oxysterol-binding protein homolog 2	5	0	0	0
54S ribosomal protein L51, mitochondrial	0	0	0	0

15 Supporting Figures with legends

Figure S1. 17-mer depth distribution of *M. natalensis* (**a**) and 17-mer depth distribution of *Termitomyces* (**b**). The distribution of sequencing depth of *M. natalensis* (**c**) and the sequencing depth distribution of *Termitomyces* **(d**). In the process of alignment, 5 mismatches were allowed for the long reads (≥90bp) and 2 mismatches for the short reads (<90bp). The GC-depth scatter plot for *M. natalensis* **(e)** and for *Termitomyces* **(f)**.

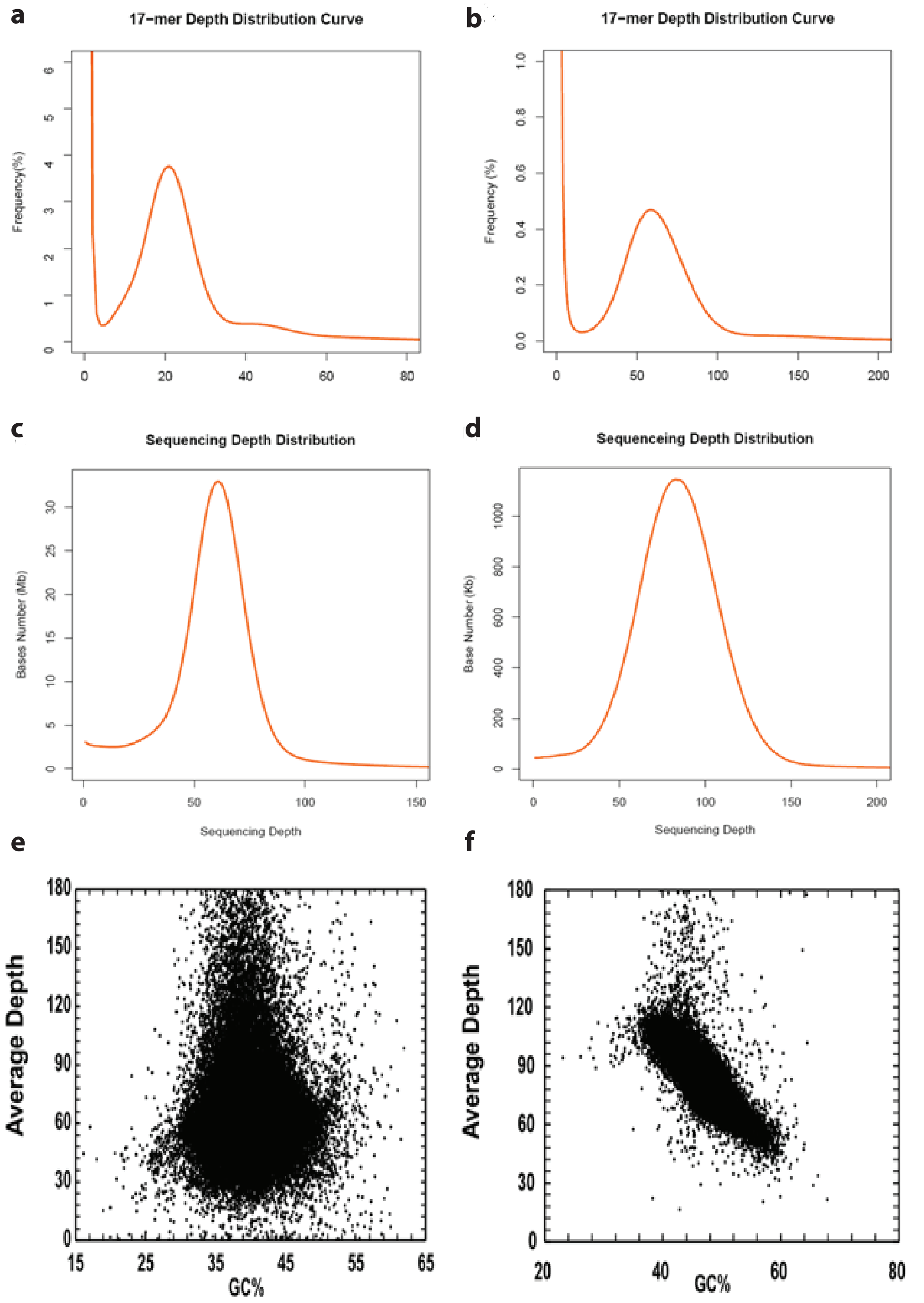


Figure S2. Contracted gene families in the *M. natalensis* genome (a) Phylogeny of the esterase EF4 gene family. **(b)** Phylogeny of the gene family of dehydrogenase / reductase SDR family member. **(c)** Phylogeny of the trypsin gene family. Abbreviations: MACNA = *M. natalensis*, NASVI = *Nasonia vitripennis*, CAMFL = *Camponotus floridanus*, DROME = *Drosophila melanogaster*, ACREC = *Acromyrmex echinatior*, ECYPI = *Acyrthosiphon pisum*, and APIME = *Apis mellifera*.

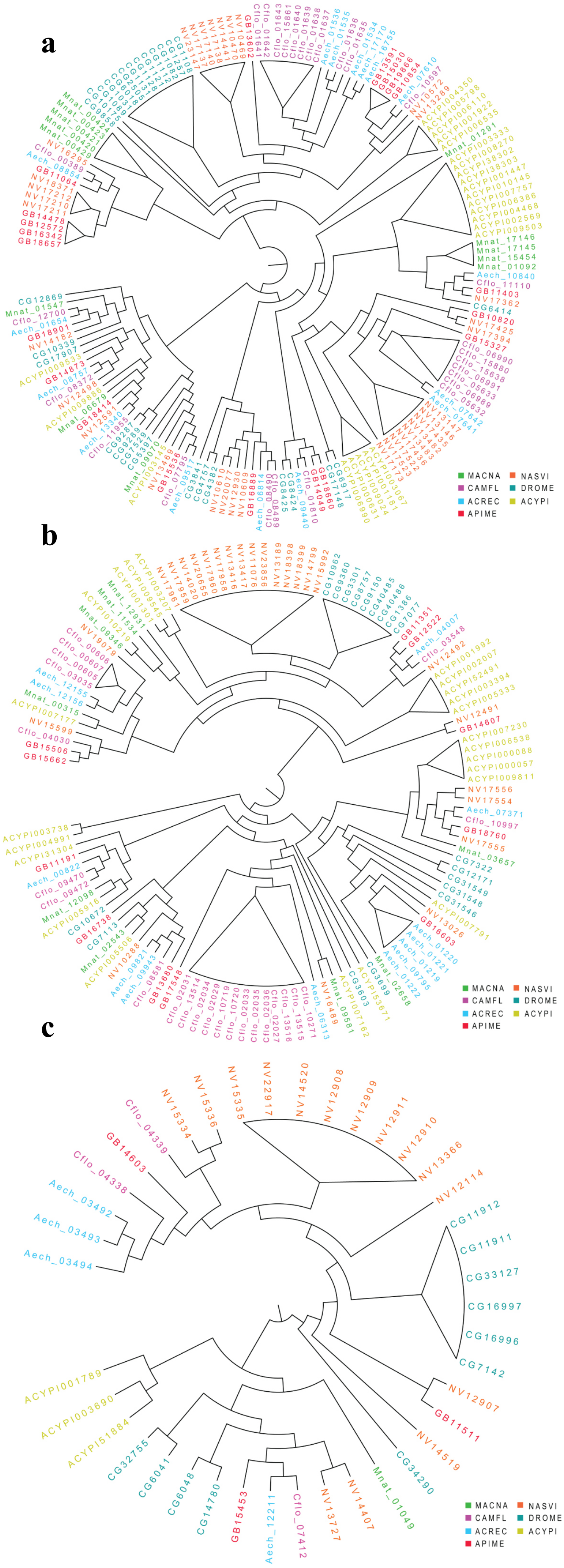


Figure S3. Maximum likelihood phylogenies of acid phosphatase **(a)**, uncharacterized protein Mb0912 **(b)**, polyamine oxidase **(c)**, and pre-mRNA-splicing factor ISY1 **(d)**.

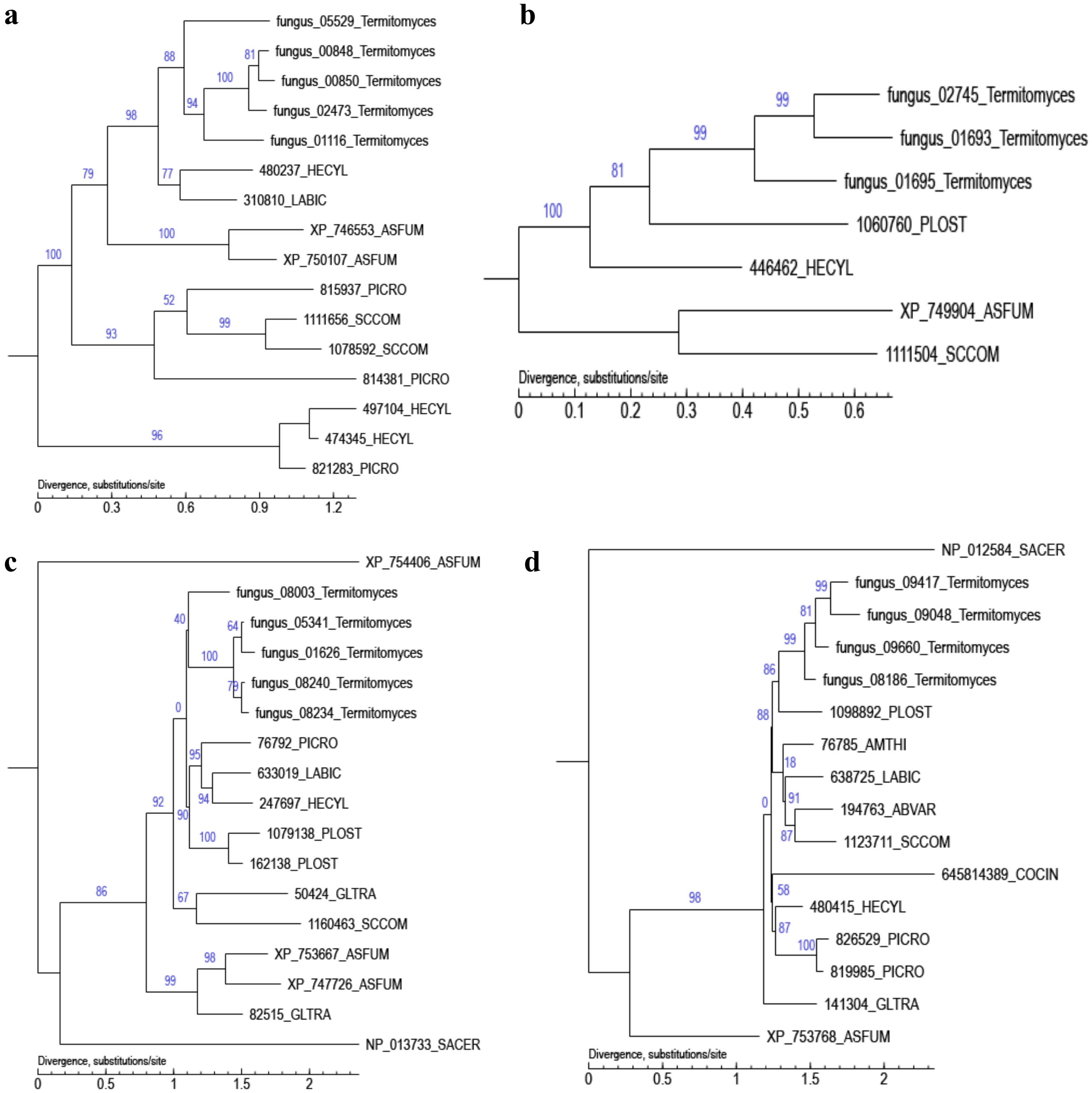


Figure S4. Maximum likelihood phylogenies of mitochondrial para-hydroxybenzoate-polyprenyltransferase **(a)** and chitinase 1 **(b)**.

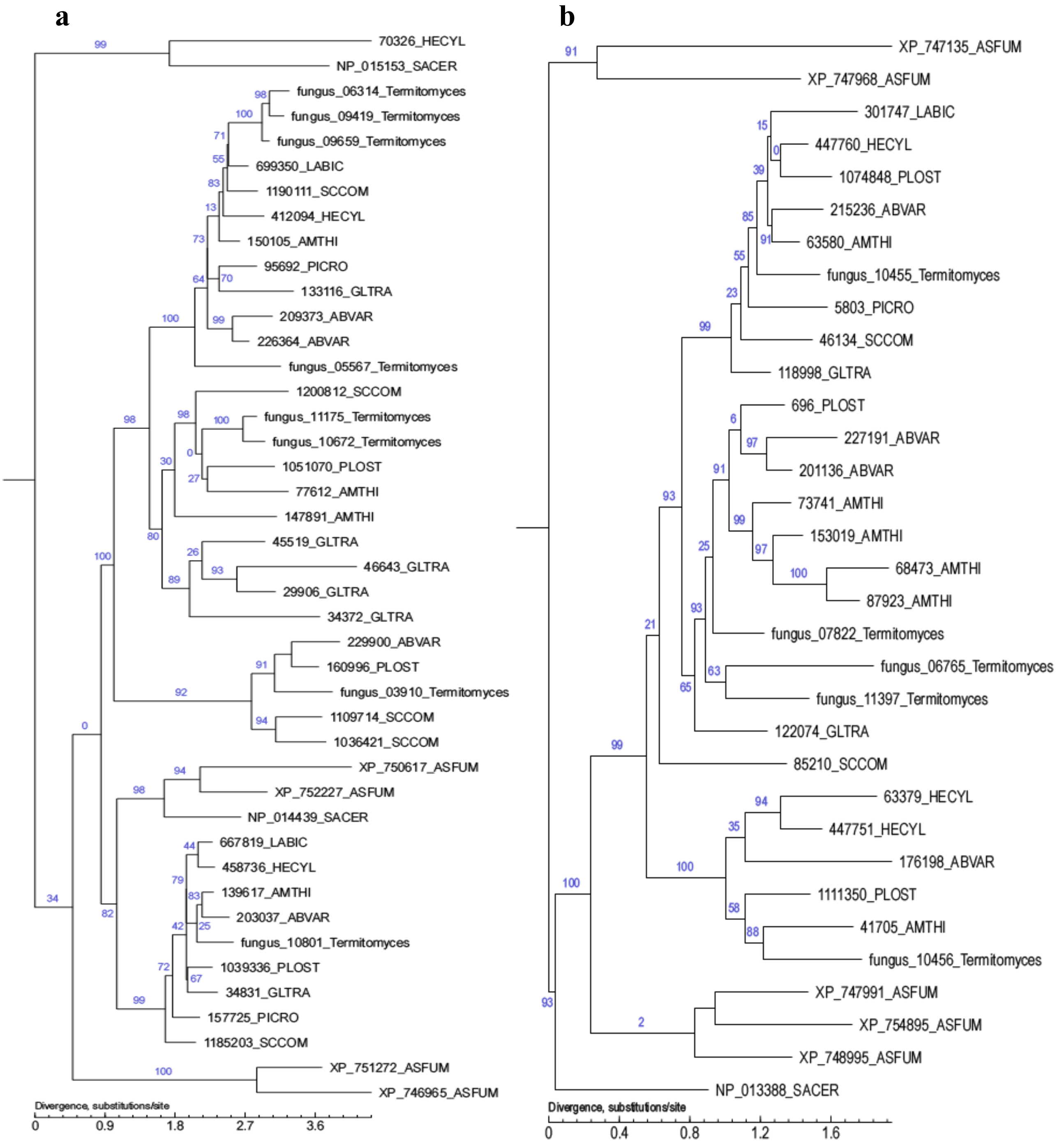


Figure S5. Maximum likelihood phylogenies of vacuolar protein sorting-associated protein 26B-like **(a)**, probable feruloyl esterase **(b)**, autophagy-related protein 13 **(c)**, and unsaturated rhamnogalacturonyl hydrolase **(d)**.

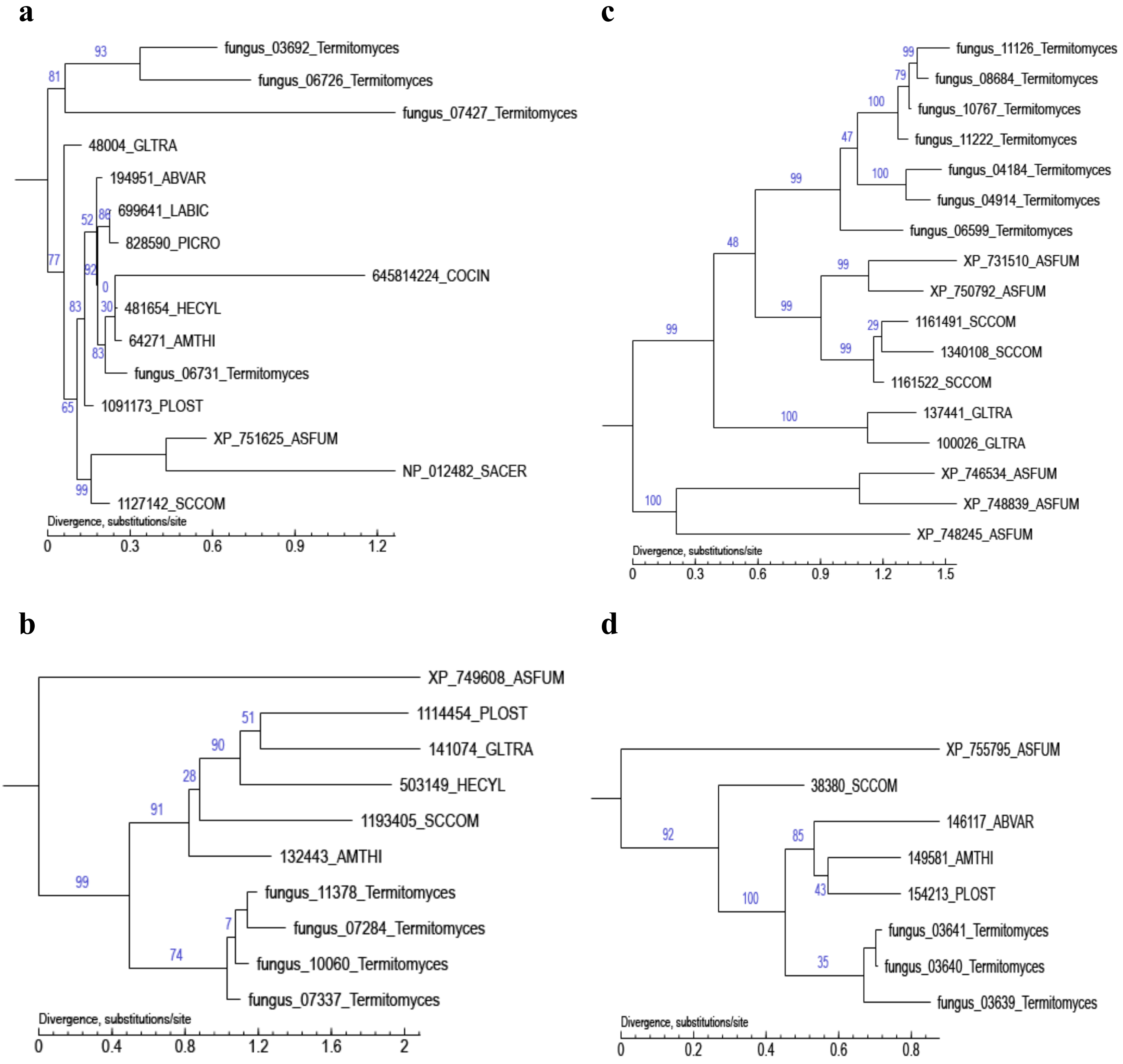
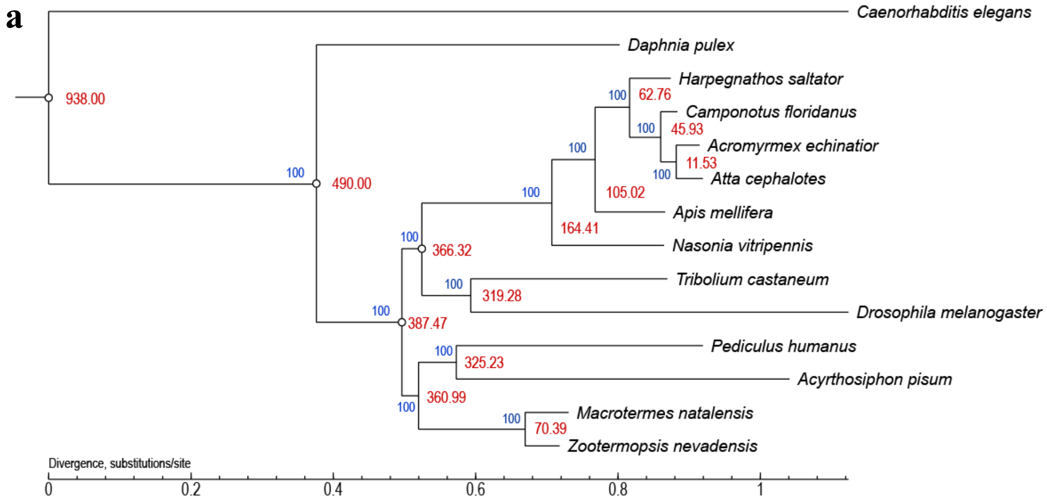

a
93
81
fungus_03692_Termitomyces
fungus_06726_Termitomyces
fungus_07427_Termitomyces
48004_GLTRA
194951_ABVAR
52
86
699641_LABIC
828590_PICRO
77
92
0
645814224_COCIN
83
30
481654_HECYL
64271_AMTHI
83
fungus_06731_Termitomyces
65
1091173_PLOST
XP_751625_ASFUM
NP_012482_SACER
99
1127142_SCCOM
Divergence, substitutions/site
0
0.3
0.6
0.9
1.2
c
99
fungus_11126_Termitomyces
79
fungus_08684_Termitomyces
100
fungus_10767_Termitomyces
47
fungus_11222_Termitomyces
99
100
fungus_04184_Termitomyces
fungus_04914_Termitomyces
48
fungus_06599_Termitomyces
99
XP_731510_ASFUM
99
XP_750792_ASFUM
29
1161491_SCCOM
99
99
1340108_SCCOM
1161522_SCCOM
100
137441_GLTRA
100026_GLTRA
XP_746534_ASFUM
100
XP_748839_ASFUM
XP_748245_ASFUM
Divergence, substitutions/site
0
0.3
0.6
0.9
1.2
1.5
b
XP_749608_ASFUM
51
1114454_PLOST
90
141074_GLTRA
28
503149_HECYL
91
1193405_SCCOM
132443_AMTHI
99
fungus_11378_Termitomyces
7
fungus_07284_Termitomyces
74
fungus_10060_Termitomyces
fungus_07337_Termitomyces
Divergence, substitutions/site
0
0.4
0.8
1.2
1.6
2
d
XP_755795_ASFUM
38380_SCCOM
92
146117_ABVAR
85
149581_AMTHI
100
43
154213_PLOST
fungus_03641_Termitomyces
35
fungus_03640_Termitomyces
fungus_03639_Termitomyces
Divergence, substitutions/site
0
0.2
0.4
0.6
0.8

Figure S6. (a) Maximum likelihood phylogeny of 14 insects with bootstrap support in blue and estimated divergence times in MYA in red. **(b)** Maximum likelihood phylogeny of eight fungi with bootstrap support in blue and estimated divergence times in MYA in red.



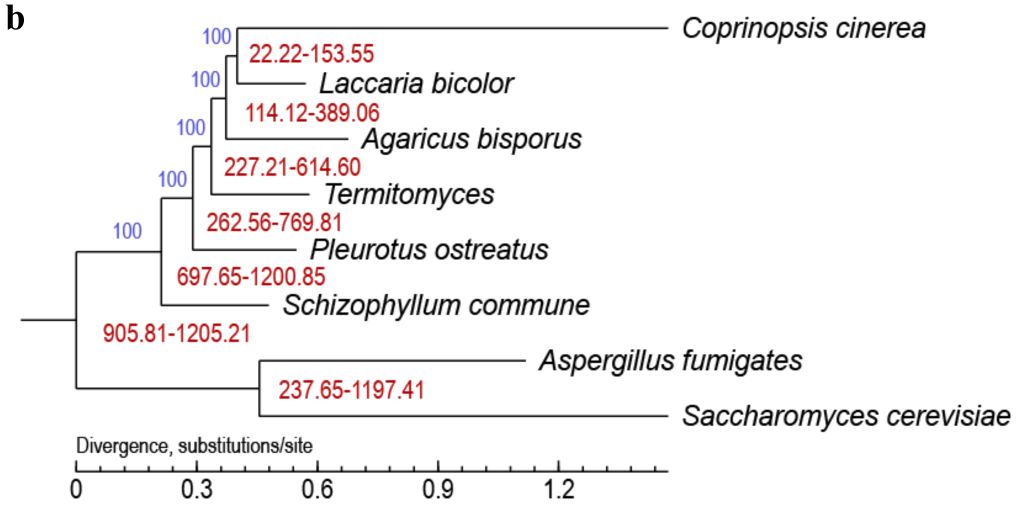


Figure S7. Rarefaction curves for the three gut metagenomes.

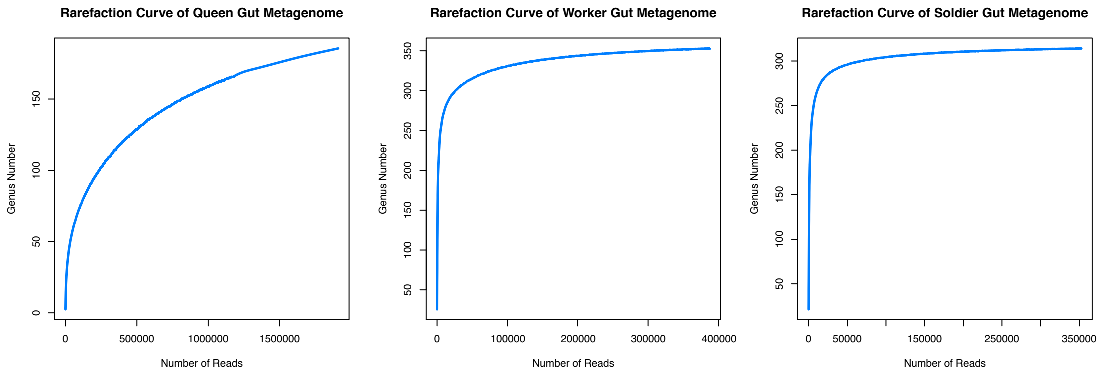
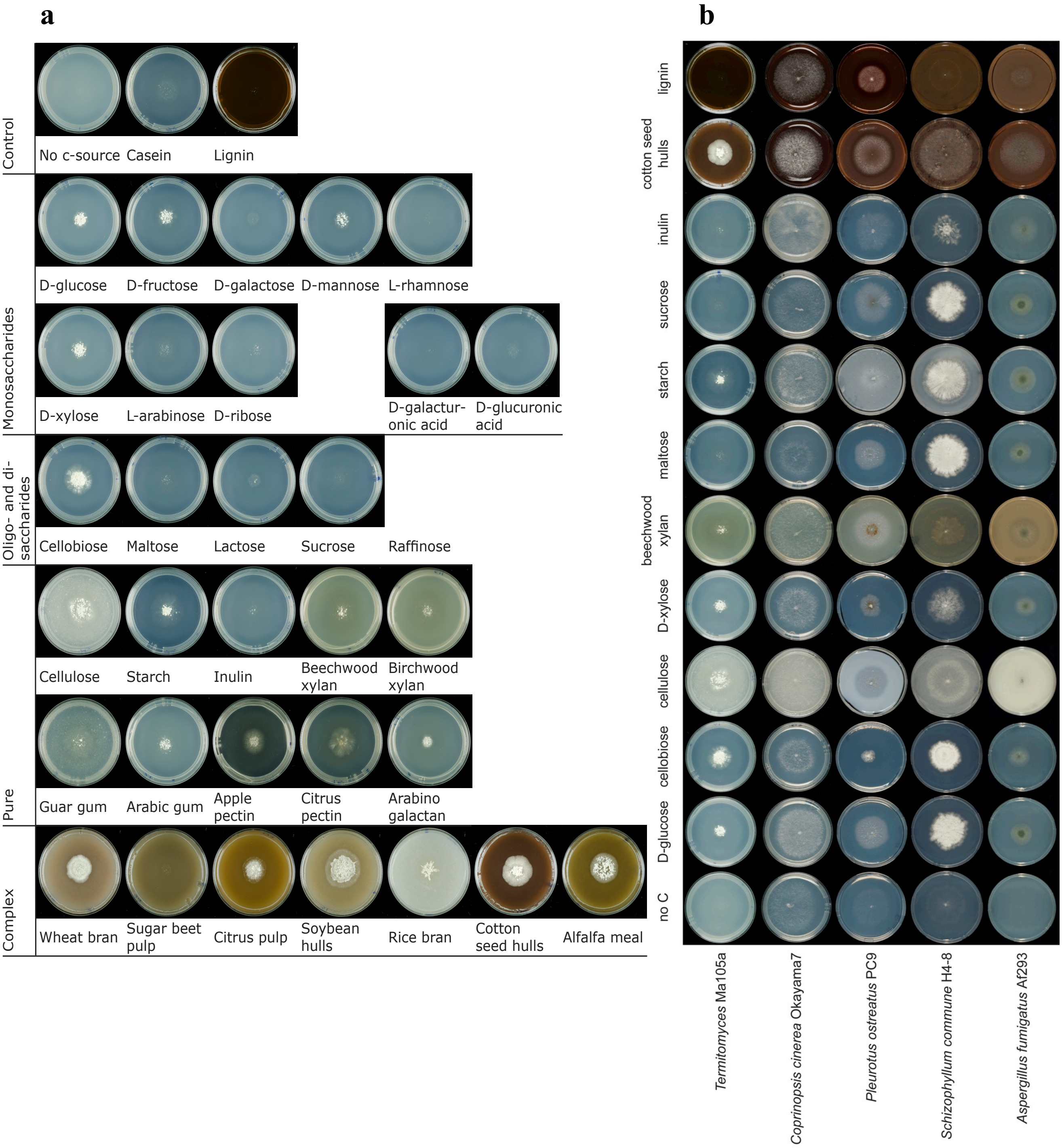
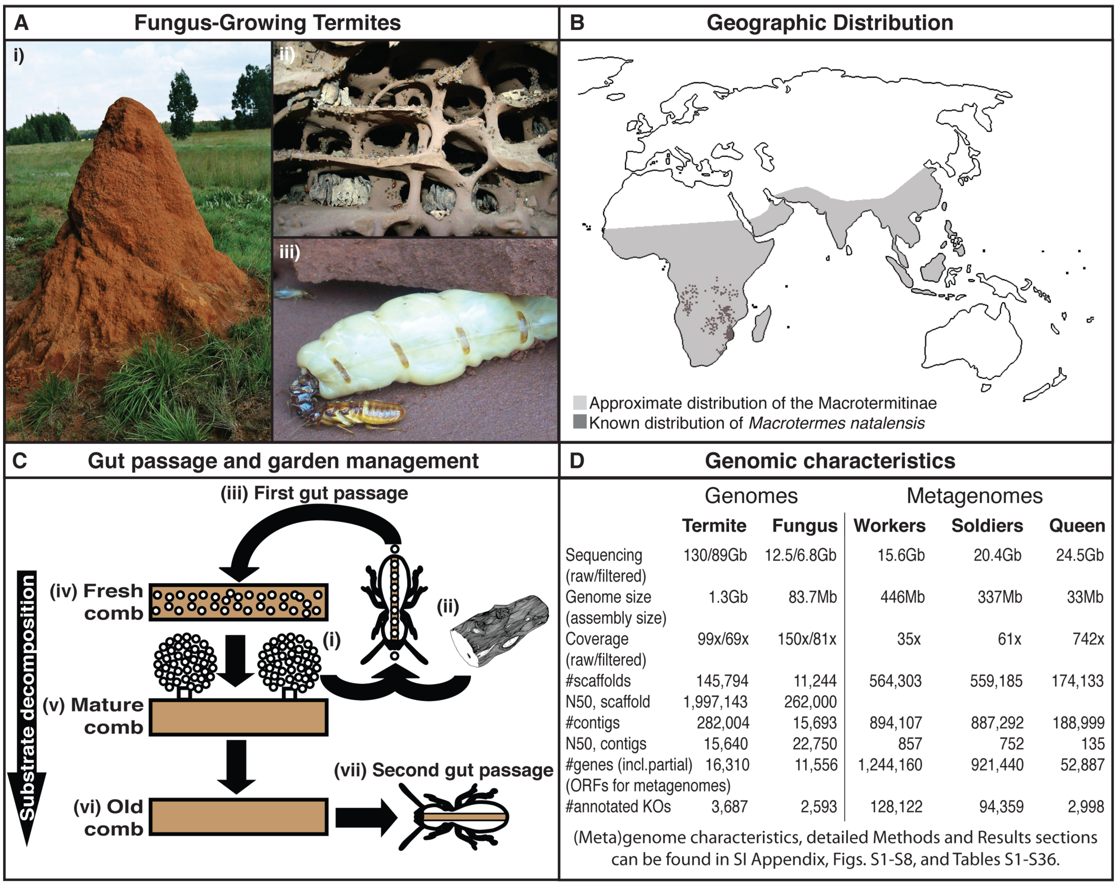


Figure S8. (a) Growth profiles of *Termitomyces* P5 on minimal and complex media. **(b)** Growth comparison of *Termitomyces* P5 with other fungi. For details and additional comparisons, see www.fung-growth.org.



16 Larger versions of main text figures with legends



	Genomes		Metagenomes		
	Termite	Fungus	Workers	Soldiers	Queen
Sequencing (raw/filtered)	130/89Gb	12.5/6.8Gb	15.6Gb	20.4Gb	24.5Gb
Genome size (assembly size)	1.3Gb	83.7Mb	446Mb	337Mb	33Mb
Coverage (raw/filtered)	99x/69x	150x/81x	35x	61x	742x
#scaffolds	145,794	11,244	564,303	559,185	174,133
N50, scaffold	1,997,143	262,000			
#contigs	282,004	15,693	894,107	887,292	188,999
N50, contigs	15,640	22,750	857	752	135
#genes (incl.partial) (ORFs for metagenomes)	16,310	11,556	1,244,160	921,440	52,887
#annotated KOs	3,687	2,593	128,122	94,359	2,998

(Meta)genome characteristics, detailed Methods and Results sections can be found in SI Appendix, Figs. S1-S8, and Tables S1-S36.

Figure 1. The fungus-growing termite symbiosis and its genomic characteristics. (A) A *Macrotermes natalensis* colony in South Africa (i), the underground fungus comb in which *Termitomyces* is maintained (ii)16, and the royal chamber with the queen and king (iii). **(B)** Geographic distribution of the Macrotermitinae (grey), with darker areas in southern Africa highlighting the known occurrences of *M. natalensis* (adapted from 61) **(C)** The substrate and recurrent *Termitomyces* inoculation within a colony centered around the termite gut: asexual *Termitomyces* spores from fungus comb nodules (i) and plant biomass substrate (ii) are mixed within the termite gut (iii; first gut passage) to become the new fungus comb substrate (iv) within which *Termitomyces* hyphae grow to maturity so that new nodules with asexual spores are produced (v) until the plant substrate is fully utilized and the old comb (vi) is consumed by the termites (vii; second gut passage). **(D)** To characterize the genetic potential of the fungus-growing termite symbiosis, we sequenced the genomes of *M. natalensis* and *Termitomyces*, and obtained gut metagenomes for workers, soldiers and the queen (for details, see SI Appendix and the GigaScience Database http://dx.doi.org/10.5524/100055).

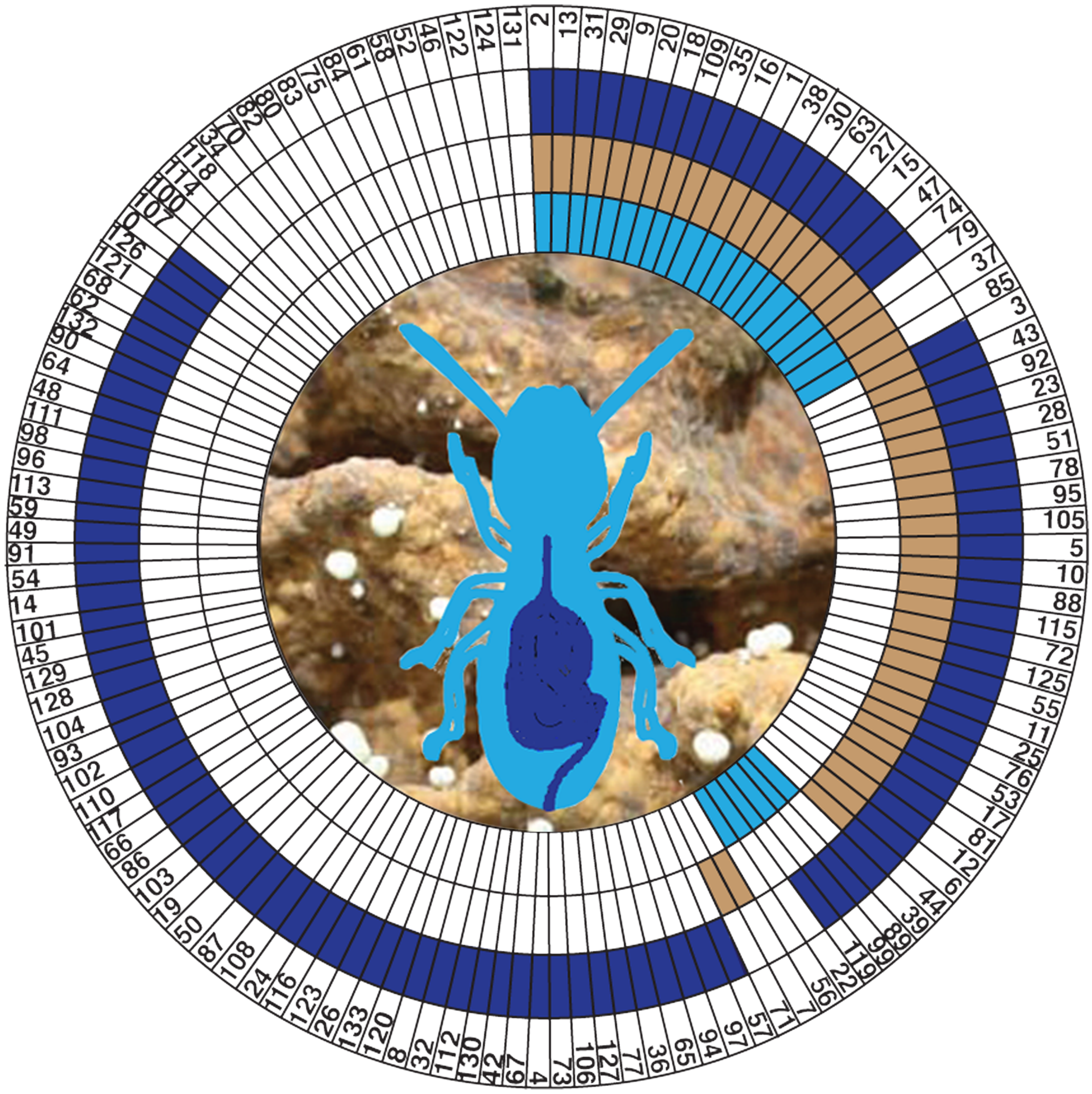


Figure 2. Functional complementary contributions to biomass degradation. Using the Carbohydrate-Active enZyme database (www.cazy.org), we classified glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate binding modules (CBM) and glycoside transferases (GTs) in the *Macrotermes*, *Termitomyces*, and worker gut microbes (Tables S27-S29). The GHs shown here were by far the most abundant enzyme class: 85 were identified in *M. natalensis* (27 GH families; light blue), 201 in *Termitomyces* (48 GH families, brown), and 15,619 in the worker gut microbiota (98 GH families; dark blue). The presence (color) / absence (white) pattern shows that the 111 GH families identified in the symbiosis represent 86.7% of all known GH families. Figures for the CBMs (73.9%), GTs (68,4%), PLs (78,3%), and CEs (100%) were of similar magnitude (Tables S27-S29). For enzyme names and key activities, including EC numbers, see Table S31.

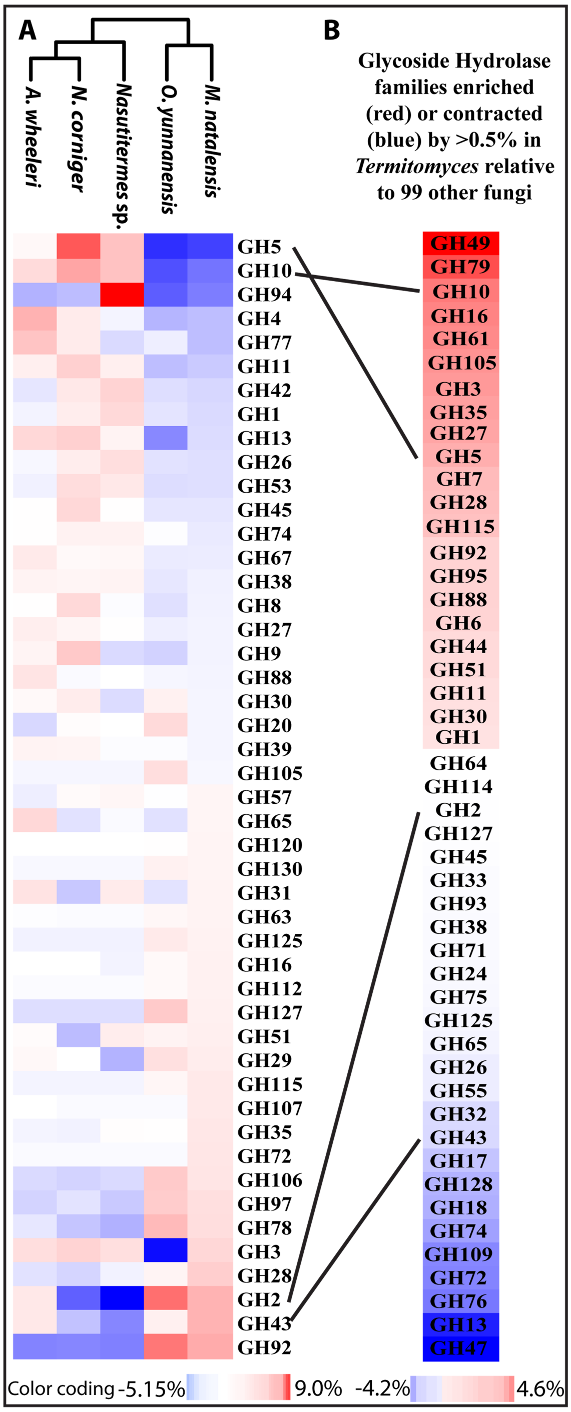


Figure 3. Complementary contributions to the spectrum of carbohydrate-active enzymes in *Termitomyces* and termite worker gut microbiota. **(a)** A heatmap of GH families enriched (red) or contracted (blue) in relative abundance across five termite species: the dung-feeding higher termite *Amitermes wheeleri* (34), two species of wood-feeding higher termites [*Nasutitermes corniger* (34); *Nasutitermes* sp. (35)], and two fungus-growing termite species (*Odontotermes yunnanensis*, 36; *M. natalensis*, this study). Only GH families with at least one termite species exhibiting >0.25% enrichment or contractions are shown (Table S33 gives the full profiles). Cluster analyses showed that the two fungus-growing termite species were more similar to each other in GH composition than to other non-farming termites (non-parametric p-value=0.03 after 10,000 Monte Carlo permutations; see SI Appendix for details). **(B)** GH families enriched (red) or contracted (blue) by >0.5% in the *Termitomyces* fungal symbiont relative to 99 fungi (62); see Table S30. GH families connected with lines were enriched in *Termitomyces* and contracted in the *M. natalensis* worker gut microbiota or *vice versa*.

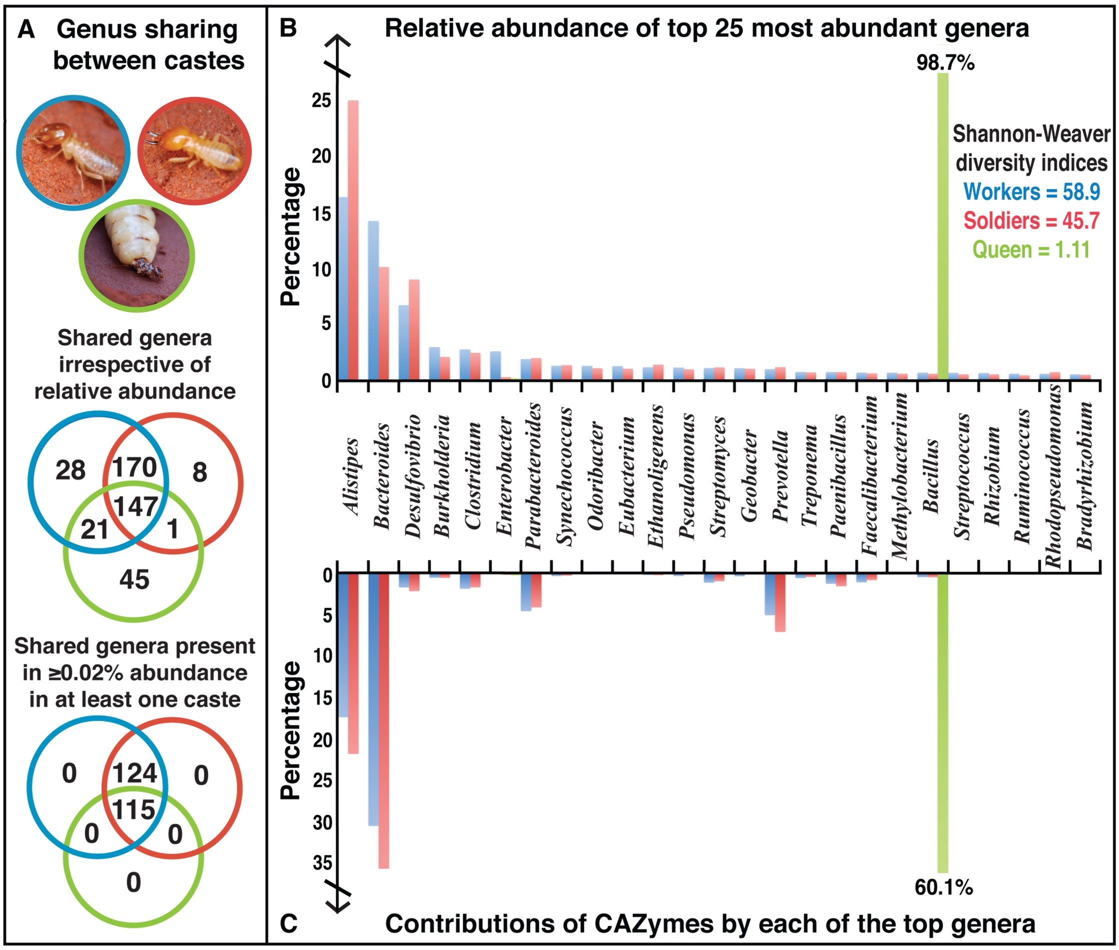


Figure 4. Diversity, distribution and CAZy potential of gut microbiotas from workers (blue), soldiers (red), and the queen (green). **(A)** Venn diagrams of the number of genera shared between the three gut metagenomes, identified using a combination of PhymmBL and BLASTn (see Methods and SI Appendix for details). The top diagram used all genera irrespective of their relative abundance within gut communities, while the bottom diagram represents a similar analysis using only genera for which at least one of the castes had ≥0.02% relative abundance, showing that none of the hits unique to only one or two castes were abundant. **(b)** The percentage of paired reads for each of the 25 most abundant bacterial genera, comprising a major portion of the total number of paired reads in workers (65.4%), soldiers (68.1%), and the queen (99.1%). Workers and soldiers shared the dominant genera *Alistipes*, *Bacteroides*, *Desulfovibrio*, *Burkholderia*, and *Clostridium*, and had relatively even distributions of reads across genera, as illustrated by similar Shannon-Weaver diversity indices. In contrast, the queen microbiota was extremely skewed towards a single dominant genus (*Bacillus*), resulting in a diversity index of only 1.11. **(B)** The percentage of CAZymes identified to originate from the 25 most abundant genera, corresponding to 68.6% of all identified CAZymes in workers, 79.2% in soldiers, and 60.4% in the queen (Table S33), showing that the majority of CAZymes originated from dominant bacterial genera for all three castes.