

Supplementary tables, figures and information

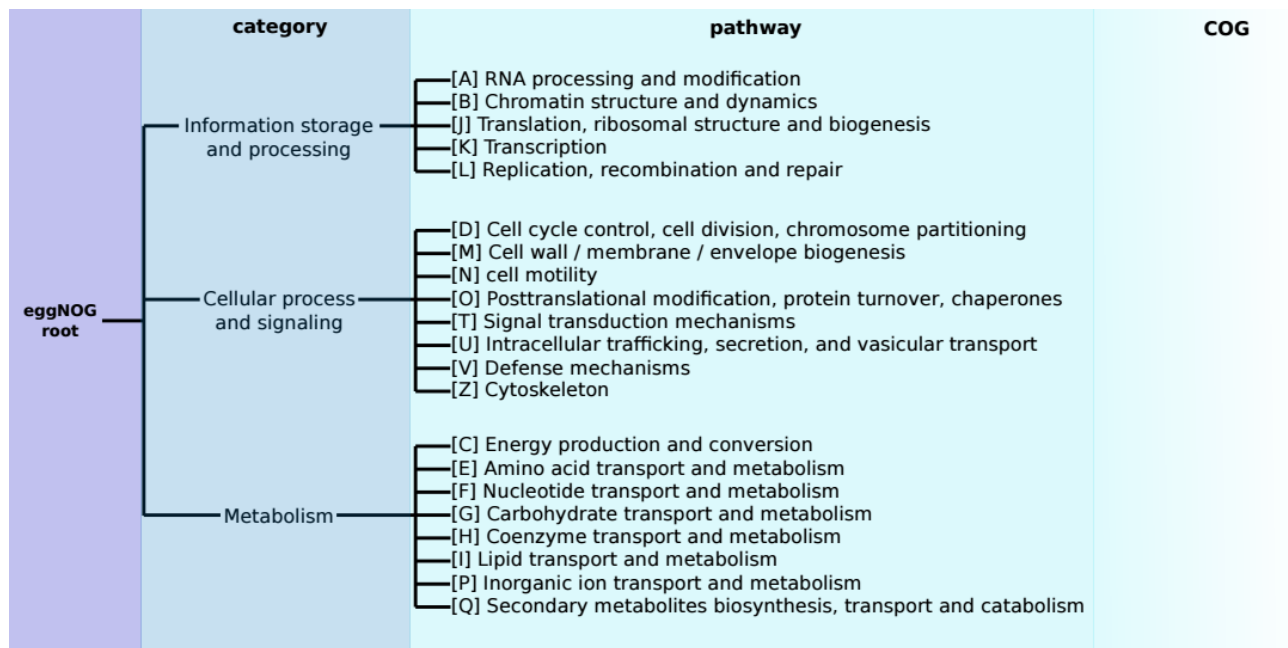


Figure S2: evolutionary genealogy of genes - non-supervised orthologous groups (eggNOG) annotation structure. The eggNOG annotation [1] is divided into three categories and 21 pathways into which the clusters of orthologous groups of genes (COGs) are grouped.

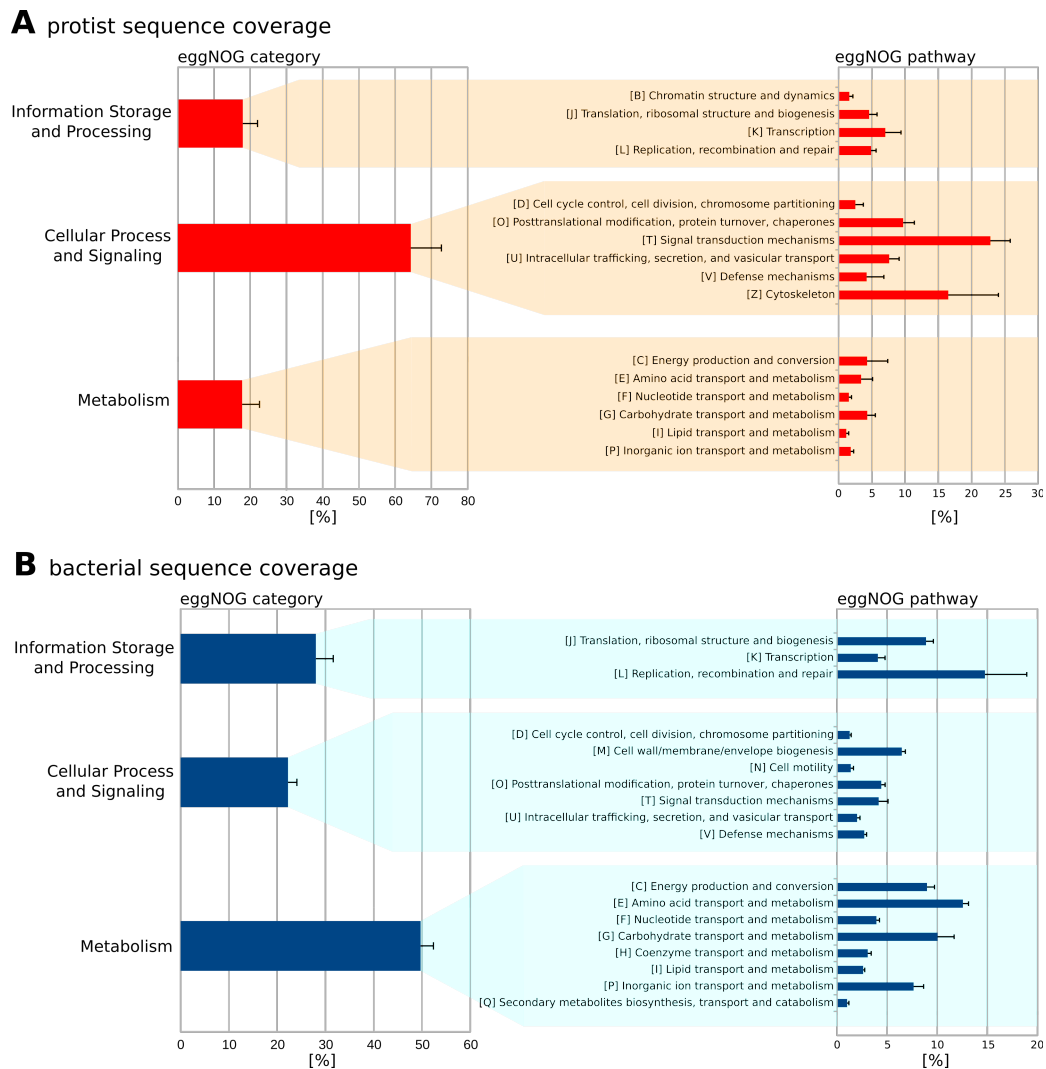


Figure S3: Metagenome sequence coverage of functional categories. A) protist sequences. B) bacterial sequences. Pathways with less than 1% coverage were removed for clarity. Background highlighting represents grouping of pathways into categories according to the eggNOG hierarchy (Figure S2). More than half (64.3%) of the protist metagenome were assigned to “cellular process and signaling”. In the bacterial metagenome sequences with annotations related to metabolic functions comprised the major group (49.7%). Error bars represent standard deviation across all samples.

Functional potential of protist and bacterial metagenomes

The largest group of functions encoded by the protist metagenome were related to signaling (Figure S3A). In particular, ABC-transporters were abundant. ABC-transporters can transport a large variety of substrates in and out of the cell [2]. In parasitic protists,

ABC transporters are essential for drug resistance and nutrient salvation [3]. They also play an important role in several insect-microbe-symbioses, in particular, when metabolic pathways are split between partners and intermediate metabolites have to be shuttled between organisms [4]. Therefore, we speculate here that shuttling of molecules is important to protists in the termite gut because they occupy a mediator position between their intracellular endosymbiotic bacteria and the termite host. Metabolites and other molecules cannot be directly exchanged between termites and intracellular bacteria, instead transport through the protist is required. This in return requires transmembrane transporters and might explain their abundance.

In contrast to the prevalence of transporters in protists, the largest fraction of bacterial sequences had metabolic functions (Figure S3B). The most common metabolic category was 'amino acid metabolism' including nitrogen metabolism. The microbiome plays a central role in nitrogen uptake and recycling because the host's primary food source, wood, is poor in nitrogen [5–7]. The second most common pathway was carbohydrate metabolism. In this pathway, we identified a total of 99 glycoside hydrolases from 31 different families (Table S12). These are the primary enzymes responsible for lignocellulose degradation. Thus, our results support a growing body of evidence that bacteria play a fundamental role in lignocellulose degradation that was formerly mainly attributed to protists in lower termites [8–11].

Protist functions

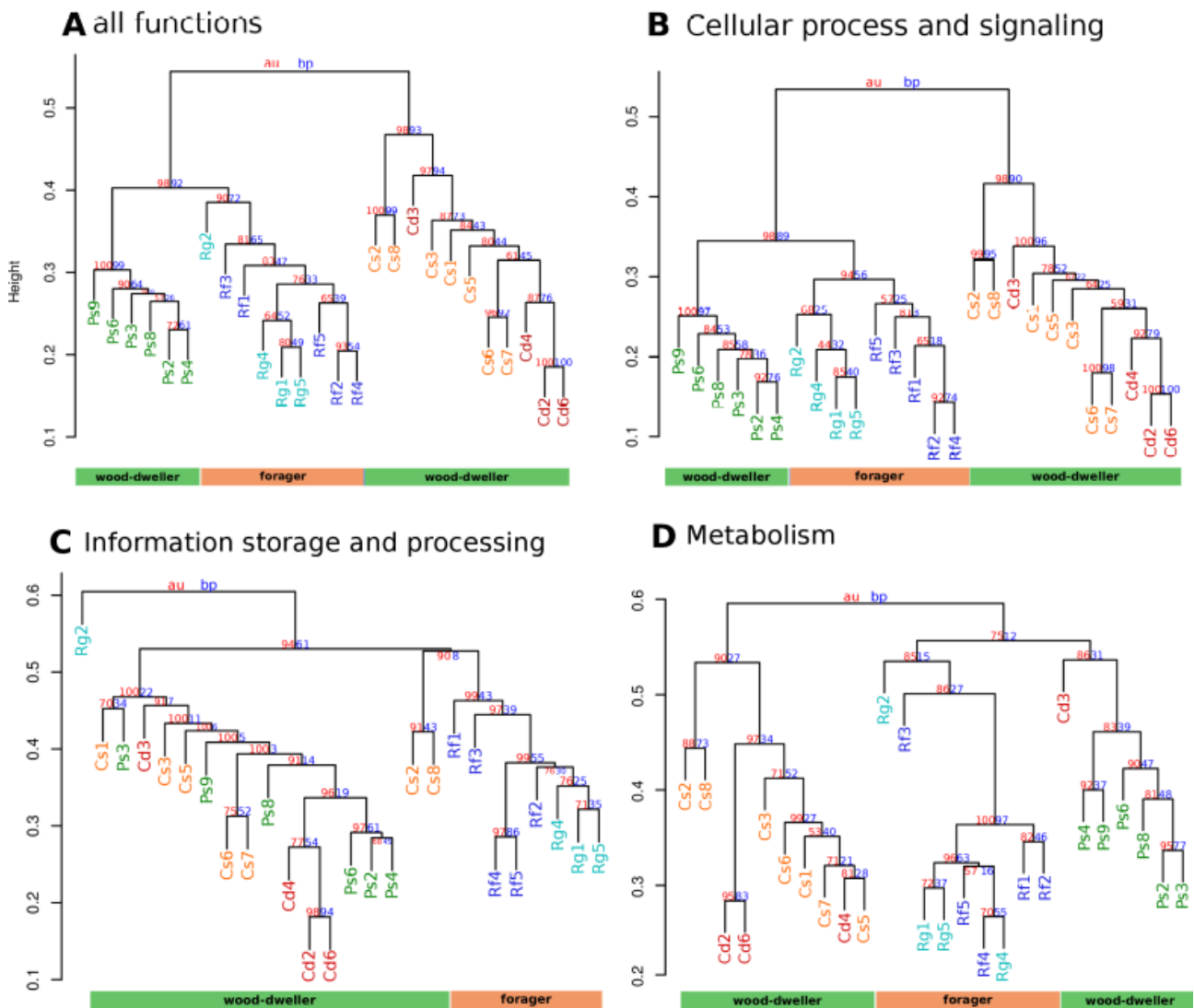


Figure S4: Cluster dendrograms of the functional profiles of the protist community.

Community distances are based on Bray-Curtis Dissimilarities. A) all functions (25,795 sequences). B) category "cellular process and signaling" (17,098 sequences). C) category "information storage and processing" (4,527 sequences). D) category "metabolism" (4,498 sequences). Cd (red) *C. domesticus* colonies; Cs (orange) *C. secundus* colonies; Ps (green) *P. simplex* colonies; Rf (blue) *R. flavipes* colonies; Rg (lightblue) *R. grassei* colonies. All functional profiles show a strong phylogenetic imprint. However, functions involved in "information storage and processing" were more similar between hosts of the same ecological life type.

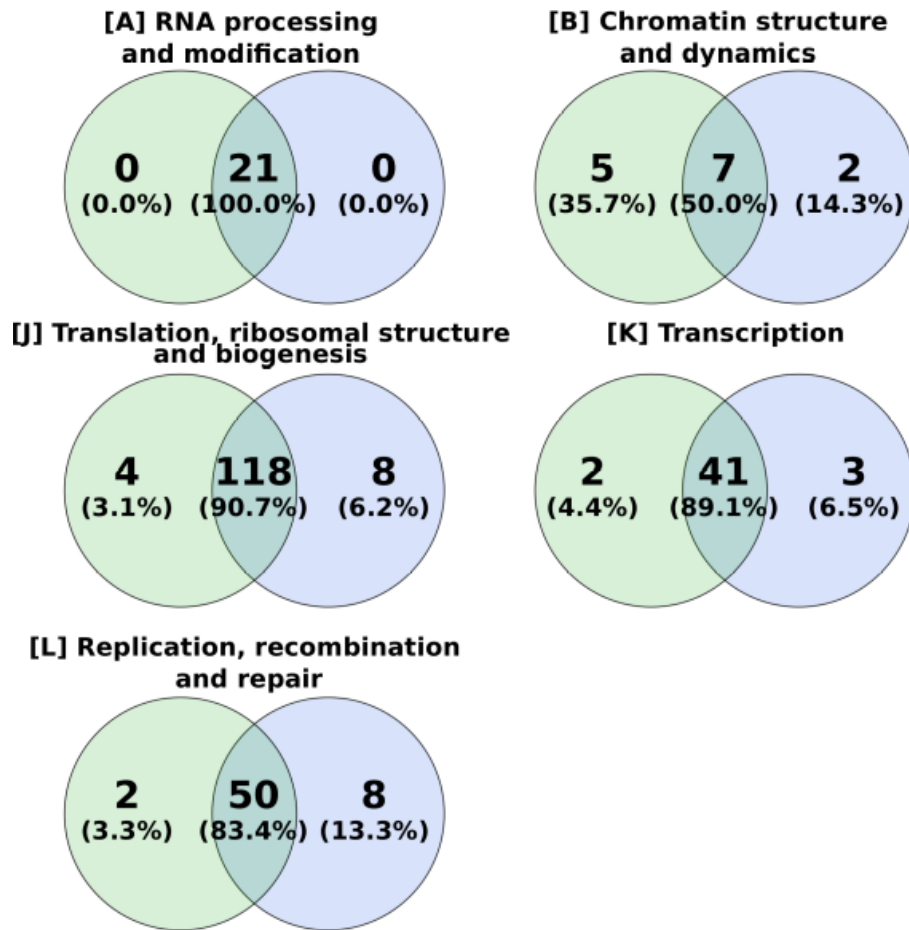


Figure S6: Venn diagram of COGs in the pathways of the eggNOG category “information storage and processing” in the protist metagenoms. Shown are numbers and percentages of over-represented and shared COGs found in the protist metagenomes of wood-dwelling (green) and foraging (blue) termite species.

Bacterial functions

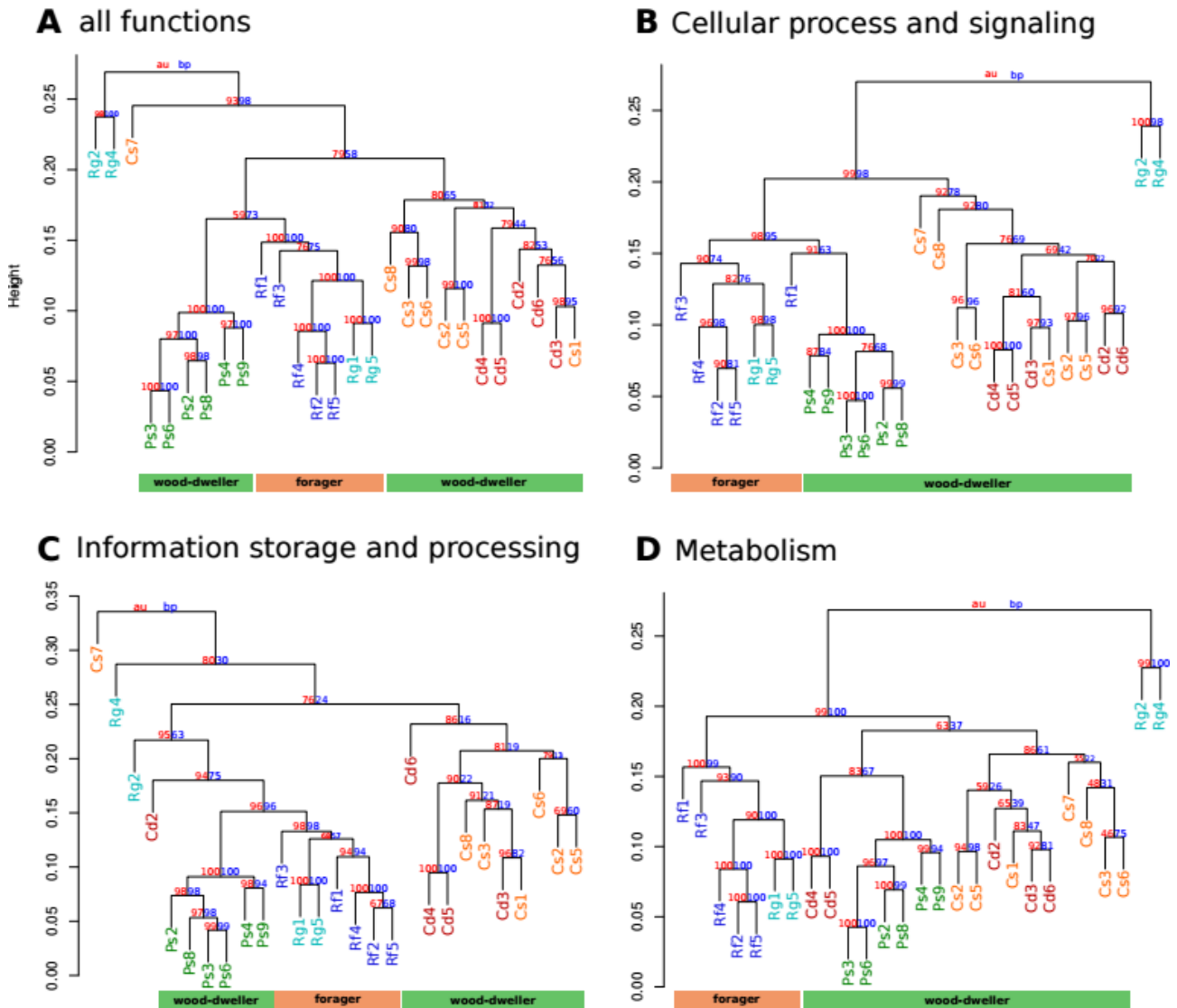


Figure S7: Cluster dendrograms of the functional profiles of the bacterial community. Community distances are based on Bray-Curtis Dissimilarities. A) all functions (21,215,480 sequences). B) category “cellular process and signaling” (4,742,380 sequences). C) category “information storage and processing” (5,954,188 sequences). D) category “metabolism” (10,586,058 sequences). Cd (red) *C. domesticus* colonies; Cs (orange) *C. secundus* colonies; Ps (green) *P. simplex* colonies; Rf (blue) *R. flavipes* colonies; Rg (lightblue) *R. grassei* colonies. Similar to the protist set, bacterial functional metagenome showed a strong phylogenetic imprint. However, the metabolic metagenomes clustered according to host ecology (life type).

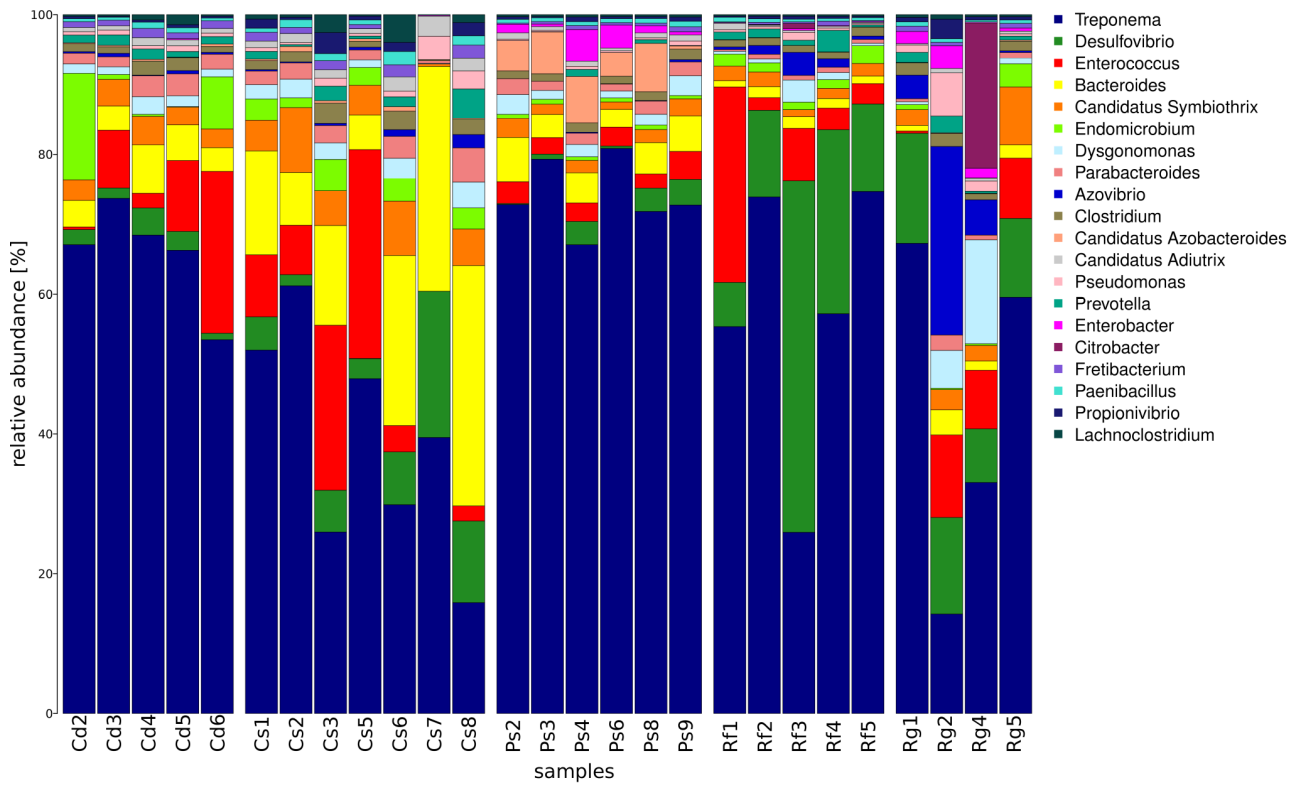


Figure S8: Frequency of bacterial genera. Cd = colony replicates of *C. domesticus*; Cs = colony replicates of *C. secundus*; Ps = colony replicates of *P. simplex*; Rf = colony replicates of *R. flavipes*; Rg = colony replicates of *R. grassei*. Shown are only the 20 bacterial genera with the highest sequence coverage. *Treponema* was the most abundant genus (14.2% (Rg2) – 80.9% (Ps6)), followed by *Desulfovibrio* (0.1% (Ps2) - 50.3% (Rf3)), *Enterococcus* (0% (Cs7) – 29.9% (Cs5)) and *Bacteroides* (0.8% (Rf1) – 34.4% (Cs8)). Like already observed in a previous study with the same samples [12], Rg2 and Rg4 showed an unusual community structure.

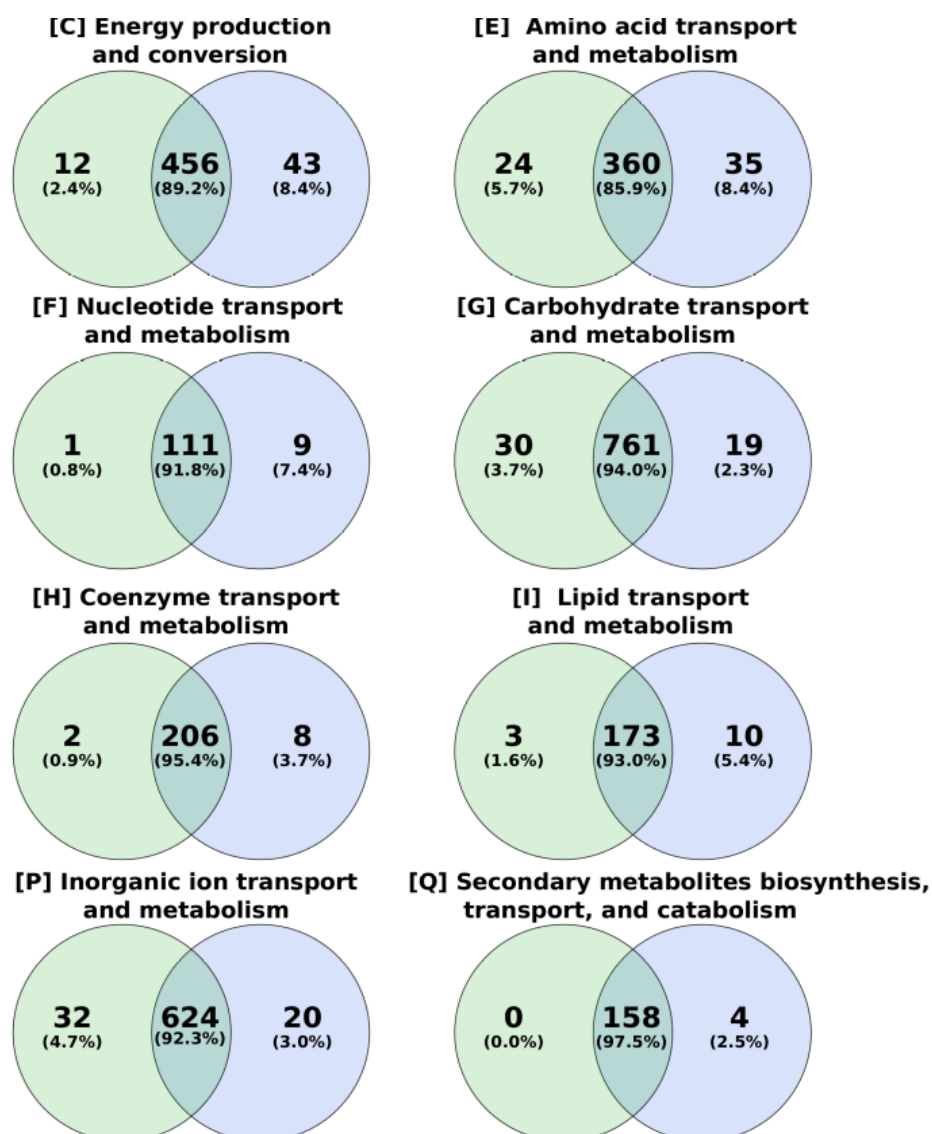


Figure S11: Venn diagram of COGs in the pathways of the eggNOG category “metabolism” in the bacterial metagenoms. Shown are numbers and percentages of over-represented and shared COGs found in the bacterial metagenomes of wood-dwelling (green) and foraging (blue) termite species.

Supplement section S13. Supplementary Material and Methods

DNA extraction

Entire guts of three workers per colony were extracted and immediately transferred to 100 μ l of CTAB solution (0.75 M NaCl, 0.05 M Tris, 0.01 M EDTA, 2% CTAB). Guts were homogenized (zirconia and glass beads, 3 min at 25*1/s), using a Tissue Lyser II (Qiagen). Additional 400 μ l of CTAB solution were added and samples were incubated for 1h at 65°C and 800 rpm on a Thermomixer Comfort (Eppendorf). After addition of 2 μ l Proteinase K (Thermoscientific, concentration: ~20 mg/mL) samples were incubated for 2h at 55°C and 800 rpm. The proteinase K reaction was terminated by heating to 98°C for 15 min. DNA was extracted with 500 μ l of chloroform:isoamyl alcohol (24:1) followed by centrifugation for 15 min at 10000 rpm. DNA, was precipitated with 325 μ l of ice cold isopropanol and overnight incubation at -20°C. Samples were washed with 300 μ l of 100% ethanol and twice with 300 μ l 70% ethanol, centrifuging for 15 min at 4°C at 14.000 rpm between each washing step. The DNA pellet was air dried and resuspended in 50 μ l of water.

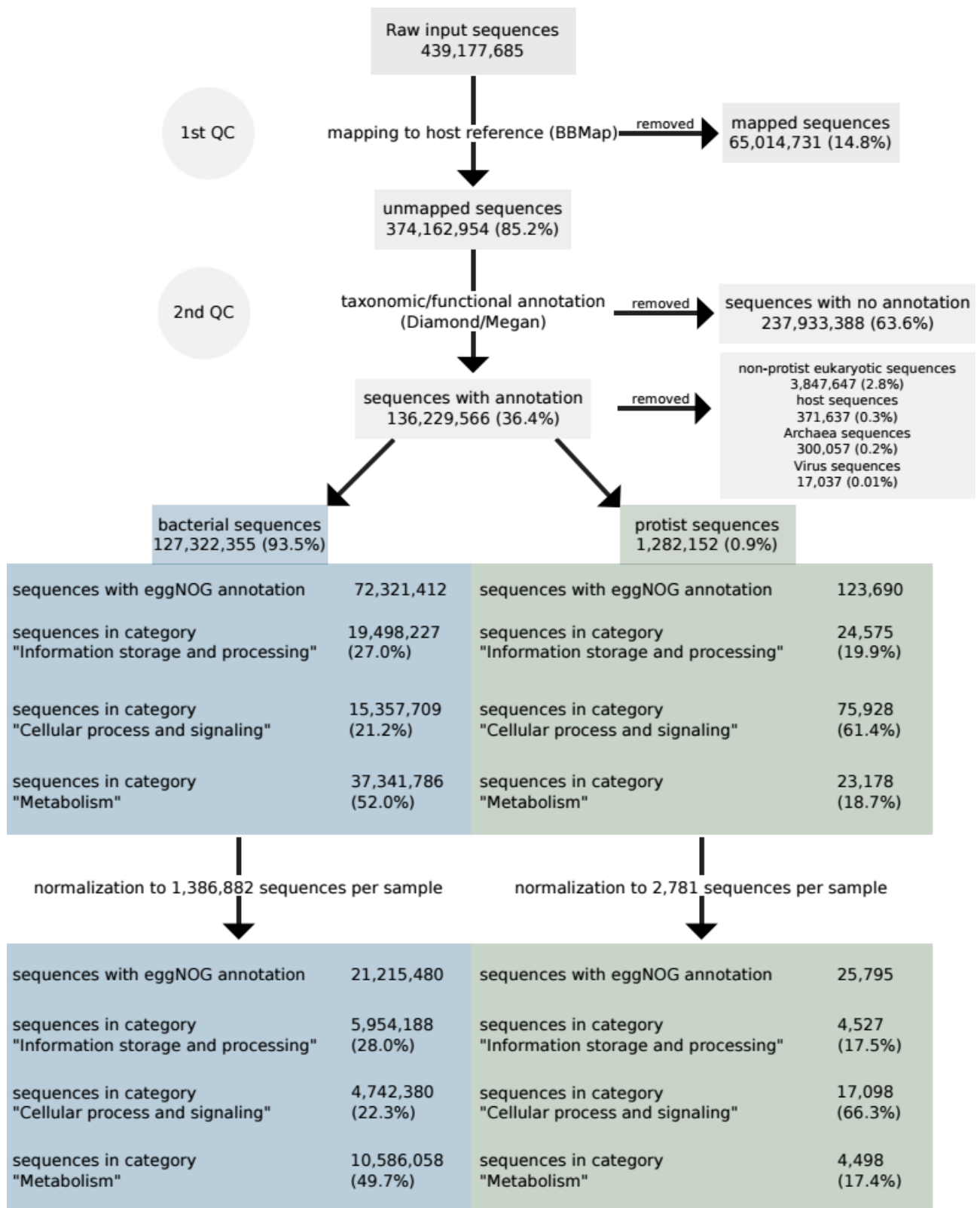


Figure S14: Analysis workflow. Of note, sequences were not assembled, since resulting contigs were not significantly longer than individual reads, likely due to the complexity of the microbiomes (data not shown). As a first filtering step, raw sequences were mapped against a host reference, using BBMap (version 37.02, [13]), with standard settings. C.

secundus and *C. domesticus* samples were mapped against the *C. secundus* [14] genome. *P. simplex* samples were mapped against the *P. simplex* transcriptome [15], *R. flavipes* and *R. grassei* samples against the *R. santonensis* transcriptome (provided by the 1KITE consortium (www.1kite.org, BioSample SAMN04005235 (for details on sequencing and removal of non-termite contaminants and cross-contaminants from the 1kite data, see [16])). Only sequences that did not match the host reference (“unmapped sequences”) were used for further analysis. Taxonomic and functional annotation of unmapped sequences was performed using Diamond (version 0.8.37.99, [17]) and Megan (MEGAN Community Edition (version 6.7.18, [18])). As a second filtering step, only sequences with taxonomic classification “bacteria” or “parabasalialia” were exported into two separate files for each sample. Therefore, contaminants such as left over host sequences, archaea, or virus sequences were removed from the datasets. In order to compare the bacterial and protist metagenomes of all samples, they were normalized to 1,386,882 and 2,781 sequences per sample, respectively. See also Supplementary section 15 below for detailed analysis steps.

Supplement section S15. Metagenomic Shotgun Data Analysis Workflow

All computing steps in this script were performed on a High Performance Computing Cluster (bwForCluster BinAC, Eberhard Karls University of Tuebingen, High Performance and Cloud Computing Group at the Zentrum fuer Datenverarbeitung of the University of Tuebingen, the state of Baden-Wuerttemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 37/935-1 FUGG).

Please note that the sequences were not assembled (see Material and Methods main text), but directly annotated using the following workflow.

1.) Mapping raw input sequences against reference genomes with BMap (version 37.02, [13]).

Sequences of *Cryptotermes secundus* and *Cryptotermes domesticus* were mapped against the *C. secundus* genome [14], sequences of *Prorhinotermes simplex* against the *P. simplex* transcriptome ([15], NCBI Bioproject ID: 219597, Assembly version GASE02000000) and sequences of *Reticulitermes flavipes* and *Reticulitermes grassei* against the *R. santonensis* (syn. *R. flavipes* [19]) transcriptome provided by the 1KITE consortium (www.1kite.org, BioSample SAMN04005235 (for details on sequencing and removal of non-termite contaminants and cross-contaminants from the 1kite data, see [16])).

```
>module load devel/java_jdk/1.8.0u112 #load required Java
environment
>cd path/to/directory #change to working directory
>/path/to/software/bbmap.sh
ref=reference_genome_or_transcriptome.fa
in1=Sample1_Read1.fastq.gz in2=Sample1_Read2.fastq.gz
outm1=Sample1_Read1_mapped.fastq.gz
outm2=Sample1_Read2_mapped.fastq.gz
outu1=Sample_Read1_unmapped.fastq.gz
outu2=Sample1_Read2_unmapped.fastq.gz #run BMap with standard
settings, do for all samples
```

This will produce four output files: Sample_Read1_mapped.fastq.gz, Sample_Read1_unmapped.fastq.gz, Sample_Read2_mapped.fastq.gz, Sample_Read2_unmapped.fastq.gz. Only “unmapped” sequences, that did not match the host reference, were used for further analysis.

2.) SYNTAX for Diamond and Megan6

This script is following the Syntax tutorial as provided by the BinAC HPC. For Diamond (version 0.8.37.99, [17]), the non-redundant NCBI database was used as reference, for

functional classification with Megan6 (MEGAN Community Edition (version 6.7.18, [18]), the mapping files prot_acc2tax-May2017.abin.gz, acc2eggnog-Oct2016X.abin.gz, acc2interpro-Nov2016XX.abin and acc2seed-May2015XX.abin (download details below) were used.

```
>wget ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz #download NCBI
nr.gz database
> diamond makedb --in nr.gz --db nr #build Diamond index for the
nr database, will generate a new file called nr.dmnd
```

The mapping files for Megan6 were downloaded from the Megan6 download page (<http://ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html>)

```
>mkdir path/to/wanted/directory/00fastq #put unmapped BMap output
files in here
>mkdir path/to/wanted/directory/10daa #will contain daa files
generated by Diamond
>mkdir path/to/wanted/directory/20rma #will contain rma files
generated by Megan6
```

For each file fastq.gz in the 00fastq directory, run Diamond as follows:

```
>module load bio/diamond/0.8.37 #load diamond environment
>module load bio/megan/6.7.18 #load Megan6 environment
>cd path/to/working/directory #change to working directory
>diamond blastx --query 00fastq/Sample1_Read1_unmapped.fastq.gz
--db nr.dmnd --daa 10daa/Sample1_Read1_unmapped.daa #run Diamond,
will generate a daa Diamond output file in the /10daa directory
```

For each daa file in your 10daa directory, run Megan's daa2rma as follows:

```
>module load devel/java_jdk/1.8.0u112 #load required Java
environment
>module load bio/diamond/0.8.37 #load Diamond environment
>module load bio/megan/6.7.18 #load Megan6 environment
>cd path/to/working/directory # change to working directory
>daa2rma -p -ps 1 -i 10daa/Sample1_Read1_unmapped.daa
10daa/Sample1_Read2_unmapped.daa -o 20rma/Sample1_p_unmapped.rma
-a2t prot_acc2tax-May2017.abin -a2eggnog acc2eggnog-Oct2016X.abin
-a2seed acc2seed-May2015XX.abin -a2interpro2go acc2interpro-
Nov2016XX.abin -fun EGGNOG SEED INTERPRO2GO #the -p function is
used for a paired read input and will generate a merged rma6
output file in the /20rma directory. This file can be used as
input for the Megan6 graphical user interface.
```

Further analysis using the Megan6 graphical user interface was performed locally. Each rma6 file of each sample was imported in Megan6 individually. Then bacterial and protist reads were extracted into separate files using the "Extract reads.." function in Megan6 (see

also Material and Method section). To compare the bacterial and protist datasets of all samples, they were imported into Megan6 using the “Compare...” function. During this step, the bacterial and protist datasets were normalized to 1,386,882 and 2,781 reads respectively.

3.) PvClust Cluster Dendrograms

Cluster dendrograms were generated in Rstudio (version 3.3.1, [20]) using the PvClust package and an external script (https://github.com/hallamlab/mp_tutorial/blob/master/taxonomic_analysis/code/mp_tutorial_taxonomic_analysis.R) to add the Bray-Curtis dissimilarity index. Functional abundance tables were extracted from Megan6 using the “Export...” function.

In Rstudio, execute:

```
>library(pvclust)
>source('path/to/working/directory/pvclust_bcdist.R') #source to
downloaded R script
>setwd("/path/to/working/directory") #set working directory
>read.csv("FunctionalAbundance.csv",header=TRUE,sep=",",row.names
= 1)->dat #read in input
>clust_dat <- pvclust(dat, method.hclust="average",
method.dist="bray-curtis", n=1000) #use pvclust
>plot(clust_dat,cex=0.8) #plot
>pdf("clust_dat.pdf") #export as pdf
>plot(clust_dat,cex=0.8)
>dev.off()
```

4.) RDA and model selection

Functional abundance tables were exported from Megan6 using “Export...” > “Text (CSV) format...” > “eggnogPath_to_count”. Bacterial functional abundances were rarefied to at least 1,000 sequences per COG, protist functional abundances to at least 10 sequences per COG. Functional abundances were subsequently transformed using Hellinger transformation. RDA and ordistep/ordiR2step for model selection were performed with the vegan package in R [21] and compared with ANOVA using the following commands:

```
#set Nullmodel
```

```

>rda0 <- rda(functional_abundance_table ~ 1, data = metadata)
#set model with 1 explanatory variable
>rda2 <- rda(functional_abundance_table ~ host_family, data =
metadata)
>rda3 <- rda(functional_abundance_table ~ host_lifetype ,data =
metadata)
#set model with both explanatory variables for modelselection
>rda1 <- rda(functional_abundance_table ~ host_family +
host_lifetype ,data = metadata)
#model selection via ordistep
>model_selection <- ordistep(rda0, scope = formula(rda1))
#model selection via ordiR2step
model_selection_inf <- ordiR2step(rda0_inf, scope =
formula(rda1_inf), trace = TRUE, permutations = how(nperm = 499),
Pin = 0.05, R2scope = FALSE)
#comparison against Nullmodel
>anova.cca(rda0, rda2)
>anova.cca(rda0, rda3)

```

Removing outlier samples from the datasets (outlier samples were Rg2, Cs2 and Cs8 in the category “information storage and processing” of the protist functional set and Rg2, Rg4 in the category “metabolism” in the bacterial set, and Rg2, Rg4 and Cs7 in all bacterial functions) did not change significant results.

5.) LEfSe

Functional abundance tables were exported from Megan6 using the “Export...” > “Text (CSV) format...” > “eggnogPath_to_count”. To use this file as input for LEfse [22], a row with the “class” “wooddweller” or “forager” was added above the sample names.

```

>python format_input.py FunctionalAbundance.csv
FunctionalAbundance.lefse.in -c 1 -s -1 -u 2 -o 1000000 #will
create the .in file
>python run_lefse.py FunctionalAbundance.lefse.in
FunctionalAbundance.lefse.res #will create the .res file

```

6.) Circular dendrogram of LEfse results using GraPhlAn [23]

Only significant overrepresented fuctions (LDA > 2.0, q-value < 0.05) were used as input.

```

>python export2graphlan.py -i FunctionalAbundance.csv -o
FunctionalAbundance.lefse.res -t tree.txt -a annot-txt --title

```

```
"Functional Abundance" --external_annotations 4 --fname_row 0
--skip_rows 1
>python graphlan_annotate.py --annot test_annot.txt test_tree.txt
test_outtree.txt
>python graphlan.py --dpi 150 test_outtree.txt test_outimg.png
--external_legends
```

7.) CAZy pathway analysis.

Full CAZy reference database was downloaded from
<http://csbl.bmb.uga.edu/dbCAN/index.php> (newest version CAZyDB.07202017.fa).

Bacterial reads were blasted against the reference using Diamond:

```
>diamond makedb --in CAZyDB.07202017.fa -d cazy_full
>diamond blastx -d cazy_full -q bacteria_sampleXY.fasta -o
matches.m8 -k 1 -e 0.00001 --max-hsps 1
```

Read matches for GHs of interest* were counted by piping the wc -l function after a grep search command.

*Read matches for GHs of interest: cellulose:
endo- β -1,4-glucanase (cellulase): GH5,8
cellobiohydrolase: GH6,7,9,48,74
 β -glucosidase: GH1,3,4,5,9,13,17,30,31,63,65,97,116,122,133
hemicellulose (see [24]):
xylan:
endo- β -1,4-xylanase: GH5,8,10,11,43
exo- β -xylosidase: GH3,39,43,52,54
mannan:
endo- β -1,4-mannanase: GH5,26
exo- β -1,4-mannosidase: GH1,2,5
arabinofuranosyl containing hemicellulose:
alpha-L-arabinofuranosidase: GH3,43,51,54,62
endo-alpha-1,5-arabinanase: GH43

Supplementary Literature

1. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 2016.
2. Rees DC, Johnson E, Lewinson O. ABC transporters: The power to change. *Nat Rev Mol Cell Biol.* 2009;10:218–27.
3. Dean P, Major P, Nakjang S, Hirt RP, Embley TM. Transport proteins of parasitic protists and their role in nutrient salvage. *Front Plant Sci.* 2014;5:153.
4. Peterson BF, Scharf ME. Metatranscriptome analysis reveals bacterial symbiont contributions to lower termite physiology and potential immune functions. *BMC Genomics.* 2016;17. doi:10.1186/s12864-016-3126-z.
5. Potrikus CJ, Breznak JA. Gut bacteria recycle uric acid nitrogen in termites: A strategy for nutrient conservation. *Proc Natl Acad Sci U S A.* 1981;78:4601–5.
6. Burnum KE, Callister SJ, Nicora CD, Purvine SO, Hugenholtz P, Warnecke F, et al. Proteome insights into the symbiotic relationship between a captive colony of *Nasutitermes corniger* and its hindgut microbiome. *ISME J.* 2011;5:161–4.
7. Brune A, Dietrich C. The gut microbiota of termites: Digesting the diversity in the light of ecology and evolution. *Annu Rev Microbiol.* 2015;69:145–66.
8. Tokuda G, Watanabe H. Hidden cellulases in termites: revision of an old hypothesis. *Biology letters.* 2007;3:336–9.
9. Do TH, Nguyen TT, Nguyen TN, Le QG, Nguyen C, Kimura K, et al. Mining biomass-degrading genes through Illumina-based de novo sequencing and metagenomic analysis

of free-living bacteria in the gut of the lower termite *Coptotermes gestroi* harvested in Vietnam. *J Biosci Bioeng.* 2014;118:665–71.

10. Peterson BF, Stewart HL, Scharf ME. Quantification of symbiotic contributions to lower termite lignocellulose digestion using antimicrobial treatments. *Insect Biochem Mol Biol.* 2015;59:80–8.

11. Yuki M, Kuwahara H, Shintani M, Izawa K, Sato T, Starns D, et al. Dominant ectosymbiotic bacteria of cellulolytic protists in the termite gut also have the potential to digest lignocellulose. *Environ Microbiol.* 2015;17:4942–53.

12. Waidele L, Korb J, Voolstra CR, Künzel S, Dedeine F, Staubach F. Differential ecological specificity of protist and bacterial microbiomes across a set of termite species. *Front Microbiol.* 2017;8. doi:10.3389/fmicb.2017.02518.

13. Bushnell B. BBMap. sourceforge.net/projects/bbmap/.

14. Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, et al. Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat Ecol Evol.* 2018;2:557.

15. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science.* 2014;346:763–7.

16. Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, et al. Evolutionary history of the Hymenoptera. *Curr Biol.* 2017;27:1013–8.

17. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12:59–60.

18. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol.* 2016;12:e1004957.
19. Austin J, Szalanski AL, Scheffrahn RH, Messenger MT, Dronnet P, Bagneres AG. Genetic evidence for the synonymy of two *Reticulitermes* Species: *Reticulitermes flavipes* and *Reticulitermes santonensis*. *Ann Entomol Soc Am.* 2005;98:395–401.
20. R Core Team. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL [http://wwwR-project.org/](http://www.R-project.org/). 2014;:R Foundation for Statistical Computing.
21. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. vegan: community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. Version 2.4-3. URL <https://CRAN.R-project.org/package=vegan>. 2017.
22. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011;12:R60.
23. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 2015;3:e1029.
24. Shallom D, Shoham Y. Microbial hemicellulases. *Curr Opin Microbiol.* 2003;6:219–28.