Termite diet rather than geographical origin determines the microbiome composition and functional genetic structure of nests from South American and African representatives, as revealed by a multiomics approach

Juan José González Plaza1,a,* , Jaromír Hradecký2, Jan Šobotník3
1 Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic
2 Extemit Department, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic
3 Faculty of Tropical Agriscience, Czech University of Life Sciences Prague, Prague, Czech Republic

a Current affiliation: International Research Centre in Critical Raw Materials-ICCRAM, University of Burgos, Plaza Misael Bañuelos García s/n, 09001 Burgos, Spain

*Author to whom correspondence must be addressed: gonzalezplazajuanjose@gmail.com

Keywords:
Metabolomics, amplicon sequencing, metagenomic sequencing, termite nest microbiome, feeding habits

Abstract
Termites represent one of the most important insect groups worldwide due to their key role as plant decomposers and proxy of carbon recycling in the tropical rainforest ecosystems. Besides, high relevance in research has been given to these social insects due to a prominent role as urban pests. However, one of the most fascinating aspects of termites are their defence strategies that prevent the growth of detrimental microbiological strains on their nests. One success factor is the key role of the nest allied microbiome. Understanding how beneficial microbial strains aid termites in pathogen biocontrol strategies could provide us with an enhanced repertoire for fighting antimicrobial resistant strains or mine for genes for bioremediation purposes.

We carried out a multiomics approach for dissecting the nest microbiome in a wide range of termite species, covering several feeding habits and three geographical locations at two tropical sides of the Atlantic Ocean, and an African savanna. Our experimental approach included untargeted volatile metabolomics, targeted evaluation of volatile naphthalene, taxonomical profile for bacteria and fungi through amplicon sequencing, and further dive into the genetic repertoire through a metagenomic sequencing approach. Volatile naphthalene was present in species belonging to the genera Nasutitermes and Cubitermes. We further assessed the apparent differences in terms of bacterial community structure, having found a stronger influence from feeding habits and genera, rather than the geographical location. Lastly, our metagenomic analysis revealed that the gene content provides both soil feeding genera with similar functional profiles, while the wood feeding genus shows a different one. These results seem to be independent of the geographical location, indicating that the nest functional profile is heavily influenced by the diet of the termite inhabiting, building, and maintaining the nest.
1 Introduction

Termites represent one of the most successful stories of adaptation in the evolution of animal kingdom. One of the main ingredients for biological success is the effective use and exploitation of a specific niche and its resources, what allows for the transference of genetic information to the next generations. Termites have adapted to use plant dead material in order to obtain their energetic resources. In the process of doing so, they have achieved a key role in the recycling of carbon in one of the most important ecosystems on Earth, the tropical rainforest. The use of lignocellulosic sources depends largely on the symbiotic relationships with a wide range of symbiotic microorganisms (Brune, 1998; Ohkuma and Brune, 2010). Not only interact termites with these microorganisms for energetic purposes, but are indeed deeply related with them because they live on their surface or at their nest material. Nests are the Achilles heel of termites, since their whole social structure turns around their central place of existence. That is so because termites live in populated colonies with a series of intricated galleries, were the conditions of humidity (Wiltz et al., 1998) and temperature create an ideal setting for the growth of pathogenic organisms, such as fungi. Not exclusively because of these two, but also due to the high content of nutrients in termite nests in comparison with the nutrient-poor surrounding soil environment (López-Hernández, 2001; van der Sande et al., 2018), yet rich in nutrient cycling (Jordan and Herrera, 2015). Among the reasons for the high content of nutrients in nests and immediate surrounding areas is the way in which nests are built. Nests are a product of a considerable amount of faeces as cementing tool (Jose et al., 1989; Chouvenc et al., 2018), and soil particles that are transported from the surrounding environment (Mizumoto et al., 2021). Considering these limited availabilities of food substrates in the surrounding soil, the conditions are set for the competitive race for survival in the microbial moiety, with a continuous input of external strains, potentially detrimental and threatening the existence of the whole colony. These external microorganisms thrive for an environment which could limit the presence of grazing organisms (Clarholm, 1981; Rønn et al., 2002), and high availability of nutrients provided by the input of termite faeces when they maintain and build their nests (Chouvenc et al., 2013). However, termites are very successful in their strategies for limiting the growth of detrimental microorganisms. Among them, it has been appointed that fumigation with certain type of compounds, such as naphthalene, aids in controlling pathogens (Chen et al., 1998a; Wright et al., 2000).

In this context, we have carried out a research effort to better characterize the microbial communities living in termite nests among several termite species across different geographical locations. We were interested in termites with two feeding strategies, wood and soil, originating at both sides of the Atlantic, including rainforest and savanna representatives. To evaluate the presence of volatile compounds, we used a coupled untargeted gas chromatography–mass spectrometry (GC-MS) analyte identification strategy, followed by a targeted naphthalene analysis. These metabolomics approaches were complemented with a taxonomic characterization of the microbial communities existing on the same material, in order to try to understand the potential co-occurrences of certain identified metabolites with bacterial or fungal genera. Lastly, we further performed shotgun metagenomic sequencing from selected termite colonies for a deeper understanding of the genetic microbial contribution to defence and naphthalene production.

Our results showed first a much lower amount of naphthalene in termite nests than those previously reported in literature. A very interesting observation found throughout the three omics used approaches (metabolomic, amplicon sequencing, metagenomic sequencing) is that genera and especially the feeding habits determine the relationships between the different microbial communities, regardless of the geographical origin.
2 Material and methods

2.1 Sample collection and initial processing

Nests from twenty-eight termite colonies were collected during 2019 and 2020 at different sampling campaigns in South America and Africa. The first sampling site is located in French Guiana, Petit-Saut Dam area, which is located in Sinnamary river (5°03’N, -53°02.76’E), and is characterized by a hot and humid weather throughout the year with two dry (February-March, July-November) and two rainy seasons (Colas et al., 2020). In Africa we collected samples in the area of Ebogo village (Cameroon, 3°23.86’N, 11°28.19’E), located in the Mbalmayo Forest Reserve. This area is characterized by a bimodal precipitation pattern (Knoben et al., 2019), with a dense evergreen forest subjected to perturbations on approximately a third of the territory due to agricultural practices (Mey and Gore, 2021). The dry periods span from December to February, and July to August (Zapfack and Engwald, 2008). The second African sampling site was Nyika National Park (Malawi, -10°32.92’N, 33°53.47’E), characterized mostly by miombo woodland, and in minor degree by montane dambos and grasslands, where mean average temperatures are of 23 °C (Allingham and Harvey, 2013), and the dry season lasts mostly from June to October, while rainfalls occur chiefly during November-March (van Velden et al., 2020). Two of the samples originate in the breeds at Faculty of Forestry and Wood Sciences (Czech University of Life Sciences Prague, Czech Republic), maintained in humid constant temperatures of 28 °C.

Nest samples were processed in clean hexane-wiped metal trays under laminar flow hood to avoid contamination. Large pieces of nests devoid of termites were broken in small fragments and stored at -80 °C until processing. Samples were further homogenized in a Retsch oscillatory mill. The hexane-wiped steel 10 mL chambers were filled with nest material, closed, submerged into liquid nitrogen until stabilized. Samples contained in steel chambers were then homogenized to fine powder for 5 minutes at 30 Hz frequency, recovered, and stored at -80 °C until further analysis.

2.2 Metabolite profiling

For initial solid phase microextraction (SPME) volatiles profiling, 200 mg of previously homogenized samples were placed into different 10 mL screw-top vials for headspace analysis, and were sealed with a magnetic cap. Sample incubation and the following metabolite extraction, were performed at 50 °C in agitator of Gerstel MPS2 autosampler (Gerstel, SUI) for 10 and 30 minutes respectively. For separation and detection gas chromatograph coupled with time of flight mass spectrometer (GC-TOF-MS) (Leco Pegasus 4D, Leco, USA) was used. Temperature programmed injector was operated in splitless mode at 275 °C. For separation, a 30 m (0.25 mm i.d., 0.25 µm film thickness) Rxi-5Sil MS (Restec, USA) column was used. The temperature program for the 1D oven was as follows: 40 °C for 1 min, then ramped at a rate of 10 °C/min to 210 °C, then at 20 °C/min to 300 °C with a hold time of 3 minutes. The total GC run time was 26 min. The mass spectrometer was operated in mass range 35-500 m/z with acquisition speed of 10 Hz.

For targeted naphthalene analysis, an SPME method based on Cao (2012) and Tsimeli et al. (2008) was used, employing standard addition quantification. One hundred mg of homogenized sample were placed into 10 mL HS vial and 0.5 mL of water (UHPLC-MS purity, Supelco, USA) was added before sealing with magnetic screwcap. Four vials per sample were prepared. Thru septum, to first two vials, 10 µL of methanol (HPLC Plus purity, Sigma-Aldrich, USA) were injected, while in resting two, 10 µL of methanol containing different amount of naphthalene were injected. Sample was heated under agitation at 50 °C for 10 minutes and then volatiles from headspace were collected onto a SPME fibre for 5 minutes.

Linear response of naphthalene was observed, when blank nest sample was spiked on levels from 1-1100 ng/g. Method accuracy and reproducibility were checked by measuring a certified reference material (CRM) of polyaromatic hydrocarbons in soil (CRM170, Lot: LRAC8900, Sigma-Aldrich, USA). Measuring six replications of CRM, relative standard deviation (RSD) of entire determination was 12%, resulting in uncertainty (U=2xRSD) of 25%. Comparing obtained naphthalene content of 563 ng/g ± 140 ng/g (average ±U) to certified value of 573 ± 45 ng/g showed good accuracy of the presented method. The only modification for CRM measurement was that due to high naphthalene concentration, sample weight was reduced to 10 mg per vial.

2.3 Species identification
DNA from termites belonging to each colony were isolated using the QIAGEN Blood and Tissue kit according to manufacturer’s instructions. Isolated DNA was used for amplification of the Cytochrome Oxidase II gene (COII) with the following program: 1 minute at 94 °C, 30 cycles of 94 °C for 15 sec, 61 °C for 1 minute, and 72 °C for 35 sec, and a final step of elongation at 72 °C for 10 minutes. PCR products were submitted to Microsynth AG (Switzerland) for Sanger sequencing. Results were compared to NCBI nr database for retrieving BLAST results (Altschul et al., 1990). The used primers were previously described by Benjamino and Graf (Benjamino and Graf, 2016).

2.4 Microbial DNA isolation

Powder homogenized samples were used for isolation of DNA. One hundred mg of soil were used as starting material using the NucleoSpin Soil DNA kit (Macherey-Nagel, Germany). The whole procedure was carried out under laminar flow, using sterile material, and kit components were exclusively opened and manipulated under sterile conditions. Samples were quantified by Qubit 2.0 Fluorometer (Thermo Scientific, USA) and quality checked through spectrophotometric methods. Samples were submitted to Novogene (Hong Kong, China) for sequencing of the regions 16S V34 and ITS2 in order to retrieve the taxonomic structure of the bacterial and fungal communities of the samples. DNA was subjected to quality control checks through DNA purity (optical density 260/280), agarose gel electrophoresis to evaluate carryover of contaminants, and Qubit (Thermo Scientific, USA) concentration evaluation.

2.5 Sequencing

2.5.1 16S rRNA gene sequencing

Amplicon sequencing for the 16S V4 region with primers 515F (GTGCCAGCMGCAGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) was carried out following standard company procedures (Novogene, Hong Kong, China). In brief, a 292 bp amplicon was obtained through PCR with the above-mentioned specific primers. PCR products were quantified, purified (Qiagen Gel Extraction Kit, Qiagen, Germany), and used for library preparation (NEBNext Ultra DNA Library Pre-Kit, Illumina, San Diego, CA, USA). Libraries were sequenced following a 250 bp sequencing strategy.

2.5.2 Internal Transcribed Spacer (ITS) Sequencing

ITS amplicon sequencing was carried out by Novogene (Hong Kong, China), following standard company procedures. In brief, 1 ng of DNA was used to amplify fungal ITS genes from ITS2 region with a specific set of primers ITS3 (GCATCGATGAAGACGCAGC), and ITS4 (TCCTCCGCTTATTGATATGC), with an amplicon size of 386 bp. The overall procedure with amplicons follows the one explained for 16S.

2.5.3 Metagenomic sequencing

We selected two genera, comprising four samples each, from those where naphthalene was produced (Cubitermes and Nasutitermes). For comparison purposes a third group with no naphthalene presence in nests was selected (Anoplotermes). DNA was isolated as above-mentioned, and submitted for Illumina PE150 sequencing at Novogene (Hong Kong, China). After quality control in standard agarose gels and quantity concentration estimation with Qubit 2.0 (ThermoFisher, USA), DNA was randomly sheared through sonication. Fragments were end-polished, A-tailed, and then the Illumina adapters (5’ Adapter: 5’-AGATCGGAAGAGCGGTGTAGGCTCTAAGG-3’; 3’ Adapter: 5’-GATCGGAAGAGCGGTGTAGGCTCTAAGG-3’) were ligated. A PCR amplification was carried out using P5 and P7 indexed oligonucleotides, and products purified with AMPure XP system. Size distribution of libraries was evaluated at Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA), and quantified through real-time PCR for achieving equimolar concentrations. Sequencing was subsequently carried out at Illumina instruments.

2.7 Microbiome data processing and analysis

2.6.1 Amplicon sequencing data: 16S and ITS

The assignment of paired-end reads to samples was based on their unique barcodes, afterwards truncated by cutting off the barcode and primer sequence. Overlaps of paired-end reads served for merging reads using
FLASH (Magoč and Salzberg, 2011). Reads were quality filtered to obtain high-quality clean tags (Bokulich et al., 2013) using QIIME (V1.7.0). Reads were then compared to a reference database (Gold DB, http://drive5.com/uchime/uchime_download.html) with the UCHIME algorithm (Edgar et al., 2011) for detection and removal of chimera sequences. Clean reads were stored in individual files in fastq format. These files were used as input in a common tool used in microbiome studies for bioinformatic treatment of clean demultiplexed reads, QIME2 (release 2021.4) (Bolyen et al., 2019). This wrap-up tool was used to import (q2-import, SingleEndFastqManifestPhred33V2) the clean demultiplexed single fastq files corresponding to our samples with the import plugin. Denoising was carried out with DADA2 (via q2-dada2-denoise single) (Callahan et al., 2016), what allows for identification of all observed amplicon sequence variants. ASVs were aligned using mafft (Katoh et al., 2002) (via q2-alignment), and used to build a phylogeny (fasttree2), via q2-phylogeny (Price et al., 2010).

After DADA2, sequences that were found in the negative control “.fastq” file were removed from the project through quality-control exclude-seqs plugins.

Core metrics were calculated with core-metrics-phylogenetic plugin, where rarefaction was performed using a sampling depth of 10,000 for 16S data, and 20,000 for ITS reads. Among metrics, we calculated Faith’s Phylogenetic Diversity (Faith, 1992), beta diversity such as UniFrac (Lozupone et al., 2007), unweighted UniFrac (Lozupone and Knight, 2005), Jaccard Distance and Bray-curtis dissimilarity. Plugins were core-metrics-phylogenetic, alpha-group-significance, and beta-group-significance, where the metadata column of choice was “geography”, “genus”, or “feeding-group”.

Taxonomy was assigned with the q2-feature-classifier classify-sklearn naïve Bayes taxonomy classifier (Bokulich et al., 2018; Kaehler et al., 2019), for the 16S we used the weighted Greengenes 13.8 99% OTUs full-length sequences (DeSantis et al., 2006), while for ITS we used silva-138-99-nb-classifier (Pruesse et al., 2012; Quast et al., 2013; Yilmaz et al., 2014). Taxonomy plots were drawn in R through the ampvis2 package, using the tutorials available at: https://sites.google.com/a/ciad.mx/bioinformatica/home/metagenomica/visualizacion/ampvis2, and https://madsalbertsen.github.io/ampvis2/articles/ampvis2.html, where we used the Heillinger transformation in order to represent the PCA (Legendre and Gallagher, 2001).

Once taxonomy was assigned and frequency calculated per sample, we performed a Least Discriminant Analysis (LDA) Effect Size (LEfSe) (Segata et al., 2011). This algorithm discovers biomarkers that identify features on the tested biological conditions. A text tab separated file in metadata-columnar format was loaded at a Galaxy instance (https://huttenhower.sph.harvard.edu/galaxy/). The OTU table was filtered in the case of 16S data to include only those that satisfied frequency > 0.001 in at least one of the samples.

### 2.8 Metagenomics bioinformatics

Sequenced reads were first cleaned with trimmomatic (Bolger et al., 2014), and then subjected to a quality control process (FASTX Toolkit 0.0.14, FASTQ/A short-reads pre-processing tools), as shown in Supplementary material S1. We followed then the Kalamazoo Protocol for metagenome assembly khmer 0.8.4 (Brown et al., 2013). We used the digital normalization protocol as stated (Brown et al., 2012, 2015), and used the normalized reads as an input for SqueezeMeta v1.4.0, v. May 2021 (Tamames and Puente-Sánchez, 2019), in order to carry out a first assembly via Megahit (Li et al., 2015). We carried out that first assembly using the seqmerge mode, but did not use the statistics or any other results. After carrying out the assembly, the generated contigs were used as external assembly reference (extassembly) in SqueezeMeta, and thus, no assembly was performed at this second step. At this second round, the input consisted on trimmed, quality control passed, no spaces (in house script) files containing separately forward and reverse pairs, gzipped. The pipeline followed the usual procedure by SqueezeMeta, with the following tools:

Assembly carried out with Megahit (Li et al., 2015), contig statistics with prinseq (Schmieder and Edwards, 2011), and redundant contig removal by cd-hit (Schmieder and Edwards, 2011). Merging of contigs was carried out with Minimus2 (Treangen et al., 2011), RNA prediction with Barnnap (Seemann, 2014), taxonomy classification of 16S rRNA sequences by RDP classifier (Wang et al., 2007), tRNA/tmRNA prediction by Aragorn (Laslett and Canback, 2004), and Prodigal for Open Reading Frames (ORF) prediction (Hyatt et al., 2010). Diamond (Buchfink et al., 2014) served as the tool to search for similar patterns at GenBank (Clark et al., 2016), eggNOG (Huerta-Cepas et al., 2016), KEGG (Kanehisa and Goto, 2000). Then HMM homology
searches were performed by HMMER3 (Eddy, 2011) for the Pfam database (Finn et al., 2016). Reads were mapped against contigs with Bowtie2 (Langmead and Salzberg, 2012), and binning was done using MaxBin2 (Wu et al., 2016) and Metabat2 (Kang et al., 2019), where binning results were combined using DAS Tool (Sieber et al., 2018). MiniPath (Ye and Doak, 2009) was used against KEGG (Kanehisa and Goto, 2000) and MetaCyc (Casp et al., 2018) databases for pathway prediction.

At the end of the pipeline, tables for results processing in R were created by means of the sqm2tables.py script provided in the SqueezeMeta package. Results were loaded in R through the use of SQMTools (Puente-Sánchez et al., 2020), and charts produced according to the SQMTools manual. Some of the R packages used were: “SQMTools”, “ggplot2”, “reshape2”, “pathview”, “data.table”.

Non-metric multidimensional scaling (NMDS) was carried out in RStudio (RStudio Team, 2020), using Transcripts per Million (TPM) calculated values obtained from SqueezeMeta pipeline, for those genes with KEGG annotation, through “vegan” package. This ordination collapses the information from all genes versus the samples, to visualize the relationships between samples.

We further used diamond (Buchfink et al., 2014) in order to search for genes and patterns of interest in our final contig list with Open Reading Frames, where the databases of reference were: AnHyDeg (Callaghan and Wawrik), AromaDeg (Duarte et al., 2014), BacMet (Pal et al., 2014), bacterbase (Hammami et al., 2010), CARD (protein_homolog_model) (Alcock et al., 2020), and CAZy (Carbohydrate Active Enzymes database, http://www.cazy.org/) (Lombard et al., 2014). We used by default an e-value threshold of 0.005, an identity percentage threshold of 40, and manually curated the results discarding those hits with lower than 50 as bit-score. We also compared our assembled contigs with a plasmid database plsdb (Galata et al., 2019) through blastn (Altschul et al., 1990) with e-value threshold of $5 \times 1^{-10}$ and an identity percentage of 50.

3 Results

3.1 Identification of termite species belonging to each collected colony

Alive termites belonging to each of the collected nests used in this study were taxonomically identified through molecular methods (COII amplification and sequencing). Results are shown in Table 1.

We have classified the nests according to the termite species habitating them. Ten other colonies were classified in Group II from three locations: our laboratory breeds (BR), Cameroon (CM), and French Guiana (FG). Another group comprising ten colonies originating from French Guiana were classified as Group III. Lastly, Group IV comprised colonies originating from Cameroon, French Guiana, and Malawi (MW). Most of the species from these colonies belong to the Cubitermes genus, comprising five different species (Table 1).

Table 1. Species identification after COII method. Feeding habits of termites, their geographical origin, and classification according to termite feeding group are shown as well. Note: For metagenomic data analysis after sequencing, we have used the generic Ban1, Ban2, and Ban3 for simplicity purposes, and have the correspondence: Ban1 = Anba 2 FG III; Ban2 = Anba 3 FG III; Ban3 = Anba 4 FG III. Note: Breeds samples were collected in our termite collection at Czech University of Life Sciences, Prague, Czech Republic. N/A: Not applicable; BR: Breeds; CM: Cameroon; FG: French Guiana; MW: Malawi.

| Genus          | Species | Feeding habits | Origin | Termite feeding group | Code   | Sequenced (metagenome) | Sampling campaign | Rainfall season at sampling |
|----------------|---------|----------------|--------|-----------------------|--------|------------------------|-------------------|---------------------------|
| Coptotermes    | formosanus | Wood           | BR     | I                     | Copfor BR I |                           | July 2019         | N/A                       |
| Nasutitermes   | sp       | Wood           | BR     | II                    | Nasp BR II |                           | July 2019         | Dry                       |
| Nasutitermes   | lujae    | Wood           | CM     | II                    | Naslu 1 CM II | Yes                       | July 2019         | Dry                       |
| Nasutitermes   | lujae    | Wood           | CM     | II                    | Naslu 2 CM II | Yes                       | July 2019         | Dry                       |
| Nasutitermes   | lujae    | Wood           | CM     | II                    | Naslu 3 CM II | Yes                       | July 2019         | Dry                       |
| Nasutitermes   | sp       | Wood           | CM     | II                    | Nasp CM II  | Yes                       | July 2019         | Dry                       |
| Microcerotermes| sp       | Wood           | CM     | II                    | Micsp CM II | January 2020              | Dry               |
| Cephalotermes  | rectangularis | Wood        | CM     | II                    | Cefre CM II | July 2019                 | Dry               |
| Constrictotermes| cavifrons | Lichens       | FG     | II                    | Conca FG II | January 2020              | Dry               |
| Nasutitermes   | similis  | Wood           | FG     | II                    | Nassi 1 FG II | May 2019                 | Rainy             |
| Nasutitermes   | similis  | Wood           | FG     | II                    | Nassi 2 FG II | May 2019                 | Rainy             |
| Anoplotermes   | banksii  | Soil           | FG     | III                   | Anba 1 FG III | May 2019                 | Rainy             |
| Anoplotermes   | banksii  | Soil           | FG     | III                   | Anba 2 FG III | Yes                      | May 2019         | Rainy                     |
| Anoplotermes   | banksii  | Soil           | FG     | III                   | Anba 3 FG III | Yes                      | May 2019         | Rainy                     |
| Anoplotermes   | banksii  | Soil           | FG     | III                   | Anba 4 FG III | Yes                      | May 2019         | Rainy                     |
| Embiratermes   | neotenicus | Soil          | FG     | III                   | Embne FG III | January 2020              | Dry               |
3.2 Untargeted metabolite profiling

Principal component analysis (PCA) was used to find separation and grouping of samples, based on non-target volatiles analysis. Differentiation of samples according to feeding material was observed, while also geographical location seems to influence the distribution, according to the grouping of samples, which is better observed when samples with less than three representatives per genus are discarded (Fig. 1).

Among the wood-feeding representatives, we observe a clearer separation in the genus *Nasutitermes* between CM and FG. Three samples belonging to the nests from the same species (Naslu 1-3) grouped together, located at the same Quadrant as another CM wood feeder nest (Nassp CM II). The four nest samples belonging to the soil feeding species (*Anoploterms banksii*, FG) grouped together.

Samples originating from MW grouped together, occupying the upper part of chart, together with most of the representatives from feeding groups III and IV.

To further investigate differences in feeding, genus, and geographical origin, a Partial least squares discriminant analysis (PLS-DA) was used.

We have observed a clear stratification according to the food origin (Fig. 2A) where all the wood-feeding species nests occupy the I and IV quadrants, and the soil feeding species nests are on the II and III. Besides, a second degree of separation is again the geography, where FG wood feeders are on top of Quadrant I, and CM wood feeders lie grouped on Quadrant II. The inverse trend can be observed in samples from soil feeding groups, with MW samples mostly on top, and a gradient of other samples from CM in the middle towards all the FG samples at the bottom. The genus component has been represented in Fig. 2B and we find defined groups according to their group, with some overlaps in the soil feeding samples with some influence of the geographical origin (Fig. 2C).

Lastly, we represented the same OPLS-DA according to the geographical origin, and we observe in a clearer manner the overlaps in the soil feeding groups, where we can still observe a certain stratification according to the location. This stratification is very clear in the case of the wood feeding samples with the clear division between both FG originating samples, and the rest within this genus coming from CM.

The 10 most decisive compounds for separation presented on Fig. 2 are identified in Table 2, where we displayed either their contribution to feeding separation, or genera.

**Table 2.** Ten most decisive analytes when considering: A) feeding habits, and B) genera. This table represents the 10 first variables on the VIP analysis. VIP: Variable Importance in Projection; RT: Retention Time; RI: Retention Index; VIP SE: VIP standard error.

| Primary ID | Variable name                  | RT  | RI (measured) | RI  | VIP  | VIP SE |
|------------|--------------------------------|-----|---------------|-----|------|--------|
| 10         | Propanal, 2-methyl-             | 357 | 562           | 558 | 1.808| 1.645  |
| 15         | 2,3-Butanediol                 | 369 | 582           | 581 | 1.844| 1.141  |
| 18         | 2-Butanol                      | 375 | 592           | 597 | 2.196| 1.437  |
| 25         | tert-Isomyl alcohol            | 400 | 633           | 647 | 1.862| 1.571  |
| 37         | 2,3-Pentanediione              | 435 | 694           | 696 | 3.163| 1.590  |
| 38         | 3-Pentanone                    | 437 | 696           | 700 | 1.864| 1.733  |
| 98         | Hexanoic acid, methyl ester    | 644 | 919           | 925 | 1.777| 0.887  |
Primary ID | Variable name | RT (measured) | RI (Nist 2017) | VIP | VIP SE
--- | --- | --- | --- | --- | ---
18 | 2-Butanol | 375 | 626 | 610 | 1.7171.695
37 | 2,3-Pentanedione | 435 | 694 | 696 | 2.1821.092
47 | Methyl Isobutyl Ketone | 473 | 737 | 739 | 1.8181.864
81 | 1-Hexanol | 591.3 | 865 | 867 | 1.6831.181
97 | Cyclohexane, 1-ethyl-2-methyl-, cis- | 641 | 916 | 919 | 1.7430.653
103 | Cyclohexane, propyl- | 663 | 938 | 941 | 1.7881.785
127 | 3-Furancarboxylic acid, 2-methyl-, methyl ester | 723 | 998 | 999 | 1.7571.593
132 | beta-Phellandrene | 734 | 1010 | 1015 | 1.8020.736
154 | Terpinolene | 807 | 1093 | 1089 | 2.0261.176
235 | Naphthalene, 1,2,3,5,6,8a-hexahydro-4,7-dimethyl-1-(1-methylethyl)-, (1S-cis)- | 1150 | 1544 | 1536 | 1.6701.454

3.3 Targeted metabolite profiling: Naphthalene content in samples

Naphthalene is a polyaromatic hydrocarbon that was observed in the volatile profiling during the non-targeted metabolite analysis (Fig. 3A). The highest abundancies were recorded in samples from the genus Nasutitermes, Coptotermes from BR, and some of the Cubitermes. There was certainly a weak presence in many of the samples belonging to Group III, especially the species Anoplotermes banksii and the three samples from Neocapritermes taracua.

A quantitative method was applied to determine the exact concentration of this compound in nest material, where the naphthalene concentration ranged from 1.5 to 8 µg/kg (Fig 3B). These targeted metabolite results indicate that volatile naphthalene is not generally present in the studied species belonging to Group III, except for the nest material from one of the colonies of Neocapritermes taracua (colony No. 1). In keeping with the untargeted results, we have observed a general presence in species belonging to feeding Groups II and IV. In detail, these included colonies belonging to three samples from the genus Nasutitermes (Group II, both Cameroon and French Guiana), and the five species belonging to the genus Cubitermes (Group IV) that were included within this study (Cameroon and Malawi). Naphthalene was found as well in the colony nest material from Constrictotermes cavifrons, belonging to Group II (French Guiana).

When we quantified naphthalene in a targeted approach with a known standard, we observe a similar trend (Fig. 3B). The samples belonging to group III do not show detectable amounts of naphthalene (with the exception of one of the Neocapritermes samples). There are two main genera where we could detect naphthalene, mainly Nasutitermes and Cubitermes. We mostly observe samples from the African locations. We observe the presence of this compound markedly in some species of Group II and IV, while in groups I and III is very weak.

After correction with blank, volatile naphthalene was only found in one nest of the FG samples belonging to group III. These results indicate that volatile naphthalene is not generally present in the studied species belonging to Group III, except for the nest material from one of the colonies of Neocapritermes taracua (colony No. 1). In contrast, we have observed a general presence in species belonging to feeding Groups II and IV. This observed presence of volatile naphthalene does not seem to be dependant of geographical location, as we have observed it in samples from all localities. There is a notable presence of this compound in colonies belonging to three species from the genus Nasutitermes in Group II (both Cameroon and French Guiana) and in the five species belonging to the genus Cubitermes in Group IV that were included within this study (Cameroon and Malawi). Naphthalene was found as well in the colony nest material from Constrictotermes cavifrons, belonging to Group II (French Guiana).

3.4 Bacterial community structure

The 16S amplicon sequencing results of the V4 region have been represented in Table S1. We have calculated an average of 131,154 effective tags, after removal of chimera sequences for our samples. These values range from 121,959 tags (Coptotermes, BR) to 142,192 non-chimeric tags (NasLu3, CM). In our DNA isolation
process, which was carried out with sterility-opened reagents inside of a laminar flow-hood during the entire isolation process, we also included a negative DNA extraction control. In this case, after chimera removal we obtained a total of 36,406 tags.

After the data treatment in Qiime2, we did a comparison of the assigned OTUs according to their representation on each, which can be accessed in Table S2 (OTUs on each sample) and Table S3 (feature table). There was a total of 531 OTUs assigned to the negative extraction control, from a total of 19,753 OTUs in our project. The overlap with other samples has been calculated, and from these 532 OTUs, we found overlap with other samples in a total of 106 OTUs (Table S3).

We then filtered the sequences after “dada2” step. Practically, all of the sequences were removed, having remained a total of 112 tags (from a total of 36,406 prior to filtering). We have also observed a decrease of sequences in the rest of samples, where the minimum amount of tags belongs to Neota2FG.III with 38,777 tags.

3.4.1 Whole set of samples
We did not find a clear stratification of samples when carrying out a PCOA plot using the Bray-Curtis dissimilarity metric. We then evaluated the α-diversity by the Abundance Based Coverage Estimator (ACE) Index. We found that at a Kruskal-Wallis pairwise comparison there are differences in terms of geography (Fig. 6) (MW vs FG, p-value = 0.0146; and BR vs FG, p-value = 0.026). Comparing feeding groups, we found some significant differences (II vs III, p-value = 0.005; and III and IV, p-value = 0.006). When we consider genera, we found also some differences in ACE Index values (Anoplotermes vs Neocapritermes, p-value = 0.034; Cubitermes vs Neocapritermes, p-value = 0.02; and Nasutitermes vs Neocapritermes, p-value = 0.03). When using all groups, Kruskal-Wallis tests indicated differences when considering feeding groups (p-value = 0.007) and geography (p-value = 0.012).

For β-diversity calculation, among other measurements we highlight the results obtained at the Unweighted Unifrac distance. We obtained significant differences between feeding groups III and IV (p-value = 0.039), and the geographical locations FG and MW (p-value = 0.039).

3.4.2 Subset of samples with at least three representatives per genus
We carried out a selection of samples where at least three representatives per genus were present. The calculation of α-diversity was performed with the ACE Index. From a statistical point of view, we found that at a Kruskal-Wallis pairwise comparison there are differences in terms of geography between FG and MW, p-value = 0.031), although this is not significant when taking all groups together (p-value = 0.077). Comparing feeding groups, we found significant differences when comparing II vs III, p-value = 0.045; and III vs IV, p-value = 0.045, and these differences are statistically significant among all groups (p-value = 0.028). When we consider genera, we found also some differences in ACE Index values (Anoplotermes vs Neocapritermes, p-value = 0.034; Cubitermes vs Neocapritermes, p-value = 0.02; and Nasutitermes vs Neocapritermes, p-value = 0.034), being as well significant among all groups (p-value = 0.03).

The β-diversity was calculated by the Unweighted Unifrac distance, testing the significance for feeding groups through a pairwise permanova (differences between groups II and III, p-value = 0.029; groups III and IV, p-value = 0.005; differences among all groups p-value = 0.001). In the case of geography, the differences were observed between FG and MW (p-value = 0.045), while differences among all groups were of p-value = 0.02. Data with at least three representatives per genus could be compared by genera as well. In this case, we found statistical differences between Anoplotermes and Cubitermes (p-value = 0.006), and also between Anoplotermes and Nasutitermes (p-value = 0.006). Lastly, the differences between groups had a p-value = 0.001.

We used then the OTU annotated data in R through the ampvis2 package to create PCA analysis (Fig. 4). Results show a clear stratification of samples both by genera and feeding strategy (Fig. 4A), while the stratification is less clear when grouping samples by geographical location (Fig. 4B). The two principal components of the analysis explain 24.6% and 17.7% of the variance.

The phylogenetic assignment of taxa has been represented as a table with the most abundant orders where the samples have been grouped by termite genera with distinction of geographies (Table 3). We observe an overall representation of Actinomycetales, Acidimicrobiales, Clostridiales, Gemmatales, Solirubrobacterales, being Spirochaetales another category well represented in many samples.
LEfSe results are shown in Fig. 5. Samples from genus Neocapritermes were characterized for the presence of the class Betaproteobacteria, order IS_44. Nasutitermes samples was characterized by representatives from the genus Streptomyces, down to species lanatus. In the case of Cubitermes, presence of genus Rhodococcus, and members of the family Intrasporangiaceae. Lastly, Anoplotermes was characterized for the presence of Chitinophaga.

3.5 Fungal community structure

3.5.1 Whole set of samples

The first results when considering all samples are not conclusive (data not shown). We then evaluated the alpha diversity by the ACE Index. We found that at a Kruskal-Wallis pairwise comparisons there are no differences in terms of geography, either as pairwise comparisons or between all groups. Comparing feeding groups, we found some significant differences (II vs III, p-value = 0.045; II and IV, p-value = 0.003; III and IV, p-value = 0.032; between groups p-value =0.007). When we consider genera, we found also some differences in ACE Index values (Anoplotermes vs Nasutitermes, p-value = 0.018; Cubitermes vs Nasutitermes, p-value = 0.01; and Nasutitermes vs Neocapritermes, p-value = 0.017), but not between groups (p-value = 0.05).

The beta diversity was evaluated by the Unweighted Unifrac distance, found significant differences for feeding groups (II and III, p-value = 0.001; II and IV, p-value = 0.006; III and IV, p-value = 0.012), and geography (CM and FG, p-value = 0.005; FG and MW, p-value = 0.001).
3.5.2 Subset of samples with more representatives per genus

We proceeded according to the previous section, and have selected only those samples where more than two representatives are present per genus.

We evaluated the alpha diversity by the ACE Index. We found that at a Kruskal-Wallis pairwise comparison in terms of geography there were no differences among any of the locations, at a level of significance $\alpha \leq 0.05$. Comparing feeding groups, we found significant differences when comparing II vs III ($p$-value = 0.004), between groups III vs IV ($p$-value = 0.01), and lastly at the whole groups’ comparison ($p$-value = 0.006). When we consider genera, we found also some differences in ACE Index values ($Anoplotermes$ vs $Nasutitermes$, $p$-value = 0.019; $Cubitermes$ vs $Nasutitermes$, $p$-value = 0.01; $Nasutitermes$ vs $Neocapritermes$, $p$-value = 0.02; and all groups comparison, $p$-value = 0.012).

The beta diversity was calculated by the Unweighted Unifrac distance, testing the significance for Feeding groups through a pairwise permonova, where we found differences between groups II and III ($p$-value = 0.003), groups II and IV ($p$-value = 0.006), and groups III and IV ($p$-value = 0.003).

In the case of geography, the differences in $\beta$-diversity were observed between CM and FG ($p$-value = 0.025), and FG vs MW ($p$-value = 0.002). Data with at least three representatives per genus could be compared by genera as well. In this case, we found statistical differences between $Anoplotermes$ and $Nasutitermes$ ($p$-value = 0.007), and also between $Anoplotermes$ and $Neocapritermes$ ($p$-value = 0.030). Additional differences exist between $Cubitermes$ and $Nasutitermes$ ($p$-value = 0.005), and $Cubitermes$ vs $Neocapritermes$ ($p$-value = 0.009).

Lastly, there were as well statistical differences between $Nasutitermes$ and $Neocapritermes$ ($p$-value = 0.013).

A PCA plot was built in R through the ampvis2 package (Fig. 6). Results show a clear stratification of samples especially by genera, and then and feeding strategy (Fig. 6A), while the stratification is less clear when grouping samples by geographical location (Fig. 6B). The two principal components of the analysis explain 10.6% and 9.5% of the variance.

The phylogenetic assignment of taxa has been represented as a table with the most abundant orders where the samples have been grouped by termite genera with distinction of geographies (Table 4). We observe an overall representation of the phylum $Ascomycota$, among them the class $Sordariomycetes$ with the order $Hypocreales$, and the class $Dothideomycetes$. Chlorophyceae were relevant in $Cubitermes$ (CM). OTUs belonging to $Basidiomycota$ were also found but in much minor proportion than $Ascomycota$.

Table 4. Taxonomic overview of the samples retained for further analyses (with more than 2 representatives per genus). Taxa with over 5% representation have been shaded and bolded. FG: French Guiana; CM: Cameroon; MW: Malawi; Anba: $Anoplotermes$ banksii; Cub: $Cubitermes$; NasLu: $Nasutitermes$ lujae; Nas: $Nasutitermes$; NasSp: $Nasutitermes$ sp.; Neo: $Neocapritermes$ taracua.

| p_Ascomycota; c_Sordariomycetes; o_Hypocreales | Anba (CM) | Cub (CM) | Cub (MW) | Nas (FG) | NasLu (CM) | NasSp | Neo |
|-----------------------------------------------|-----------|----------|----------|----------|------------|-------|-----|
| p_Ascomycota; c_Sordariomycetes; o_Hypocreales | 31.30      | 3.74     | 6.20     | 3.18     | 9.74       | 81.75 | 0.98|
| p_Ascomycota; c_Sordariomycetes; o_Hypocreales | 16.43      | 28.30    | 25.57    | 31.53    | 53.69      | 6.01  | 7.66|
| p_Ascomycota; c_Dothideomycetes; c_Exobasidiomycetes | 33.94      | 40.02    | 30.97    | 12.03    | 2.79       | 3.75  | 37.95|
| p_Ascomycota; c_Dothideomycetes; c_Exobasidiomycetes | 6.45       | 6.26     | 12.87    | 28.92    | 5.69       | 5.91  | 23.25|
| Unidentified Eukaryote | 7.82 | 11.24 | 10.01 | 7.10 | 22.79 | 1.68 | 16.94|
| p_Chorophyta; c_Chorophyceae; o_Chorophyceae | 0.55       | 5.03     | 0.55     | 0.03     | 0.00       | 0.00  | 0.00|
| p_Phragmoplastophyta; c_Embryophyta; o_Magnoliophyta | 0.07       | 0.00     | 0.00     | 5.08     | 0.00       | 0.00  | 0.02|
| p_Basidiomycota; c_Exobasidiomycetes | 0.02       | 0.01     | 0.02     | 0.00     | 0.01       | 0.01  | 6.47|
| p_Basidiomycota; c_Agaricomycetes; o_Incertae_Sedis | 0.02       | 0.00     | 0.01     | 4.81     | 0.00       | 0.00  | 0.00|
| p_Basidiomycota; c_Dothideomycetes; o_Pleosporales | 0.12       | 0.21     | 4.77     | 0.08     | 0.02       | 0.00  | 3.22|
| p_Basidiomycota; c_Dothideomycetes; o_Capnodiales | 1.04       | 1.90     | 1.94     | 4.20     | 0.21       | 0.00  | 1.38|
| Unassigned | 0.61 | 0.66 | 1.94 | 0.11 | 4.83 | 0.86 | 0.72|
| p_Basidiomycota; c_Dothideomycetes; o_Capnodiales | 1.08       | 0.76     | 0.74     | 2.61     | 0.03       | 0.01  | 0.43|
| p_Basidiomycota; c_Dothideomycetes; o_Capnodiales | 0.15       | 1.20     | 2.92     | 0.12     | 0.00       | 0.00  | 0.00|
| p_Basidiomycota; c_Eurotiomycetes; o_Chaetothyriales | 0.03       | 0.31     | 0.57     | 0.00     | 0.00       | 0.00  | 0.90|
The LEfSe analysis was carried out, and we found no significant taxa.

3.6 Metagenomic sequencing
3.6.1 Cleaning, assembly, SqueezeMeta pipeline
We obtained a total of 625,536,948 raw reads (Table 5). Clean quality control passed reads were digitalized as described in methodology, and used as input in SqueezeMeta for the first assembly. This assembly yielded a 5,632,566 contigs, totalling 4,093,561,329 bp, with a minimum of 200 bp, and maximum length of 146,620 bp, average size of 726.77 bp, and N50 of 776 bp. The reads were assigned taxonomically, and the most abundant hits per sample are shown in Table 6, while the statistics on Open Reading Frames (ORFs) is presented on Table 7.

Binning rendered a total of 47 bins (Table 8), with 33 over 50% completeness, from which 16 ≥75%, and finally 2 over 90% completeness. Twenty-two bins had a contamination lower than 10% and 3 over 50%.

### Table 5. Statistics on metagenomic sequencing. Raw reads: total amount of reads of raw data, each four lines taken as one unit. For paired-end sequencing, it equals the amount of read1 and read2, otherwise it equals the amount of read1 for single-end sequencing. Seq.: sequencing; Geo.: Geographical. Raw data: (Raw reads) * (sequence length), calculating in G. For paired-end sequencing like PE150, sequencing length equals 150, otherwise it equals 50 for sequencing like SE50. Effective: (Clean reads/Raw reads) * 100. Error: base error rate. Q20, Q30: (Base count of Phred value > 20 or 30) / (Total base count). GC: (G & C base count) / (Total base count).

| Geo. location | Species | Code | Sample seq. code | Data treatment code | Raw reads | Raw data | Effective (%) | Error (%) | Q20 (%) | Q30 (%) | GC (%) |
|---------------|---------|------|------------------|--------------------|-----------|----------|---------------|-----------|---------|---------|--------|
| FG            | Anoplotermes banksii | Anba 2 FG III | Ban2 | Ban1 | 53,288,002 | 8 | 99.67 | 0.02 | 98.3 | 95.44 | 66 |
| FG            | Anoplotermes banksii | Anba 3 FG III | Ban3 | Ban2 | 55,905,619 | 8.4 | 99.7 | 0.02 | 98.4 | 95.59 | 65.8 |
| FG            | Anoplotermes banksii | Anba 4 FG III | Ban4 | Ban3 | 66,039,800 | 9.9 | 99.59 | 0.03 | 97.6 | 93.83 | 68.3 |
| CM            | Nasutiters mes lajae | Naslu 1 CM II | Nal1 | NasLu1 | 63,479,926 | 9.5 | 99.68 | 0.02 | 98 | 94.97 | 68.3 |
| CM            | Nasutiters mes lajae | Naslu 2 CM II | Nal2 | NasLu2 | 53,196,396 | 9 | 99.67 | 0.02 | 98.1 | 95.17 | 68.3 |
| CM            | Nasutiters mes lajae | Naslu 3 CM II | Nal3 | NasLu3 | 63,131,466 | 9.5 | 99.66 | 0.02 | 98.1 | 95.14 | 67.5 |
| CM            | Nasutiters mes sp. | Nassp CM II | NasSp | NasSp | 56,112,754 | 8.4 | 99.7 | 0.02 | 98.2 | 95.34 | 68.1 |
| CM            | Cubitermes planifrons | CubPL CM IV | CubPL | CubPl | 49,692,682 | 7.5 | 99.74 | 0.02 | 98.2 | 95.35 | 67.1 |
| CM            | Cubitermes sulcifrons | CubSu CM IV | CubSu | CubSu | 55,052,860 | 8.3 | 99.71 | 0.02 | 98.2 | 95.25 | 67 |
| MW            | Cubitermes maneris | Cubmu 1 MW IV | MLW3 | CubMu | 56,193,314 | 8.4 | 99.75 | 0.02 | 98.2 | 95.24 | 66.4 |
| MW            | Cubitermes inclitus | Cubin MW IV | MLW2 | CubIn | 53,443,938 | 8 | 99.7 | 0.02 | 98.2 | 95.34 | 69.1 |

Total 625,536,948
Table 6. Statistics on metagenomic sequencing. Abund.: abundant; Superk: superkingdom; Actinob.: Actinobacteria; Actinom.: Actinomycetia; Streptosp.: Streptosporangiaceae; Hyphom.: Hyphomicrobiales; Microc.: Micrococcales; Microb.: Microbacteriaceae; Gryllot.: Gryllotalpicola; Dormib.: Dormibacteraeota; Alphaprot.: Alphaproteobacteria. Note: “Data treatment code” (Table 4) has been used to name samples, for simplicity purposes.

| Most abund. taxa | Ban1 | Ban2 | Ban3 | CubIn | CubMu | CubPl | CubSu | NasLu1 | NasLu2 | NasLu3 | NasSp |
|------------------|------|------|------|-------|-------|-------|-------|--------|--------|--------|-------|
| Super. phylum    | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria | Bacteria |
| Actinob. class   | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. | Actinob. |
| Actinom. order   | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. | Actinom. |
| Streptosp. family| Pseudonocardiales | Pseudonocardiales | Pseudonocardiales | Actinomycetia | Actinomycetia | Actinomycetia | Actinomycetia | Actinomycetia | Actinomycetia | Actinomycetia | Actinomycetia |
| Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium | Mycobacterium |
| Micrococcales |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |Microc. |
| Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |Microb. |
| Gryllot. genus |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |Gryllot. |
| Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium | Actinobacterium |
| Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |Bradyrhizobium |
| Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae | Acidimicrobiaceae |

Table 7. Statistics on ORFs after SqueezeMeta pipeline.

| Number of ORFs | Whole Assembly | Ban1 | Ban2 | Ban3 | CubIn | CubMu | CubPl | CubSu | NasLu1 | NasLu2 | NasLu3 | NasSp |
|----------------|----------------|------|------|------|-------|-------|-------|-------|--------|--------|--------|-------|
| Number of rRNAs | 2,390 | 1,024 | 1,276 | 1,415 | 1,221 | 1,447 | 1,339 | 1,380 | 995 | 1,039 | 1,223 | 1,257 |
| Number of tRNAs/tmRNAs | 27,874 | 4,555 | 6,898 | 6,420 | 4,890 | 6,658 | 8,385 | 8,613 | 4,384 | 4,378 | 4,941 | 4,643 |
| ORFs by Aragorn | 27,874 | 4,555 | 6,898 | 6,420 | 4,890 | 6,658 | 8,385 | 8,613 | 4,384 | 4,378 | 4,941 | 4,643 |
| ORFs by Prodigal | 8,097,351 | 1,391,578 | 2,459,488 | 2,384,504 | 1,571,393 | 2,462,813 | 2,691,629 | 2,758,078 | 1,101,858 | 1,039,555 | 1,161,645 | 1,147,044 |
| ORFs by barrnap | 2,390 | 1,024 | 1,276 | 1,415 | 1,221 | 1,447 | 1,339 | 1,380 | 995 | 1,039 | 1,223 | 1,257 |
| Orphans (no hits) | 1,343,453 | 148,874 | 277,464 | 255,731 | 172,699 | 267,778 | 287,824 | 304,858 | 193,024 | 170,776 | 187,135 | 190,547 |
| No tax assigned (with hits) | 18,951 | 2,078 | 4,110 | 3,915 | 2,625 | 4,320 | 4,687 | 4,828 | 3,372 | 2,910 | 3,849 | 3,642 |
| KEGG annotations | 4,095,584 | 801,355 | 1,412,923 | 1,388,974 | 899,823 | 1,423,940 | 1,552,121 | 1,588,244 | 597,678 | 580,592 | 648,183 | 639,147 |
| COG annotations | 5,811,435 | 1,111,402 | 1,946,729 | 1,902,975 | 1,237,057 | 1,941,438 | 2,133,173 | 2,173,849 | 801,962 | 768,568 | 862,097 | 848,836 |
| PFAM annotations | 2,844,947 | 646,229 | 1,072,204 | 1,029,204 | 697,782 | 1,045,021 | 1,161,233 | 1,169,171 | 455,815 | 440,914 | 488,195 | 483,351 |
Table 8. Statistics on bins after SqueezeMeta pipeline. Qual.: quality; contam.: contamination.

| Number of bins          | 47  |
|-------------------------|-----|
| Complete ≥ 50%          | 33  |
| Complete ≥ 75%          | 16  |
| Complete ≥ 90%          | 2   |
| Contamination < 10%     | 22  |
| Contamination ≥ 50%     | 3   |
| Congruent bins          | 7   |
| Disparity >0            | 40  |
| Disparity ≥ 0.25        | 21  |
| Hi-qual. bins (>90% complete, <10% contam.) | 1 |
| Good-qual. bins (>75% complete, <10% contam.) | 3 |

3.6.2 Overall characterization of samples

We used the TPM values for all KEGG annotated genes, to evaluate the general structure of samples and their relationships through non-metric multidimensional scaling (NMDS), which was carried out in R (Fig. 7). We observe that all Nasutitermes samples cluster together, while the other two groups of nests belonging to genera having a soil feeding strategy are in the same area, although having two dissimilar patterns.

We then analysed the taxonomic signature of samples after metagenomic sequencing. The level Phylum/genus was represented as a bar chart in Fig. 8.

We can observe that the four samples belonging to the genus Nasutitermes (all originating in CAM) have a similar profile at the genus level, while the two soil-feeding genera (Cubitermes, from MW and CAM; Anoplotermes, FG), despite differences at the species level, present a closer profile. An outstanding exception is the biological replicate Ban1, with a lower representation of Proteobacteria than the rest of samples from this genus (all samples as well belonging to the species Anoplotermes banksii). For Nasutitermes, we find differences between the three Nasutitermes lujiae replicates, and Nasutitermes sp. The samples belonging to the genus Cubitermes have a more variable profile. Among those, CubPl and CubSu belong to the CM sampling site, while CubIn and CubMu originate in MW.

We further carried out a subset of the present genera belonging to the “aromatic AA” representation (“Phenylalanine, tyrosine, and tryptophan biosynthesis”) (Fig. 9). The most important genera for soil feeders seem to lie within an unclassified Streptosporangiales genus with higher abundance for Cubitermes and Anoplotermes, while Nasutitermes had higher representation of Microlunatus, an unclassified genus within Microbacteriaceae, and Gryllotalpicola.

A LDA Effective Size (LEfSe) was calculated using the PFAM annotation, and differences were found among the three represented genera. We obtained 61 classes of annotated genes that are specific biomarkers for each of the three (Fig. 10, Table 9). Anoplotermes was characterized by several biomarkers, among the first PF00196: Bacterial regulatory proteins luxR family, PF13424: Tetratricopeptide repeat, PF03704: Bacterial transcriptional activator domain, and PF13191: AAA ATPase domain. Cubitermes had PF00872: Transposases, Mutator family, and PF0084: sulfatase as the only biomarker categories found. One of the most represented biomarker categories in the genus Nasutitermes was PF00296: Luciferase-like monooxygenase followed by PF02515: CoA transferase family III, PF00873: AcrB/AcrD/AcrF family, and PF00171: Aldehyde dehydrogenase family.

Table 9. LEfSe biomarker categories for the three genera included in the metagenomic sequencing.

| Term                          | Genus   | LDA Score Log10 |
|-------------------------------|---------|-----------------|
| PF00196 Bacterial regulatory proteins luxR family | Anoplotermes | 2.836470466 |
| PF13424 Tetratricopeptide repeat | Anoplotermes | 2.590311713 |
| PF03704 Bacterial transcriptional activator domain | Anoplotermes | 2.445492463 |
| PF13191 AAA ATPase domain      | Anoplotermes | 2.442521088 |
| PF07719 Tetratricopeptide repeat | Anoplotermes | 2.411987663 |
PF04392 ABC transporter substrate binding protein
PF00515 Tetrapiricerotide repeat
PF13374 Tetrapiricerotide repeat
PF07228 Stage II sporulation protein SpoIIIE
PF00553 Cellulose-binding domain
PF13176 Tetrapiricerotide repeat
PF14200 Ricin type beta trefoil lectin domain like
PF00652 Ricin type beta trefoil lectin domain
PF13185 GAF domain
PF04000 WD domain G beta repeat
PF00990 Diguanylate cyclase GGDEF domain
PF00211 Adenylate and Guanylate cyclase catalytic domain
PF13683 Integrase core domain
PF00872 Transposase Mutator family
PF00884 Sulfatase
PF0296 Luciferase-like monoxygenase
PF02515 CoA transferase family III
PF00873 AcrB/AcrD/AcrF family
PF00171 Aldehyde dehydrogenase family
PF04185 Phosphoesterase family
PF00126 Bacterial regulatory helix turn helix protein lysR family
PF12833 Helix turn helix domain
PF03466 LysR substrate-bindin domain
PF13458 Periplasmic-binding protein
PF03972 MmgE PrpD family
PF13181 Tetrapiricerotide repeat
PF00593 TonB dependent receptor
PF05685 Putative restriction endonuclease
PF01408 Oxidoreductase family NAD-binding Rossmann fold
PF13432 Tetrapiricerotide repeat
PF19305
PF13431 Tetrapiricerotide repeat
PF01391 Collagen triple helix repeat 20 copies
PF02826 D isomer specific2 hydroxoyacid dehydrogenase NAD-binding domain
PF13428 Tetrapiricerotide repeat
PF13414 TPR repeat
PF09084 NMT1 THIS-like
PF00664 ABC transporter transmembrane region
PF07715 TonB dependent Receptor Plug Domain
PF13433 Periplasmic-binding protein domain
PF13371 Tetrapiricerotide repeat
PF02661 Fic DOC family
PF01436 NHL repeat
PF04261 Dyp-type peroxidase family
PF02321 Outer membrane efflux protein
PF12849 PBP superfamily domain
PF14559 Tetrapiricerotide repeat
PF07804 HipA likeC terminal domain
PF04255 Protein of unknown function DUF433
3.6.3 Differential gene content analysis

We further evaluated the differences among samples in a similar way to transcriptomics Differentially Expressed Genes. In this case, we termed them “Differentially Present Genes” since we are studying the metagenome. Firstly, we approached the structure of the data through a hierarchical clustering using Euclidean distance (Fig. 11A). We find two clearly separated nodes, one with all Nasutitermes samples (belonging to nests of wood feeding termites from the genus Nasutitermes, CAM), and those belonging to the soil feeding genera forming the second group. In the second group, there is also a subdivision in two nodes. One comprises Ban1 and Cubin. The other is subdivided into a branch where we find the other two Anoplotermes samples (Ban2 and Ban3), and another branch by the rest of the Cubitermes samples. Among them, at an inner position we find the Cubitermes samples originating in CAM. These differences can be clearly observed through a correlation heatmap (Fig. 11B), and the overall gene content can be observed at Fig. S2, for PFAM annotation.

The main differences existing between pair of samples in terms of Differentially Present Genes (DPG) have been displayed through volcano plots (Fig. 12). When using COG annotation, at a fold change of 2 (|log2FoldChange| ≥ 1; DESeq2 padj ≤ 0.05) we found from a total of 2,099 DPGs, 1,147 upregulated genes (more present in Cubitermes) when comparing the two soil feeders (Cubitermes vs. Anoplotermes), and 952 COG annotated genes with higher presence in A. banksii samples (Fig. 12A). When comparing Nasutitermes and Cubitermes, we found 6,771 upregulated DPGs, and 7,871 downregulated ones (statistically significant higher presence in the soil feeder) (Fig. 12B), the last comparison was carried out between Nasutitermes and Anoplotermes (Fig. 12C). We found 7,152 upregulated genes, while there were 7,478 downregulated DPGs.

We then carried out a similar analysis using the PFAM annotated genes (heatmap showing the overall in Fig. S2). For the comparison between the functional profiles of microbial communities present in the nests from both soil feeding genera, we found a total of 194 PFAM annotated contigs with statistically significant over-presence in Cubitermes, while Anoplotermes had 265 PFAM-annotated contigs with over-presence. We could equally state that these 294 genes have a decreased presence in the microbial communities from Cubitermes genus, in comparison with the microbial communities on those nests from Anoplotermes. We found some genes with extreme differences in presence, such as PF03945: delta endotoxin, N-terminal domain (FC = 16,326,654.29; p-adj = 1.55 $10^{-67}$) and PF14470: Bacterial PH domain (FC = 5,562,101.72; p-adj = 1.34 $10^{-15}$). On the other side, with higher presence in Anoplotermes PF02183: Homeobox associated leucine zipper (FC = 3,509.92; p-adj = 9.10$^{-14}$).

The comparisons between microbial communities from the two soil feeding termites with the wood feeding termite nest provided again a higher number of DPGs, as previously observed with the COG annotation. The first comparison was carried out between Nasutitermes and Cubitermes. We found a total of 2,002 PFAM annotated contigs with statistically significant over-presence in Nasutitermes, while Cubitermes had 1,114 PFAM-annotated contigs with over-presence. We could equally state that these genes have a decreased presence in the microbial communities from Nasutitermes genera, in comparison with the microbial communities on those nests from Cubitermes genera. Among the functions most present in Nasutitermes we find PF16937: Type III secretion system translocator protein, HrpF (FC = 4,569.92; p-adj = 7.20 $10^{-13}$), PF17667: Fungal protein kinase HrpF (FC = 2,032.43; p-adj = 2.26 $10^{-97}$), and PF05420: Cellulose synthase operon protein C C-terminus (BCSC_C) (FC= 1,943.41; p-adj = 1.01 $10^{-15}$). The other direction, higher presence in Cubitermes, provides a number of highly present PFAM functions (Table S1), for instance PF03945: delta endotoxin, N-terminal domain (FC= 299,812,460.54, p-adj = 8.66 $10^{-14}$).

The comparison between Nasutitermes and Anoplotermes yielded 2,034 upregulated genes, while 1,207 were downregulated.
3.6.4 Annotation of ORFs in specialized databases
We carried out further annotation of the ORFs retrieved by SqueezeMeta with several specialized databases of interest (view Mat. Methods section) (Suppl. material S2 to S5). The conditions that had to be satisfied for being included in our list was a minimum identity $\geq 40\%$, bit-score $\geq 50$. An arbitrary threshold of $\geq 90\%$ identity (diamond blastx) was selected firstly for considering genes as known or already deposited in databases, while under this threshold genes were considered novel.

3.6.4.1 Differential gene content: AromaDeg
We used the annotation obtained through diamond blastx (AromaDeg DB) as an input for differential gene content analysis. We firstly curated the results to eliminate those that were duplicated, where the total AromaDeg annotated genes counted for a total of 8,132. The comparison Cubitermes vs Anoplotermes yielded 14 upregulated DPGs, and 21 downregulated DPGs. The comparison Nasutitermes vs Cubitermes was carried out, and 468 genes were upregulated or significantly more present in Nasutitermes genus, while 82 were downregulated or with higher presence in Cubitermes. For the comparison Nasutitermes vs Anoplotermes we found 224 upregulated genes, and 12 downregulated ones.

3.6.4.2 Differential gene content: CAZy
We carried out the analysis with the CAZy DB annotated genes. In the comparison between both soil feeders we found no statistical differences. Nasutitermes vs Cubitermes yielded 287 genes upregulated, while 2 were downregulated. We found a larger number of genes for the Nasutitermes vs Anoplotermes comparison. This one provided 519 upregulated genes, while 218 were downregulated.

3.6.4.3 Differential gene content: CARD
The comparison between the soil feeders yielded 54 upregulated genes, and 241 genes with higher presence in Anoplotermes. The correlation among samples and hierarchical clustering divided clearly all samples by genera. The comparison Nasutitermes vs Cubitermes resulted in 1,587 CARD annotated genes with higher presence in the wood feeder nest community, while the soil feeder nest community had 988 upregulated genes. When Nasutitermes was compared with Anoplotermes, we found 1,403 upregulated genes, while 1,745 genes were downregulated (higher presence in the soil feeder).
4 Discussion

A very important aspect of termite ecological success is their ability to maintain a pathogen-free nest. This is of paramount importance since they live in colonies with very high density of individuals, where conditions are humid, nutrients are widespread, and temperatures can favour the growth of microbial strains. Among different strategies, fumigation with organic compounds such as naphthalene can aid (Chen et al., 1998b). We wanted to evaluate through an untargeted metabolomic approach the presence of volatile compounds, including naphthalene. We studied termite nests originating in three geographical locations, displaying different feeding strategies (Table 1). Some of the difficulties that we have faced came as a consequence of the taxonomic identification of samples. While we for instance thought that Cubitermes samples from MW were the same species, it was further confirmed to be three different ones. This fact creates an additional constraint since the samples are unique and sampling campaigns are costly.

We have then carried out a multi-omics approach with metabolite untargeted and targeted analysis, being the second omic level the taxonomic microbial profile, and lastly a metagenomic sequencing of selected samples with detected naphthalene presence.

One of the most important findings, is that the metabolite untargeted approach was able to stratify the samples according to feeding strategies, with a certain influence of the geographical location. Samples with one representative were eliminated from the analysis, especially due to bias from samples originating in artificial environments (breeds). The 10 most decisive analytes were several organic compounds, most of them of type alcohol. Naphthalene was also found when we consider the separation between genera (Table 2B).

One of these compounds (Table 2), 2-butanol, has resemblances with 2-methyl-butanol, previously reported to have antifungal activity (Matsuura and Matsunaga, 2015). We found it important for separation of samples according to feeding habits and genera.

Terpenes are a group of aromatic compounds that have been related to mediation of information transmission (Gershenzon and Dudareva, 2007), or have been tested specifically as a toxic compound against termites (Seo et al., 2014). Among them, terpinolene was found as a decisive analyte in our study. Terpenes have being reported repeatedly in termite literature as a chemical component of the defensive secretions in Nasutitermes soldiers (Prestwich et al., 1981), not related with feeding habits as explained by its absence in workers. Besides, terpenes have been appointed as an antimicrobial substance, as confirmed by Mitaka and collaborators (2017). Terpinolene was firstly identified in the cephalic secretion from Amitermes herbertensis (Moore, 1968), and later in small quantities of a Nasutitermitinae, Grallatotermes africanus (Prestwich, 1979), in the defensive secretion from Nasutitermes ephratae soldiers (Valterová et al., 1986), soldier frontal glands in Termitogeton planus (Dolejšová et al., 2014), and significant amounts in Constrictotermes cyphergaster soldiers (De Mello et al., 2021). According to our results, this compound is mostly found in the nests from Nasutitermes, with neglectable concentrations in the nests from the other genera. Considering literature, we believe that this compound is used as an antimicrobial defensive mechanism, and that the secretions may be carried out by soldiers, as extensively reported in several studies.

A PLS-DA analysis showed a better distinction according to the feeding habit and genera. Most of the soil feeding genera cluster together. Thus, we hypothesized that the geographical origin has a lower effect on the nest microbial communities than the feeding habits or the genera, and a great distinction among metabolite content in the nest material from our analysed termite species reflects the termite dietary habits.

At the targeted approach, our data indicated a lower level of naphthalene in comparison with published literature (Chen et al., 1998b; Wilcke et al., 2000). In overall, two genera concentrate the readings of naphthalene (Nasutitermes and Cubitermes), while the rest of samples had a scattered representation in some genera. We did not record naphthalene in the soil feeding nests from Anoplotermes from FG. We believe that the type of sample preparation could have caused partly these readings, although a soil standard indicated correctness in our method. In some cases, the results could have been affected by the time elapsed from sampling to processing, matter of a few days due to travelling and transportation. Even if that is the case, we could still observe clear differences between samples at the volatile level, which were further confirmed at the other omic levels.

4.2 Sequencing

4.2.1 Bacteria
Our analyses included a negative extraction control (Eisenhofer et al., 2019). Bacteria are ubiquitous and even dust on a tube can produce a signal after PCR amplification, while many times there is as well an input of DNA in the reagents (Salter et al., 2014). We have deployed specific measurements for avoiding contamination inputs, for example carrying out the nest sampling under laminar-flow conditions, equally as those put in practice when performing the DNA isolation protocol. The reagents were treated as if they were sterile microbial growth media, putting in practice all aseptic methods that are normally used when working in microbiological experiments. While reagents are not sterile, we only used them for this project, and all bottles were exclusively opened under aseptic conditions and using sterile filtered tips. These measures are of paramount importance when working with samples with potentially small quantities (Weyrich et al., 2019), but as well for any other type of projects, a cross contamination can lead to incorrectly establish a core microbiome among samples. Despite these precautions, stringent conditions, and wide experience working in aseptic conditions by the sole operator, we have observed a number of OTUs in negative samples. After initial analysis of these data, and we confirm that the OTUs do not overlap with the ones in our samples, except for a really minor amount which would not affect the downstream results. However, we carried out a filtration step according to common procedures in Qiime2, which resulted in the elimination of virtually all contaminating tags. This elimination was effective for the negative control sample (some residual tags remained with zero overlap with the rest of samples), and for the rest of samples.

4.2.1 Whole set of samples

We firstly analysed the microbial communities in the nests, with a quick overview on the bacterial community compositions through calculation of the Bray-Curtis dissimilarity metric (Sorensen, 1948; Bray and Curtis, 1957). This allowed us to obtain a quick classification of the studied colonies given the broad species origin. This index provides an idea of similarity/dissimilarity due to the product of present taxa and their relative abundances (van Rensburg et al., 2015). When studying the whole set of samples, we observed similar results than those obtained with the overall untargeted profile (Fig. S1), without a very clear stratification of microbial communities. The ACE index of α-diversity provides an overview of the microbial richness within one sample, but does not take in account the OTUs abundance (Qiao et al., 2017). We observed an effect of geographical origin in the richness of the microbial communities associated to different nests, where the samples with highest values were those from Breeds. This could be owed to an artificial input of different strains in several samples and colonies kept in similar laboratory conditions and close proximity. Besides, there is an input of human microbiome that could be at play. The second highest level was that of MW samples. We hypothesize that the savannah conditions could pose a higher degree of stress over soil microbial communities, which could find a better shelter in these nests.

In terms of feeding strategies, wood feeders showed statistical differences with soil feeders from group III. While this could be owed to the different type of diet, these differences were as well present between groups III (with origin in FG) and IV (only African representatives).

Genera was a factor with more differences among microbial communities, where we found differences mainly in those ones with at least three representatives. Among wood or soil feeders, we observed only ACE index differences between Nasutitermes and Neocapritermes. Kruskal-Wallis comparisons among all groups indicated differences when considering feeding groups and geography.

At β-diversity the weighted UniFrac distance can represent the differences between samples according to the evolutionary history (Lozupone and Knight, 2005; Lozupone et al., 2007, 2011), and how different are two microbial communities. Our results show significant statistical differences between feeding groups III and IV, and between FG and MW. That could be surprising considering that groups III and IV are both soil feeders, but can be explained because all of the samples from group III originate in FG, while the ones from group IV originate in Africa, with a great number of them from MW. These two locations are precisely the ones yielding differences when considering geography, and agree with the reported statistical differences among all groups observed at the α-diversity, which were observed on feeding strategies and geographical origin. For β-diversity, we could not compute distances based on genera, as some groups had only one representative.

4.2.1.2 Samples with at least three representatives per genus
Unfortunately, we could not increase the number of samples from other genera than *Anoplotermes, Cubitermes, Nasutitermes, and Neocapritermes*, since we largely rely on bringing nests from distant and remote geographical locations, and not all times they survive transportation. Soil feeding species are especially sensitive. A long-time observation in our laboratory, is the fascinating fact that nests without the presence of termites are quickly populated by pathogenic fungi. At the moment we ignore if it is infection what drives the death of these colonies, or these species are especially sensitive to the artificial laboratory conditions. What is clear is that in matter of scarcely two weeks fungi dominates the colony and impairs sample processing or the use of the material. This is another factor that limits our potential for bringing a higher number of colonies to our home laboratory.

Their stratification in a PCOA plot using Bray Curtis values (plot not shown), indicate an important component of variability in bacterial microbial populations explained by the type of feeding habits, as most of the samples are clustering in proximal areas of the representation, with only one exception from one sample belonging to group II within group IV. Additionally, we got the idea that the genera can be another important factor in the explanation of variability. The potential lack of influence of the geographical location is indicated by the presence of *Cubitermes* samples in the same sector, despite originating in Malawi and Cameroon (with known differences between tropical rainforest and savannah). However, we cannot discard it completely, since some samples at the feeding group III (*Neocapritermes*) are closer to a representative of group II from FG. These results are indicative that feeding group and genus have a stronger role in stratification than geographical location. We cannot ascertain that naphthalene presence could be driving the stratification.

Among other biodiversity calculations we obtained the Faith Phylogenetic Diversity (Faith’s PD) index (Faith and Baker, 2006), indicating that among geographical locations exists a statistical difference between those samples from Malawi and those in French Guiana, as reported as well for the ACE index. This indicates that differences in diversity at the nest bacterial microbial communities have, until some extent, influences due to the different conditions at geographical locations. The lack of differences between the rest of groups, for instance CM and MW, could point towards a conservation of the diversity in Africa, despite the marked ecological differences between the two sites. These differences across the ocean are of higher entity, accounting for the statistically significant differences between FG (tropical rainforest) and MW (savannah), although not enough when comparing the two tropical sites at both sides of the Atlantic.

We found as well statistical differences in diversity at the genera level, accounting for differences between *Cubitermes* (MW-CM, SF) and *Neocapritermes* (FG, SF), and at the other hand between *Nasutitermes* (CM, WF) vs *Neocapritermes* (FG, SF), and among *Anoplotermes* (FG) and *Neocapritermes* (FG). Here we can observe as well that geography is not a strong factor determining richness differences (comparison between *Anoplotermes* and *Neocapritermes*), but genera seems to take a greater role. Furthermore, when considering feeding habits we must remember that *Cubitermes* and *Neocapritermes* are in two distinct feeding groups (Group IV and Group III respectively), were we also report significant statistical differences. We could point towards the feeding strategy as another difference to account for the bacterial diversity. In terms of β-diversity results are mainly in agreement with the α-diversity, except for the differences among genera.

The PCA plot based on the annotated OTUs further confirmed that the differences between microbial communities are determined strongly by the genera and then the feeding strategy (Fig. 6A). While *Anoplotermes* and *Neocapritermes* belong to the same feeding group, they occupy distant locations in the plot, what provides the idea that genus is a stronger component than feeding strategy. The second part of the plot (Fig. 6B) indicates that the geographical location is weak when explaining the variability among nest bacterial communities based on 16S amplicon data.

The results from our study agree with the published termite literature regarding the bacterial taxonomic composition of nests. *Actinomycetales* were reported as dominant in termite nests (Krishanti et al., 2018). These are Gram-positive bacteria capable of forming branching hyphae, and related to nutrient recycling roles (Goodfellow and Williams, 2003), and have been used for production of secondary metabolites, accounting for a very large fraction of all bioactive secondary metabolites in use at industry (Olano et al., 2008). This group has been also reported in the study from Chouvenc et al. (2018), where authors studied the origins of acquisition of *Streptomycetes*, the largest genus within *Actinobacteria* (El-Naggar, 2021), in *Coptotermes* nests. Authors have reported a recruitment from surrounding soil. Similar results regarding *Streptomycetes* were reported in nests of *Reticulitermes flavipes* (Aguero et al., 2021). Another group of bacteria notably described in literature is that of...
Rhizobiales, Gram-negative bacteria belonging to the Alphaproteobacterial class, and with many members capable of fixing nitrogen (Beeckmans and Xie, 2015) and associated with many plants, lichens, and mosses (Erlacher et al., 2015). This order was also reported as enriched in termite nests of Coptotermes testaceus, Heterotermes tenuis, and Nasutitermes octopilis (Soukup et al., 2021). In our case the proportion is limited. Clostridiales are gram-positive anaerobic bacteria (Bowman, 2011) that have capability to form resistant spores (Paredes-Sabja et al., 2011), which have been described as thermophilic ethanologens capable of fermenting lignocellulosic biomass (Arora et al., 2019). This bacterial order includes species that are lignocellulolytic (Auer et al., 2017). This order has been associated with the termite gut community of soil feeders (Ohkuma and Brune, 2010), and also Reticulitermes santonensis, through the presence of Clostridia related clones within Firmicutes (Yang et al., 2005), and was appointed as one of the groups carrying out recycling of uric acid (Thong-On et al., 2012). In Odontotermes yunnanensis Long et al. (2010) report the presence of a wide representation of bacteria from Clostridium genus in fungus comb structures. We found that this group was of the highest proportion in two of the soil feeders (Anoplotermes and Cubitermes), but also in nests from Nasutitermes. Regarding Acidimicrobiales, we have found no specific information in termite literature. Solirubrobacterales, an order with few species described as Gram-positive mesophilic bacteria growing in a moderate range of temperature (Whitman and Suzuki, 2015). Our results are in agreement with those described by Enagbonma et al. (2020), where they report this group at termite mounds from Coptotermes species. Uncultured representatives of Solirubrobacterales have been associated to the degradation of lignin in soils (Wilhelm et al., 2018; Silva et al., 2021), previously reported as well in sugarcane farm soils (Ogola et al., 2021).

Spirochaetales are Gram-negative that can comprise both anaerobic and strict aerobic bacteria (Smibert and Johnson, 1973). They have been reported in increased numbers when manure is applied to soil (Rieke et al., 2018). This group was also reported in R. santonensis gut (Yang et al., 2005) and gut from Coptotermes gestroi (Do et al., 2014), and have been appointed as active cellulose degraders (Xia et al., 2014), with abundance in wood-feeding termites (Tokuda et al., 2018; Hu et al., 2019).

Gemmatales are Gram-negative chemoheterotrophic bacteria (Dedysh et al., 2020), which could act as an indicator of elevated phosphorus content in soil (Mason et al., 2021), and in a complex microbial community dedicated to the utilization of cellulose could be providing endocellulases and β-glucosidases activities (McDonald et al., 2019).

The characterization of samples through biomarkers (LEfSe analysis, Fig. 5) indicated that certain genera were representative of specific samples when compared together. Streptomycetes was associated with Nasutitermes, an observation that is in agreement with the study of Chouvenc (2018), and Aguero (2021), and could be indicative of a defensive role of this bacterial order in the termite nest.

Neocaprimermes with Beta-proteobacteria, which have been described as characteristic cellulolytic bacteria in acidic forest soils in temperate areas (Štursová et al., 2012), and belonging to this class the genera Burkholderia and Collimonas have been characterized as capable of carrying out mineral-wheathering (Lepleux et al., 2012). Nests microbial communities from Cubitermes were associated through LDA with the genus Rhodococcus, comprising aerobic Gram-positive bacteria. Members from this genus have been associated with degradation of nitroaromatic compounds (Subashchandrabose et al., 2018), but more importantly with lignin degradation activities (Chong et al., 2018). This observation is aligned with the fact that lignin degradation is not important in wood feeders (Ohkuma, 2003; Griffiths et al., 2013), and could be an indication that members of this genus carry this important task in the Cubitermes-associated communities. and Intrasporangiaceae.

Chitinophaga has been described as a chitinolytic and lignolytic genus (Funnicelli et al., 2021), besides Spirochaeta, where one of the species was reported to degrade complex plant polysaccharides and depolymerize lignin (Pandit et al., 2016). That observation agrees with the type of diet in Anoplotermes, which will find lignin enriched sources in the soil, and by construction of the nest this type of bacteria can be enriched. Additionally, we could speculate that this bacteria is in charge of eliminating the growth of hyphae from fungi due to its ability to grow on fungal material, as shown by McKee and collaborators (McKee et al., 2019) when studying different roles of Chitinophaga pinensis. Perhaps, the role of this genus in Anoplotermes nest could be related to recycling of the dead material from fungi.
4.2.2 Amplicon sequencing: fungi

4.2.2.1 Whole set of samples

PCoA Emperor plots based on Bray Curtis diversity metric explains lower variability than in the case of 16S data. In overall, a few explanations emerge from this representation. We observe a grouping of samples according to feeding strategy, and geographical location, since most of the samples from each sampling site cluster in the proximity. There is still a clustering of samples according to genera, for Anoplotermes, Neocapritermes, Nasutitermes, while Cubitermes are divided by geographical location. ACE index calculations informed of a lack of differences in terms of OTU richness when we compared samples based on their geographical origin. Differences existed at the genus and feeding strategy levels. At this occasion, the feeding groups where we found differences were II vs III or IV, but also III vs IV. This indicates that the feeding strategy is capable of driving a differential richness in the fungal communities present in the termite nest. Furthermore, this seems to be reflected as well for the differences between genera, since Anoplotermes, Cubitermes, and Neocapritermes have each one differences towards Nasutitermes in ACE metrics. In terms of β-diversity, we found similar differences at statistical level as those observed in ACE metrics (same categories in feeding groups and genera), while in this case there were as well statistical differences in the pairwise comparisons at geographical level in CM-FG, and FG-MW. These results could indicate that the fungal communities are more susceptible of the geographical location, especially for the found differences between CM or MW vs FG at both sides of the Atlantic. The observed differences seem to be stronger in overall for fungi than for bacteria, and in this case driven both by the feeding strategies, especially wood vs soil feeding, and geographical origin (S. America vs Africa).

4.2.2.2 Samples with at least three representatives per genus

We also proceeded to eliminate all those nest microbial communities with less than three representatives per genus, and as it happened with the 16S data, we got very clear results. The PCoA representation provides clearer results than those for 16S, although with less explanation of variability. The factors that drive variability and stratification of samples are mainly feeding strategy, but especially genera. The representation of the latter stratifies the samples in four sectors. This is an indication that the fungal profile has a marked accent in the differentiation of fungal microbial communities. For the ACE metrics differences were reduced, while those in terms of genera remained. The β-diversity showed again these differences between group II and III or IV, between III and IV. We have here the case where the richness seems to be less affected by the feeding groups, but these strategies are driving differences in the composition of the communities, or how different are one from each other. Richness was affected by the genera, somewhat affected by the termite diet. The β-diversity was as well patent among the different genera. Because we also observed statistical significance of the location, we can conclude that differences are a complex product of the genera, feeding strategy, and ultimately also geographical origin with certain differences at both sides of the Atlantic, but not within Africa.

The PCA plot built in R (Fig. 6) was indicative of the stronger effect of genera and feeding strategy. We believe that genera were even stronger in sample stratification than type of diet, because even two genera in the same feeding group are clearly separated (Anoplotermes and Neocapritermes). Lastly, while existing, the geographical influence was weaker than the other two. These results are in overall agreement with our observations for the bacterial communities.

At the observed taxon profiles, we found that nests are dominated by the presence of OTUs belonging to the phylum Ascomycota. Within this phylum, the class Sordariomycetes and Dothideomycetes are the most abundant, as reported for nest walls from Cornitermes cumulans (Menezes et al., 2018). From the phylum Basidiomycota we found the appearance of two categories, the classes Exobasidiomycetes and Agaricomycetes. In keeping with our observations, these two phyla were the most frequent and represented in fungal isolates from O. formosanus (Xu et al., 2020). Similar results have been also reported by other recent studies (Větrovský et al., 2020)

Two phyla that belong to the kingdom Plantae appear as well in the first positions, but with distinct distribution among samples, down to the orders Chlorophyta and Magnoliophyta. We can explain the appearance of these organisms belonging to a different kingdom due to the reference used to annotate the ITS through QIIME2, in
this the Silva database which also contains Eukaryotic entries. Algae and other eukaryotes that contain 18S rRNA gene can be detected through amplification (Nakai et al., 2012; Banerji et al., 2018) even being used for classification (Buchheim et al., 2012; Mikhailyuk et al., 2020). We decided not to filtrate it as it could provide important information for our study.

There are some differences between samples belonging to the same genus but at different locations. According to what we described above for the different diversity metrics, it seems that geography is also a factor to have in account when considering differences between nest fungal microbial communities.

### 4.2.3 Metagenomic sequencing

There was a further selection on the samples used at metagenomic sequencing, as reported in the results section. For the case of *Anoplotermes*, all three samples belong to the same species: *Anoplotermes banksii*. For *Nasutitermes*, we had three representatives of *Nasutitermes lujae*, and one from *Nasutitermes* sp. Lastly, for *Cubitermes* we had representatives of four different species (Table 1). While it could be argued that we should have included only representatives of the same species on each sequenced group, that is very difficult when we talk about termites. There are constraints due to the inherent issues bringing the material in adequate conditions from the tropics to our home laboratory. Not every colony is able to arrive in good conditions, and when termites die, the nest is quickly invaded by pathogenic fungi. If we take in account the suboptimal conditions that we face in the field, it is clear that in-situ processing is not feasible. We hope that in future sequencing portable devices currently existing can be applied in our field, as it could help enormously in termite research. One aspect where it can help researchers, is the identification of colonies. The case of *Cubitermes* reflects those emerging issues with clarity. While all colonies and termites look the same at the bare or even taxonomically trained eye, it is very difficult to make accurate decisions. That is even an arduous task for termite taxonomists with good equipment in the laboratory, and it is not always possible to have these trained scientists on the field. Even at that level, our originally thought 4 equal representatives of *Cubitermes*, turned out to be different species. That was until certain extent happening with the genus *Nasutitermes*.

Additional limitations that drove our selection was the computational power, since the predicted amount of data and information was vast, and later on it turned out to be of a very notable magnitude. It also presents an enormous potential for discussion and description of novel genes of great interest for different biotechnological purposes. Computational restrictions have to be also taken in account, and while we could carry out the amplicon sequencing bioinformatic analyses without issues, the whole metagenomic data treatment and analysis required significant resources and computing time. This process delayed notably our pace especially when any step had to be repeated in order to introduce modifications that were not possible to foresee. We found that the SqueezeMeta pipeline leverages the process, and produces a vast amount of information that could be interpreted through SQM tools (Puente-Sánchez et al., 2020). Among other data management, SQM tools allow for sample sub-setting.

We have used different ways of presenting this data due to the massive amount of information that could be extracted from it. In general, we observe an equal trend for samples to stratify or group according to the feeding group and genera, with limited influence from the geographical location (Fig. 7). Despite small variability between nests from species belonging to the same genus, the differences are surprisingly stable, and even communities from soil feeding termite nests are having closer genomic and taxonomic composition at both sides of the Atlantic, as revealed at the NMDS analysis, which takes in account the full metagenome of each sample.

At the phylum level (Fig. 8) we still observe a comparable situation, where all *Nasutitermes* have a closer profile between themselves than those of *Cubitermes* or *Anoplotermes*. There are yet differences between samples within the same genera (with the exception of *Nasutitermes*). CubPI and CubSu both originating in CM have a closer profile than the other two from MW, while CubMu does still resemble the previous ones with incorporation of some OTUs not apparently present in CM (*Chloroflexi*). One of the samples from *Anoplotermes banksii* shows large differences with the other two samples from this species, probably owed to a worst state of that sample when arriving to our laboratory. The three samples from *Nasutitermes lujae* are very stable in terms of taxonomic representation. We can conclude that the taxonomic composition of the nests reflects the genera (where it seems to be stable), and especially eating habits.
In agreement with our previous taxonomic metabarcoding through 16S amplicon sequencing and ITS sequencing, we observe here similar trends. For instance, among fungi presence of *Ascomycota*, which was the most represented phylum, but in this case seems to have higher representation in wood feeders. For bacterial annotations, we find as well a very notable presence of *Actinobacteria* as reported in the first omic level. However, for 16S amplicon data we did not find that relevant proportion of *Proteobacteria* as that obtained by metagenomic sequencing. We found that the most abundant order in many of the sequenced soil feeders was *Streptosporangiales*, while in Nasutitermes they belong to class *Actinomycetia*, either to genera *Gryllotalpicola* in three cases or *Mycobacterium* in the last one.

But metagenomics has greater capabilities than just being able to provide another overview of taxonomical structure of the community. The part with most interest is the possibility to evaluate the functional potential, the metagenome.

Given a certain function, we could observe if at that level exists a differential composition in terms of taxa. We carried out a subset of functions (Puente-Sánchez et al., 2020) related to “Phenylalanine, tyrosine and tryptophan biosynthesis”. These three aromatic α-amino acids are involved in the metabolism of secondary metabolites (Parthasarathy et al., 2018). We found apparent differences according to the feeding strategies (Fig. 9). Once more, both soil feeding genera resemble each other, while the *Nasutitermes* keep a closer profile among them. This indicates that the differences are not only at the metabolomic level, or at the overall taxonomic structure of the community (both amplicon and metagenomic sequencing data). The differences go beyond and certain functions, as the one exemplified here, are carried out by different microorganisms. While nests from soil feeding termites display similar profile at both sides of the Atlantic, wood vs soil feeding are more dissimilar. Considering that, the geographical component seems to have a weaker contribution than the feeding habits in driving the functional differences among communities. There seems to be though a small reflection of a component that relates communities from the African environment, as we observe a group of unclassified *Actinobacteria* with similar proportion between *Nasutitermes* and *Cubitermes*. Our explanation here is that the feeding habits of termites and possibly some termite-active antimicrobial strategies are creating a set of conditions that allow the growth of certain type of genera, which will carry the required set of community functions, in this case the metabolism of aromatic amino-acids. We mentioned the termite active strategies, due to the presence of terpene as a selective metabolite in our study, with antimicrobial properties and secreted by *Nasutitermes*, as both evidenced in literature. The defensive or protective component of faeces was greatly reviewed by Cole, Ceja-Navarro, and Mikaelyan (2021). In our study, we infer that the combination of active strategies and passive constraints imposed by the feeding strategy create a specific growth-media for selected microorganisms. We could establish a parallelism with microbiology work in the laboratory, and state that termites with different feeding habits have formulated a growth media for their allied microbiome to thrive. A number of iterations of this process through trial and error in evolutionary time have probably fine-tuned these strategies, where the most successful in ensuring the nest and colony survival have been fixed. Until such an extent that microbial communities associated to two different genera of soil feeders resemble in structure at two sides of the Atlantic Ocean, where the different flora composition impose a different soil community (Lima-Perim et al., 2016). However, despite these stated differences even at short distance, the differences between the *Cubitermes* from CM and MW are minor, even when belonging to two clearly distinct biotopes. Thus, this is one more observation that helps us to associate a lower discrimination power to the geographical location.

Among the communities with higher presence for this function in *Nasutitermes*, the genus *Gryllotalpicola* has been found in larval galleries from the pine beetle *Monochamus alternatus* collected in the Chinese Jiangxi province (Chen et al., 2020), and was previously isolated from guts of the wood-feeding *Reticulitermes chinensis* Snyder, sampled in Wuhan, China (Fang et al., 2015). The interesting connection is that members from this genus have been found in insects with wood-feeding habits, at distant locations. *Streptosporangiales* is an order of bacteria that can obtain carbon from green waste (Cai et al., 2018), especially at the middle composting stage (Li et al., 2021). They can secrete cellulases and hemicellulases for lignocellulosic degradation, besides of being producers of antibiotics that can suppress pathogens (Gomes et al., 2017; Waglechner et al., 2019). These organisms are highly abundant in soil (McCarthy and Williams, 1992). Thus, since soil feeding genera use soil as their substrate, *Streptosporangiales* may originate at their surrounding environment and be enriched due to the substrate type and nutrients present. They could provide a cleaner environment at the termite nest due to production of antimicrobial compounds. This strategy, which we
cannot ascertain if it is an active strategy or derived from the termite activity, repeats itself at both sides of the Atlantic Ocean.
The genus *Microbacterium* (*Actinobacteria*) has been previously reported in *Mycocepurus smithii* ants (Kellner et al., 2015), a fungus-farming ant reported to acquire its microbiome from the surrounding environment. Bacteria from this genus have been found from deep-sea sediments (Jroundi et al., 2020) to rhizospheric soil (Wang et al., 2008), and it has been indicated that they have potential for pollution management (Zhang et al., 2017), especially for capabilities of phosphorus accumulation (Zhong et al., 2018).

*Microbacteriaceae* is a family of Gram-positive bacteria with members that have been reported among the dominant cellulase producing strains isolated from yellow stem borers (*Scirpophaga incertulas*) (Bashir et al., 2013), but much less proportion in the guts from termites at the same study identified as *Odontotermes hiananensis*, which were feeding on decomposed trees and leaf litter.

The LEFSE analysis on the ORFs annotated at PFAM, and have found that each of the genera is having a specific signature.

But because this information is vast, we carried out a different approach for gene characterization. Are these apparent differences in specific functions translated into differential content of certain genes beyond a LEFSE analysis? Has the microbial community gained or being enriched in certain functions? In transcriptomics, statistical testing is applied for finding differentially expressed genes (DGEs), what provides a clear idea of which genes are used in given conditions, being a molecular reflection of the phenotype. If this is applied in the case of metagenomics, the reflection is not exactly about what genes are being used, but more of the potential. However, this is not as static as if we were to evaluate the genome of a single organism. Being true that not all present genes are going to be expressed, we can infer that a certain function elevating its content in comparison with the same function in a second community, is relevant and this is not product of a random event. For instance, in the field of environmental antibiotic resistance qPCRs are routinely used to quantify the differences in gene content due to pharmaceutical pollution among different sites (González-Plaza et al., 2019; Milaković et al., 2019). We can then bring this knowledge to our current case and adopt the term “Differentially Present Genes” (DPGs).

While TPM is a normalization generally used in RNAseq experiments and its data treatment, according to Puente-Sánchez et al. (2020), it can be very useful in metagenomic experiments. In that regard, we found it as a very important value that allows for sample comparison and allowed us to get the first sample overview (Fig. 7). However, because we understand the limitations on TPM use (Zhao et al., 2020), we have carried out additional normalizations for comparison purposes, using procedures that use raw count values and are normalized differently, methodology taken from transcriptomics research (González Plaza et al. 2022, in preparation).

Testing for significant statistical differences will illustrate if a certain function is highly abundant in a nest microbial community, when compared with a second one. When carrying out these analyses, we can firstly establish a comparison between all samples based on the normalized count values for the PFAM annotated ORFs, which clearly separates *Nasutitermes* from the soil feeders (Fig. 11), again a division of microbial nest communities by the termite diet. We used for this analysis that reduced list of ORFs due to limitations in computing power. We additionally suffered the lack of a mixed approach between shotgun sequencing and long-read sequencing that could have notably improved the process and decreased the number of final contigs from where the ORFs where drawn.

When the analysis was brought to the DPGs content, we found that the differences between microbial communities in wood feeding termites nests and those from soil feeding termite nests are much stronger, given the amount of up-regulated (in terms of presence) genes, than those existing differences between the two soil feeding genera. Up-regulated in this study does not mean “higher expression” since we are working with DNA material, but it means a higher presence on that community.

The previously addressed recruitment of external or surrounding strains reported in literature, can be different in terms of specific strains, but they will be fundamentally equivalent in the functions that they carry out. Soil feeders “will recruit” microorganisms that are performing or carrying the same genes. We want to highlight that “recruit” may not be an active termite program, but a consequence of the complex cross-talk between active antimicrobial strategies, and the effect of diet.
Besides, the differences due to species appear to be of a lower entity, due to the relatively lower number of DPGs between both soil feeding genera. Among both soil feeder there were more than 2,000 DPGs, but those numbers rose to more than 14,000 DPGs when Nasutitermes was compared with either Anoplotermes banksii, or the Cubitermes genus.

We have carried out differential gene content analysis with the results obtained in specialized databases. We believe that this is important as a source of novel genes that could have a really important application in the biotechnological industry. Furthermore, our analyses have shown further a similarity among the genomic profiles between soil feeders and great differences to the wood feeder. Furthermore, the rationale for this additional genetic prospecting is that these databases can reflect different aspects of the microbial communities, which can be manifold and related with the metabolism as a cornerstone (Schmidt et al., 2019). In our study not only genes annotated in databases such as AnHyDeg or AromaDeg could serve as protection purposes when organic antimicrobial compounds such as terpenes are used, but also as part of a metabolism that could drive to social microbial interactions. For instance, antibiotic resistance (CARD database) can be used among members of the community to defend themselves against antibiotics, either from those produced by them, or from competing organisms (Scherlach and Hertweck, 2020). Additionally, at the usual amounts of antibiotic concentrations found at natural communities, these genes can serve for communication purposes rather than survival (Romero et al., 2011).

A first characterization of these results was to count for those known genes over a 90% identity threshold. We used a blasting approach with a lower identity threshold down to 40% but bit-score ≥ 50, based on Pearson (Pearson, 2013) to discover homologous genes. However, this threshold does not appear to be an adequate filtering value. Cases with 100% identity, had a length of alignment under 50 aminoacids. When using the maximum bit-score values, e-value was zero in 28 hits at the BacMet list, with alignment lengths varying from 605 – 1,118 amino-acids for those 28 genes, and identity ranging from 55% to 99%. Therefore, we suggest that the use of bit-score is a much better indication of the homology to described genes, rather than just using the percentage of identity.

We found that mostly known genes belong to the CAZy database, probably owed to the great interest of the cellulose and lignocellulose degrading field, where many sequencing approaches have been directed towards mining the microbial communities associated to insects that use cellulose. We believe that the other tools are also important, since for example the antibiotic resistance genes harboured in wildlife associated microbial communities has not been properly addressed, in comparison with the development of the field regarding man-made environments (Dolejska and Literak, 2019). Furthermore, the low number of described genes at BactiBase or AnhyDeg indicates the potential of these termite nest microbial communities for being a source of novel genes that can be used in fighting antimicrobial resistance (bacteriocins). The results from AnhyDeg mining can relate as well to a lower use of anaerobic degradation of hydrocarbons, since these nests environments should be mainly oxygenic. However, we carried out a thorough homogenization of the nest material, and that could have brought as well microorganisms thriving in deeper layers of biofilm structures or within the nest material.

When using three of these lists, we still found differences in gene content, again sharply evident between wood and soil feeders. The differences in the use of CAZy annotated genes were of small volume, and interestingly those were absent between both soil feeders. However, we must note that this could be a resemblance of the smaller number of genes that we used at this analysis due to computing limitations. If proportions are kept in regard to what we observe in the other comparisons, always higher number of DPGs between Nasutitermes and each of the soil feeders, it is logical to think that we did not have enough number of “events” for statistical differences to appear. Thus, while each of the lists referring to different type of microbial functional features, they maintain the observed differences when using the general PFAM annotated profile.

5 Conclusion

In our study we have evaluated a wide range of termite nests originating at the two sides of Atlantic Ocean, including nests belonging to species including soil and wood feeders. Sampling locations encompassed tropical rainforest and savannah habitats. We have faced constraints in our study that have limited the resolution, as some samples were unique representatives. Therefore, we have first characterized them, but for the main conclusions have been discarded. These limitations were imposed by the type of material and its source. We
believe that further studies with higher number of samples per genus will have equal statistical power and reach similar findings.

We firstly found differences in the overall metabolite profile that indicated a strong effect of the feeding habits of termites inhabiting those nests than the geographical location. Some of the most selective metabolites at VIP analysis were found to be reported in literature from cephalic glands in Nasutitermes wood feeders. These compounds may have an antimicrobial role in those nests. When we evaluated naphthalene content through a targeted approach, we were surprised to find much lower content than those previously reported in literature. The naphthalene presence seems to be prominent in few of the tested microbial communities, especially those related to Cubitermes and Nasutitermes genera.

The evaluation of the bacterial and fungal taxonomical profiles from those microbial communities had no indication of any effect of naphthalene on them, but rather pointed towards a stratification owed to the termite feeding habits. This is probably owed to the fact that nest material is composed of termite faeces and surrounding soil material. We could not assess differences due to the geographical location. While the differences between both tropical rainforest soils could be limited, the microbial communities’ differences between the savannah and tropical soil should be of a very strong nature. However, even the differences between the species in nests from soil feeding termites from both distinct locations in Africa were minor, probably owed to the effect of genus and again the diet.

We carried out a metagenomic profiling of two nest containing naphthalene from two distinct genera (soil vs wood feeding) and an additional soil feeding termite nest from French Guiana with no detected amount of naphthalene. The results from our analyses support the previous observations, where the major drive for genetic functional differences are the feeding habits of those termites inhabiting the studied nests. We could still establish genetic differences between both soil feeding termites, but these were of a more limited entity than those between each of the soil feeding communities and the wood feeding one. It is to be highlighted the vast genetic richness of functions and genes of biotechnological application, as observed by our mining approaches with specialized databases. This is only one example of the virtually unlimited genetic diversity that tropical rainforests harbour, and should bring attention on the necessity of preserving this resource that could be applied in fields such as bioremediation of organic compounds, or the improvement in the synthesis of biofuel or bioethanol.

Gene content analyses have provided an overview of the apparent changes between distinct feeding habits and even allow us to indicate that the feeding habit is a stronger component into the gene diversity and overall functions, rather than the geographical location. In that regard, the results at both sides of the Atlantic for soil feeding strategies (reflected in the microbial composition of the termite) indicate that the local surrounding microbial communities have less effect, although we in this study ignore their composition in comparison with the communities studied here. We found certain differences between the Cubitermes representatives at species level, although still clustering together. Even when half of the representatives originate from a humid tropical location, while the Malawi representatives belong to a savannah ecosystem. Those are closer to Anoploterms, despite the geographical distance. This is happening both at the gene content level, but also at the taxonomical composition level. The taxonomy composition arises from metagenomic sequencing, but also from 16S and ITS amplicon sequencing. We found clearer associations between the feeding habits and bacterial microbial communities than fungal communities.

We conclude that the differences among microbial communities at termite nests depend largely on the type of diet of the termite species, and not on the surrounding environment. This observation was already presented by Hu and collaborators (Hu et al., 2019), as authors suggested the linkage between diet and microbiome in termite guts. This observation is widespread with other animal models (Colman et al., 2012; David et al., 2013; Baker et al., 2014; Flint et al., 2015; Otani et al., 2019; Leite-Mondin et al., 2021). However, we state here that these differences go beyond the gut to be even reflected at the nest community level, consequence of the type of building strategy. The nest community structure-feeding habits linkage is here as thoroughly stated through three omics approaches.

A wide range of published studies have indicated that termites recruiting microbial members from the environmental communities. Our combined results are not contradicting that, but indicate that the genetic functions that microbial communities belonging to soil feeding genera are carrying out are similar across the Atlantic, even in two different habitats, savannah and rainforest. While exists a recruitment of bacterial strains
from the environment, what ultimately fixes the functional genomic content is the type of diet, and subsequently the nutrients that are available on each case. That observation has been repeated in our study at the three omic approaches. Further studies considering the different microbial niches (gut, environment, food substrate, and nest) could shed more light on the relationship between the microbial communities under influence of termites and those in the surrounding environment.

6 Acknowledgements
This project has been funded by grants IGA No. B_19_04 from the Faculty of Forestry and Wood Sciences (FFWS, Czech University of Life Sciences, Prague), and “Advanced research supporting the forestry and wood-processing sector’s adaptation to global change and the 4th industrial revolution,” OP RDE, Ministry of Education Youth and Sports of Czechia, Grant No. CZ.02.1.01/0.0/0.0/16_019/0000803. JGP thanks Dr. María Suárez at Wageningen University (The Netherlands) for support with scripts, to SurfBio Project (EU Horizon 2020 grant agreement No. 952379) for the training material provided at the Online Training School, to the EMBO Practical Course Microbial “Metagenomics: A 360º Approach”, and ICCRAM-Universidad de Burgos (Spain). We thank the EVA4.0 management team at FFWS.

7 Authors contribution
JGP and JŠ designed the study. JGP and JŠ participated in sample collection. JGP design the sampling strategy, isolated DNA and COII identification of many termites used in this study, processed nest material to powder, carried out all experiments regarding nest DNA isolation, sequencing data processing and analysis. JH performed metabolite profiling setup, experiments, data processing, and writing of the corresponding metabolite sections at the manuscript. JGP wrote the main body of the manuscript, JH and JŠ participated in the article writing and data interpretation.

8 Conflicts of interests
JGP discloses an International PCT Application No. PCT/CZ2022/050072, „A bacteriocin composition for coding Linocin-M18-like protein with antimicrobial activity and its usage”, filed on 05.08.2022 and related with the results of this research.

9 Ethical statement
Sampling was carried out following Nagoya guidelines for biodiversity sampling (Convention on Biological Diversity, 2010) with the required research permits to JŠ and other members of FFWS, and following ethical behaviour towards local communities.

References
Aguero, C. M., Eyer, P. A., Crippen, T. L., and Vargo, E. L. (2021). Reduced Environmental Microbial Diversity on the Cuticle and in the Galleries of a Subterranean Termite Compared to Surrounding Soil. Microb. Ecol. 81, 1054–1063. doi:10.1007/S00248-020-01664-W/FIGURES/5.
Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 48, D517–D525. doi:10.1093/nar/gkz935.
Allingham, S. M., and Harvey, M. (2013). Effects of Different Fire Regimes on Amphibian Communities in the Nyika National Park, Malawi. https://doi.org/10.5358/hsj.32.1 32, 1–8. doi:10.5358/HSJ.32.1.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.
Arora, R., Sharma, N. K., Kumar, S., and Sani, R. K. (2019). Lignocellulosic Ethanol: Feedstocks and Bioprocessing. Bioethanol Prod. from Food Crop., 165–185. doi:10.1016/B978-0-12-813766-6.00009-6.
Auer, L., Lazuka, A., Sillam-Dussès, D., Miambi, E., O’Donohue, M., and Hernandez-Raquet, G. (2017). Uncovering the potential of termite gut microbiome for lignocellulose bioconversion in anaerobic batch bioreactors. Front. Microbiol. 8, 2623. doi:10.3389/FMICB.2017.02623/BIBTEX.
Baker, R., Buckland, A., and Sheaves, M. (2014). Fish gut content analysis: robust measures of diet
composition. *Fish Fish.* 15, 170–177. doi:10.1111/FAF.12026.

Banerji, A., Bagley, M., Elk, M., Pilgrim, E., Martinson, J., and Santo Domingo, J. (2018). Spatial and temporal dynamics of a freshwater eukaryotic plankton community revealed via 18S rRNA gene metabarcoding. *Hydrobiologia* 818, 71–86. doi:10.1007/S10750-018-3593-0/FIGURES/4.

Bashir, Z., Kondapalli, V. K., Adlakha, N., Sharma, A., Bhatnagar, R. K., Chandel, G., et al. (2013). Diversity and functional significance of cellulolytic microbes living in termite, pill-bug and stem-borer guts. *Sci. Reports* 2013 31, 1–11. doi:10.1038/srep02558.

Beecmanns, S., and Xie, J. P. (2015). Glyoxylate Cycle. *Ref. Modul. Biomed. Sci.* doi:10.1016/B978-0-12-801238-3.02440-5.

Benjamino, J., and Graf, J. (2016). Characterization of the Core and Caste-Specific Microbiota in the Termite, Reticulitermes flavipes. *Front. Microbiol.* 7, 171. doi:10.3389/fmicb.2016.00171.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin. *Microbiome* 6, 1–17. doi:10.1186/s40168-018-0470-Z/TABLES/3.

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi:10.1038/nmeth.2276.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9.

Bowman, J. P. (2011). Protein-based analysis and other new and emerging non-nucleic acid based methods for tracing and investigating foodborne pathogens. *Tracing Pathog. Food Chain*, 292–341. doi:10.1533/9780857090508.3.292.

Bray, J. R., and Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* 27, 325–349. doi:10.2307/1942268.

Brown, C. T., Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., et al. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research* 2015 4900 4, 900. doi:10.12688/f1000research.6924.1.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. doi:10.48550/arxiv.1203.4802.

Brown, C. T., Scott, C., Crusoe, M., Sheneman, L., Rosenthal, J., and Howe, A. (2013). khmer-protocols documentation. figshare. Available at: http://dx.doi.org/10.6084/m9.figshare.878460.

Brune, A. (1998). Termite guts: The world’s smallest bioreactors. *Trends Biotechnol.* 16, 16–21. doi:10.1016/S0166-7799(97)01151-7.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2014 121 12, 59–60. doi:10.1038/nmeth.3176.

Buchheim, M. A., Sutherland, D. M., Schleicher, T., Förster, F., and Wolf, M. (2012). Phylogeny of Oedogoniales, Chaetophorales and Chaetopeltidales (Chlorophyceae): inferences from sequence-structure analysis of ITS2. *Ann. Bot.* 109, 109–116. doi:10.1093/aob/MCR275.

Cai, L., Gong, X., Sun, X., Li, S., and Yu, X. (2018). Comparison of chemical and microbiological changes during the aerobic composting and vermicomposting of green waste. *PLoS One* 13, e0207494. doi:10.1371/JOURNAL.PONE.0207494.

Callaghan, A. V, and Wawrik, B. AnHyDeg: a curated database of anaerobic hydrocarbon degradation genes. Available at: https://github.com/AnaerobesRock/AnHyDeg/tree/v1.0.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869.

Cao, X. L. (2012). A REVIEW RECENT DEVELOPMENT ON ANALYTICAL METHODS FOR DETERMINATION OF BISPHENOL A IN FOOD AND BIOLOGICAL SAMPLES. http://dx.doi.org/10.1080/10826076.2012.720325 35, 2795–2829. doi:10.1080/10826076.2012.720325.
Caspi, R., Billington, R., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46, D633–D639. doi:10.1093/NAR/GKX935.

Chen, H., Hao, D., Wei, Z., Wang, L., and Lin, T. (2020). Bacterial Communities Associated with the Pine Wilt Disease Insect Vector Monochamus alternatus (Coleoptera: Cerambycidae) during the Larvae and Pupae Stages. *Insects* 2020, Vol. 11, Page 376, 11, 376. doi:10.3390/INSECTS11060376.

Chen, J., Henderson, G., Grimm, C. C., Lloyd, S. W., and Laine, R. A. (1998a). Naphthalene in Formosan Subterranean Termite Carton Nests. *J. Agric. Food Chem.* 46, 2337–2339. doi:10.1021/jf9709717.

Chen, J., Henderson, G., Grimm, C. C., Lloyd, S. W., and Laine, R. A. (1998b). Termites fumigate their nests with naphthalene. *Nature* 392, 558–559. doi:10.1038/33305.

Chong, G. G., Huang, X. J., Di, J. H., Xu, D. Z., He, Y. C., Pei, Y. N., et al. (2018). Biodegradation of alkali lignin by a newly isolated Rhodococcus pyridinivorans CCZU-B16. *Bioprocess Biosyst. Eng.* 41, 501–510. doi:10.1007/S00449-017-1884-X/0/FIGURES/3.

Chouvent, T., Efstathion, C. A., Elliott, M. L., and Su, N.-Y. (2013). Extended disease resistance emerging from the faecal nest of a subterranean termite. *Proc. R. Soc. B Biol. Sci.* 280, 20131885. doi:10.1098/rspb.2013.1885.

Chouvent, T., Elliott, M. L., Šobotník, J., Efstathion, C. A., and Su, N.-Y. (2018). The Termite Fecal Nest: A Framework for the Opportunistic Acquisition of Beneficial Soil Streptomyces (Actinomycetales: Streptomycetaceae). *Environ. Entomol.* 47, 1431–1439. doi:10.1093/ee/nvy152.

Clarholm, M. (1981). Protozoan grazing of bacteria in soil—impact and importance. *Microb. Ecol.* 1981 74 7, 343–350. doi:10.1007/BF02341429.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/NAR/GKV1276.

Colas, F., Chanudet, V., Daufresne, M., Buchet, L., Vigouroux, R., Bonnet, A., et al. (2020). Spatial and Temporal Variability of Diffusive CO2 and CH4 Fluxes From the Amazonian Reservoir Petit-Saut (French Guiana) Reveals the Importance of Allochthonous Inputs for Long-Term C Emissions. *Global Biogeochem. Cycles* 34, e2020GB006602. doi:10.1029/2020GB006602.

Cole, M. E., Ceja-Navarro, J. A., and Mikaelyan, A. (2021). The power of poop: Defecation behaviors and social hygiene in insects. *PLOS Pathog.* 17, e1009964. doi:10.1371/JOURNAL.PPAT.1009964.

Colman, D. R., Toolson, E. C., and Takacs-Vesbach, C. D. (2012). Do diet and taxonomy influence insect gut bacterial communities? *Mol. Ecol.* 21, 5124–5137. doi:10.1111/j.1365-294X.2012.05752.X.

Convention on Biological Diversity (2010). The strategic plan for Biodiversity 2011-2020. in Tenth meeting of the Conference of the Parties to the Convention on Biological Diversity.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nat. 2013 505, 559–563. doi:10.1038/nature12820.

De Mello, A. P., De Moraes, M. M., Da Cámara, C. A. G., and Vasconcellos, A. (2021). Effect of Spatial Variation on Defensive Substances of Constrictotermes Cyphergaster Soldiers (Blattaria, Isoptera). *J. Chem. Ecol.* 47, 544–551. doi:10.1007/S10886-021-01271-0/FIGURES/2.

Dedysh, S. N., Kulichevskaya, I. S., Beletsky, A. V., Ivanova, A. A., Rijpstra, W. I. C., Damsté, J. S. S., et al. (2020). Lacipirellula parvula gen. nov., sp. nov., representing a lineage of planctomycetes widespread in low-oxygen habitats, description of the family Lacipirellulaceae fam. nov. and proposal of the orders Pirellulales ord. nov., Gemmatales ord. nov. and Isosphaerales ord. nov. *Syst. Appl. Microbiol.* 43, 126050. doi:10.1016/J.ISYAPM.2019.126050.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi:10.1128/AEM.03006-05.

Do, T. H., Nguyen, T. T., Nguyen, T. N., Le, Q. G., Nguyen, C., Kimura, K., et al. (2014). 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.* 118, 665–671. doi:10.1016/J.JBIOSC.2014.05.010.

Dolejska, M., and Literak, I. (2019). Wildlife is overlooked in the epidemiology of medically important...
antibiotic-resistant bacteria. *Antimicrob. Agents Chemother.* 63. doi:10.1128/AAC.01167-19.

Dolejšová, K., Krasulová, J., Kutalová, K., and Hanus, R. (2014). Chemical alarm in the termite Termitoteron planus (Rhinotermitidae). *J. Chem. Ecol.* 40, 1269–1276. doi:10.1007/S10886-014-0515-0/FIGURES/6.

Duarte, M., Jauregui, R., Vilchez-Vargas, R., Junca, H., and Pieper, D. H. (2014). AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database* 2014. doi:10.1093/database/bau118.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7. doi:10.1371/journal.pcbi.1002195.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi:10.1093/bioinformatics/btr381.

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.* 27, 105–117. doi:10.1016/j.tim.2018.11.003.

El-Naggar, N. E.-A. (2021). Streptomyces-based cell factories for production of biomolecules and bioactive metabolites. *Microb. Cell Factories Eng. Prod. Biomol.* 183–234. doi:10.1016/B978-0-12-821477-0.00011-8.

Enagbonma, B. J., Ajilogba, C. F., and Babalola, O. O. (2020). Metagenomic profiling of bacterial diversity and community structure in termite mounds and surrounding soils. *Arch. Microbiol.* 202, 2697–2709. doi:10.1007/S00203-020-01994-W/TABLES/3.

Erlacher, A., Cernava, T., Cardinale, M., Soh, J., Sensen, C. W., Grube, M., et al. (2015). Rhizobiales as functional and endosymbiotic members in the lichen symbiosis of Lobaria pulmonaria L. *Front. Microbiol.* 6, 53. doi:10.3389/FMICB.2015.00053/ABSTRACT.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi:10.1016/0006-3207(92)91201-3.

Faith, D. P., and Baker, A. M. (2006). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evol. Bioinform. Online* 2, 121. doi:10.1177/117693430602000007.

Fang, H., Lv, W., Huang, Z., Liu, S. J., and Yang, H. (2015). Gryllotalpicola reticulitermitis sp. nov., isolated from a termite gut. *Int. J. Syst. Evol. Microbiol.* 65, 85–89. doi:10.1099/IJS.0.062984-0/CITE/REFWORKS.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi:10.1093/NAR/GKV1344.

Flint, H. J., Duncan, S. H., Scott, K. P., and Louis, P. (2015). Links between diet, gut microbiota composition and gut metabolism. *Proc. Nutr. Soc.* 74, 13–22. doi:10.1017/S0029665114001463.

Funnicelli, M. I. G., Pinheiro, D. G., Gomes-Pepe, E. S., de Carvalho, L. A. L., Campanharo, J. C., Fernandes, C. C., et al. (2021). Metagenome-assembled genome of a Chitinophaga sp. and its potential in plant biomass degradation, as well of affiliated Pandoraea and Labrys species. *World J. Microbiol. Biotechnol.* 37, 1–17. doi:10.1007/S11274-021-03128-W/FIGURES/3.

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* 47, D195–D202. doi:10.1093/NAR/GKY1050.

Gershenzon, J., and Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nat. Chem. Biol.* 2007 37 3, 408–414. doi:10.1038/nchembio.2007.5.

Gomes, K. M., Duarte, R. S., and Bastos, M. do C. de F. (2017). Lantibiotics produced by Actinobacteria and their potential applications (A review). *Microbiol. (United Kingdom)* 163, 109–121. doi:10.1099/MIC.0.000397/CITE/REFWORKS.

González-Plaza, J. J., Blau, K., Milaković, M., Jurina, T., Smalla, K., and Udiković-Kolić, N. (2019). Antibiotic-manufacturing sites are hot-spots for the release and spread of antibiotic resistance genes and mobile genetic elements in receiving aquatic environments. *Environ. Int.* 130, 104735. doi:10.1016/J.ENVINT.2019.04.007.

Goodfellow, M., and Williams, S. T. (2003). ECOLOGY OF ACTINOMYCETES. https://doi.org/10.1146/annurev.mi.37.100183.001201 37, 189–216. doi:10.1146/ANNUREV.MI.37.100183.001201.

González-Plaza, J. J., Blau, K., Milaković, M., Jurina, T., Smalla, K., and Udiković-Kolić, N. (2019). Antibiotic-manufacturing sites are hot-spots for the release and spread of antibiotic resistance genes and mobile genetic elements in receiving aquatic environments. *Environ. Int.* 130, 104735. doi:10.1016/J.ENVINT.2019.04.007.
Griffiths, B. S., Bracewell, J. M., Robertson, G. W., and Bignell, D. E. (2013). Pyrolysis–mass spectrometry confirms enrichment of lignin in the faeces of a wood-feeding termite, Zootermopsis nevadensis and depletion of peptides in a soil-feeder, Cubitermes uugandensis. *Soil Biol. Biochem.* 57, 957–959. doi:10.1016/J.SOILBIO.2012.08.012.

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J., and Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.* 10, 22. doi:10.1186/1471-2180-10-22.

Hu, H., da Costa, R. R., Pilgaard, B., Schiøtt, M., Lange, L., and Poulsen, M. (2019). Fungiculture in Termites Is Associated with a Mycolytic Gut Bacterial Community. *mSphere* 4. doi:10.1128/MSHERE.00165-19/SUPPL_FILE/MSHERE.00165-19-ST003.XLSX.

Huerta-Cepas, J., Szklarczyk, D., Forsslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi:10.1093/NAR/GKV1248.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 1–11. doi:10.1186/1471-2105-11-119/TABLES/5.

Jordan, C. F., and Herrera, R. (2015). Tropical Rain Forests: Are Nutrients Really Critical? https://doi.org/10.1086/283696 17, 7–13. doi:10.1086/283696.

Jose, J. J. S., Montes, R., Stansly, P. A., and Bentley, B. L. (1989). Environmental Factors Related to the Occurrence of Mound-Building Nasute Termites in Trachypogon Savannas of the Orinoco Llanos. *Biotropica* 21, 353. doi:10.2307/2388286.

Jroundi, F., Martinez-Ruíz, F., Merroun, M. L., and Gonzalez-Muñoz, M. T. (2020). Exploring bacterial community composition in Mediterranean deep-sea sediments and their role in heavy metal accumulation. *Sci. Total Environ.* 712, 135660. doi:10.1016/J.SCITOTENV.2019.135660.

Kellner, K., Ishak, H. D., Linksvayer, T. A., and Mueller, U. G. (2015). Bacterial community composition and diversity in an ancestral ant fungus symbiosis. *FEMS Microbiol. Ecol.* 91, 73. doi:10.1093/FEMSEC/FIV073.

Knoben, W. J. M., Woods, R. A., and Freer, J. E. (2019). Global bimodal precipitation seasonality: A systematic overview. *Int. J. Climatol.* 39, 558–567. doi:10.1002/JOC.5786.

Krishanti, N. P. R. A., Zulfina, D., Wikantyoso, B., Zulfitri, A., and Yusuf, S. (2018). Antimicrobial Production by an Actinomycetes Isolated from The Termite Nest. *J. Trop. Life Sci.* 8, 279–288. doi:10.11594/JTLS.08.03.10.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/NMETH.1923.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi:10.1093/NAR/GKH152.

Legendre, P., and Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia* 2001 1292 129, 271–280. doi:10.1007/S004420100716.

Leite-Mondin, M., DiLegge, M. J., Manter, D. K., Weir, T. L., Silva-Filho, M. C., and Vivanco, J. M. (2021). The gut microbiota composition of Trichoplusia ni is altered by diet and may influence its polyphagous behavior. *Sci. Reports* 2021 111 11, 1–16. doi:10.1038/s41598-021-85057-0.
from azithromycin-manufacturing promotes macrolide-resistance gene propagation and induces spatial and seasonal bacterial community shifts in receiving river sediments. *Environ. Int.* doi:10.1016/j.envint.2018.12.050.

Mitaka, Y., Mori, N., and Matsuura, K. (2017). Multi-functional roles of a soldier-specific volatile as a worker arrestant, primer pheromone and an antimicrobial agent in a termite. *Proc. R. Soc. B Biol. Sci.* 284. doi:10.1098/RSPB.2017.1134.

Mizumoto, N., Gile, G. H., and Pratt, S. C. (2021). Behavioral Rules for Soil Excavation by Colony Founders and Workers in Termites. *Ann. Entomol. Soc. Am.* 114, 654–661. doi:10.1093/AESA/SAAA017.

Moore, B. P. (1968). Studies on the chemical composition and function of the cephalic gland secretion in Australian termites. *J. Insect Physiol.* 14, 33–39. doi:10.1016/0022-1910(68)90131-5.

Nakai, R., Abe, T., Baba, T., Imura, S., Kagoshima, H., Kanda, H., et al. (2012). Eukaryotic phylotypes in aquatic moss pillars inhabiting a freshwater lake in East Antarctica, based on 18S rRNA gene analysis. *Polar Biol.* 35, 1495–1504. doi:10.1007/S00300-012-1188-1/TABLES/2.

Ogola, H. J. O., Selvarajan, R., and Tekere, M. (2021). Local Geomorphological Gradients and Land Use Patterns Play Key Role on the Soil Bacterial Community Diversity and Dynamics in the Highly Endemic Indigenous Afrotemperate Coastal Scarp Forest Biome. *Front. Microbiol.* 12, 281. doi:10.3389/FMICB.2021.592725/BIBTEX.

Ohkuma, M. (2003). Termite symbiotic systems: Efficient bio-recycling of lignocellulose. *Appl. Microbiol. Biotechnol.* 61, 1–9. doi:10.1007/S00253-002-1189-Z/FIGURES/3.

Ohkuma, M., and Brune, A. (2010). “Diversity, Structure, and Evolution of the Termite Gut Microbial Community,” in *Biology of Termites: a Modern Synthesis* (Dordrecht), 413–438.

Olano, C., Lombó, F., Méndez, C., and Salas, J. A. (2008). Improving production of bioactive secondary metabolites in actinomycetes by metabolic engineering. *Metab. Eng.* 10, 281–292. doi:10.1016/J.YMBEN.2008.07.001.

Otani, S., Zhukova, M., Koné, N. A., da Costa, R. R., Mikaelyan, A., Sapountzis, P., et al. (2019). Gut microbial compositions mirror caste-specific diets in a major lineage of social insects. *Environ. Microbiol. Rep.* 11, 196–205. doi:10.1111/1758-2229.12728.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D. G. J. (2014). BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* 42, D737–D743. doi:10.1093/nar/gkt1252.

Pandit, P. D., Gulhane, M. K., Khardenavis, A. A., and Purohit, H. J. (2016). Mining of hemicellulose and lignin degrading genes from differentially enriched methane-producing microbial community. *Bioresour. Technol.* 216, 923–930. doi:10.1016/J.BIORTECH.2016.06.021.

Paredes-Sabja, D., Setlow, P., and Sarker, M. R. (2011). Germination of spores of Bacillales and Clostridiales species: mechanisms and proteins involved. *Trends Microbiol.* 19, 85–94. doi:10.1016/J.TIM.2010.10.004.

Parthasarathy, A., Cross, P. J., Dobson, R. C. J., Adams, L. E., Savka, M. A., and Hudson, A. O. (2018). A Three-Ring circus: Metabolism of the three proteogenic aromatic amino acids and their role in the health of plants and animals. *Front. Mol. Biosci.* 5, 29. doi:10.3389/FMOLB.2018.00029/BIBTEX.

Pearson, W. R. (2013). An Introduction to Sequence Similarity (“Homology”) Searching. *Curr. Protoc.* *Bioinformatics* 0 3. doi:10.1002/0471250953.BIO301S42.

Prestwich, G. D. (1979). Defence secretion of the black termite, Grallatotermes africanus (Termitidae, Nasutitermitinae). *Insect Biochem.* 9, 563–567. doi:10.1016/0020-1790(78)90093-3.

Prestwich, G. D., Jones, R. W., and Collins, M. S. (1981). Terpene biosynthesis by nasute termite soldiers (Isoptera: Nasutitermitinae). *Insect Biochem.* 11, 331–336. doi:10.1016/0020-1790(81)90011-1.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5. doi:10.1371/JOURNAL.PONE.0009490.

Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi:10.1093/BIOINFORMATICS/BTS252.

Puente-Sánchez, F., García-García, N., and Tamames, J. (2020). SQMtools: automated processing and visual analysis of ’omics data with R and anvi’o. *BMC Bioinformatics* 21, 358. doi:10.1186/s12859-020-03703-2.

Qiao, Q., Wang, F., Zhang, J., Chen, Y., Zhang, C., Liu, G., et al. (2017). The Variation in the Rhizosphere Microbiome of Cotton with Soil Type, Genotype and Developmental Stage. *Sci. Reports 2017* 7, 1–10.
Analysis Pipeline. *Front. Microbiol.* 9, 3349. doi:10.3389/fmicb.2018.03349.

Thong-On, A., Suzuki, K., Noda, S., Inoue, J. I., Kajiwara, S., and Ohkuma, M. (2012). Isolation and Characterization of Anaerobic Bacteria for Symbiotic Recycling of Uric Acid Nitrogen in the Gut of Various Termites. *Microbes Environ.* 27, 1202010349–1202010349. doi:10.1264/JSME2.ME11325.

Tokuda, G., Mikaelyan, A., Fukui, C., Matsuura, Y., Watanabe, H., Fujishima, M., et al. (2018). Fiber-associated spirochetes are major agents of hemicellulose degradation in the hindgut of wood-feeding higher termites. *Proc. Natl. Acad. Sci. U. S. A.* 115, E11996–E12004. doi:10.1073/PNAS.1810550115/SUPPL_FILE/PNAS.1810550115.SD04.XLSX.

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011). Next Generation Sequence Assembly with AMOS. *Curr. Protoc. Bioinforma.* 33, 11.8.1-11.8.18. doi:10.1002/0471250953.BI1108S33.

Tsimeli, K., Triantis, T. M., Dimotikali, D., and Hiskia, A. (2008). Development of a rapid and sensitive method for the simultaneous determination of 1,2-dibromoethane, 1,4-dichlorobenzene and naphthalene residues in honey using HS-SPME coupled with GC–MS. *Anal. Chim. Acta* 617, 64–71. doi:10.1016/J.ACA.2008.03.049.

Valterová, I., Vasičková, S., Buděšínský, M., and Vrkoč, J. (1986). Constituents of frontal gland secretion of peruvian termites *Nasutitermes ephratae*. *Collect. Czechoslov. Chem. Commun.* 51, 2884–2895. doi:10.1135/CCCC19862884.

van der Sande, M. T., Arets, E. J. M. M., Peña-Claros, M., Hoosbeek, M. R., Cáceres-Siani, Y., van der Hout, P., et al. (2018). Soil fertility and species traits, but not diversity, drive productivity and biomass stocks in a Guyanean tropical rainforest. *Funct. Ecol.* 32, 461–474. doi:10.1111/1365-2435.12968/SUPPINFO.

van Rensburg, J. J., Lin, H., Gao, X., Toh, E., Fortney, K. R., Ellinger, S., et al. (2015). The Human Skin Microbiome Associates with the Outcome of and Is Influenced by Bacterial Infection. *MBio* 6, e01315-15. doi:10.1128/mBio.01315-15.

van Velden, J. L., Wilson, K., Lindsey, P. A., McCallum, H., Moyo, B. H. Z., and Biggs, D. (2020). Bushmeat hunting and consumption is a pervasive issue in African savannas: insights from four protected areas in Malawi. *Biodivers. Conserv.* 29, 1443–1464. doi:10.1007/S10531-020-01944-4/FIGURES/4.

Větrovský, T., Soukup, P., Stiblík, P., Votýpková, K., Chakraborty, A., Larrañaga, I. O., et al. (2020). Termites host specific fungal communities that differ from those in their ambient environments. *Fungal Ecol.* 48, 100991. doi:10.1016/J.FUNECO.2020.100991.

Waglechner, N., McArthur, A. G., and Wright, G. D. (2019). Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat. Microbiol.* 2019 411 4, 1862–1871. doi:10.1038/s41564-019-0531-5.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi:10.1128/AEM.00062-07.

Wang, Y. X., Cai, M., Zhi, X. Y., Zhang, Y. Q., Tang, S. K., Xu, L. H., et al. (2008). Microlunatus aurantiacus sp. nov., a novel actinobacterium isolated from a rhizosphere soil sample. *Int. J. Syst. Evol. Microbiol.* 58, 1873–1877. doi:10.1099/ijs.0.65518-0/CITE/REFWORKS.

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., et al. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi:10.1111/1755-0998.13011.

Whitman, W. B., and Suzuki, K. (2015). Solirubrobacteres. *Berger’s Man. Syst. Archaea Bact.*, 1–3. doi:10.1002/9781118960608.OBM00025.

Wilcke, W., Amelung, W., Martius, C., Garcia, M. V. B., and Zech, W. (2000). Biological Sources of Polycyclic Aromatic Hydrocarbons (PAHs) in the Amazonian Rain Forest. *J. Plant Nutr. Soil Sci.* 163, 27–30. doi:10.1002/(SICI)1522-2624(200002)163:1<27::AID-JPLN27>3.0.CO;2-E.

Wilhelm, R. C., Singh, R., Eltis, L. D., and Mohn, W. W. (2018). Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing. *ISME J.* 2018 132 13, 413–429. doi:10.1038/s41396-018-0279-6.

Wiltz, B. A., Henderson, G., and Chen, J. (1998). Effect of Naphthalene, Butylated Hydroxytoluene, Dicoyl Phthalate, and Adipic Dioctyl Ester, Chemicals Found in the Nests of the Formosan Subterranean Termite
(Isoptera: Rhinotermitidae) on a Saprophytic Mucor sp. (Zygomycetes: Mucorales). *Environ. Entomol.* 27, 936–940. doi:10.1093/ee/27.4.936.

Wright, M. S., Lax, A. R., Henderson, G., and Chen, J. (2000). Growth response of *Metarhizium anisopliae* to two Formosan subterranean termite nest volatiles, naphthalene and fenchone. *Mycologia* 92, 42–45. doi:10.1080/00275514.2000.12061128.

Wu, Y. W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi:10.1093/BIOINFORMATICS/BTV638.

Xia, Y., Wang, Y., Fang, H. H. P., Jin, T., Zhong, H., and Zhang, T. (2014). Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. *Sci. Reports* 2014 41 4, 1–9. doi:10.1038/srep06708.

Xu, X., Shao, M., Yin, C., Mao, Z., Shi, J., Yu, X., et al. (2020). Diversity, Bacterial Symbionts, and Antimicrobial Potential of Termite-Associated Fungi. *Front. Microbiol.* 11, 300. doi:10.3389/FMICB.2020.00300/BIBTEX.

Yang, H., Schmitt-Wagner, D., Stingl, U., and Brune, A. (2005). Niche heterogeneity determines bacterial community structure in the termite gut (*Reticulitermes santonensis*). *Environ. Microbiol.* 7, 916–932. doi:10.1111/J.1462-2920.2005.00760.X.

Ye, Y., and Doak, T. G. (2009). A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. *PLOS Comput. Biol.* 5, e1000465. doi:10.1371/JOURNAL.PCBI.1000465.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi:10.1093/NAR/GKT1209.

Zapfack, L., and Engwald, S. (2008). Biodiversity and spatial distribution of vascular epiphytes in two biotopes of the Cameroonian semi-deciduous rain forest. *Plant Ecol.* 195, 117–130. doi:10.1007/S11258-007-9308-7/FIGURES/4.

Zhang, C., Yu, L., and Zhang, Y. (2017). [Research progress on the genus Microlunatus]. *Wei Sheng Wu Xue Bao* 57, 179–187. Available at: https://europepmc.org/article/med/29750480 [Accessed March 12, 2022].

Zhao, S., Ye, Z., and Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 26, 903–909. doi:10.1261/RNA.074922.120.

Zhong, C., Fu, J., Jiang, T., Zhang, C., and Cao, G. (2018). Polyphosphate metabolic gene expression analyses reveal mechanisms of phosphorus accumulation and release in *Microlunatus phosphovorus* strain JN459. *FEMS Microbiol. Lett.* 365, 34. doi:10.1093/FEMSLE/FNY034.
Figure 1. PCA plot using the complete untargeted profiling analysis. Samples are coloured according to feeding habits: green, soil; blue, wood
Figure 2. PLS-DA scores plot with classes wood feeders and soil feeders. The scores plot was further coloured according to A) type of food substrate, B) genera, and C) geographical origin.

Figure 3. Naphthalene analysis. A) Results from untargeted metabolite profiling. B) Results from targeted metabolite profiling, naphthalene concentration expressed in ng per gram of nest material. Naslu: Nasutitermes lujiae; Nassi: Nasutitermes similis; Cubpl: Cubitermes planifrons; Nassp: Nasutitermes sp.; Cubmu: Cubitermes muneris; Cubsu: Cubitermes sulcifrons; Cubug: Cubitermes aff. ugandensis; Conca: Constrictotermes cavifrons; Cubin: Cubitermes inclitus; CM: Cameroon; FG: French Guiana; MW: Malawi.
Figure 4. PCA analysis with 16S data after elimination of samples with less than 3 representatives per genus, and the tags from the negative sample control. A) Samples were subjected to Hellinger transformation, sample constrain and colour were “Genus”, shape by Feeding Strategy. B) Samples were subjected to Hellinger transformation, sample constrain was “Genus”, while sample colour was Geography, shape by Feeding Strategy. Genera indications were kept as shown in Fig. 4A.

Figure 5. LEfSe results. LDA Score (log 10) for each category per genus is shown.
Fig. 6. PCA analysis with ITS data after elimination of samples with less than 3 representatives per genus, and the tags from the negative sample control. A) Samples were subjected to Hellinger transformation, sample constrain and colour were “Genus”, shape by Feeding Strategy. B) Samples were subjected to Hellinger transformation, sample constrain was “Genus”, while sample colour was Geography, shape by Feeding Strategy. Genera indications were kept as shown in Fig. 6A.

Fig. 7. Non-metric multidimensional analysis considering all the samples sequenced in this study (metagenomic shotgun sequencing).
Fig. 8. Taxonomic representation at the phylum level in those samples submitted for metagenomic sequencing. Percentages were calculated in through SQMtools after SqueezeMeta pipeline. Phylum levels were obtained after annotation of sequences.
Fig. 9. Taxonomic representation at the genus level in those samples submitted for metagenomic sequencing, which belong to the subset “aromatic aa” (‘Phenylalanine, tyrosine and tryptophan biosynthesis’). Percentages were calculated in through SQMtools after SqueezeMeta pipeline. Genus levels were obtained after annotation of sequences.
Fig. 10. LEfSe analysis for the contigs annotated in PFAM at LDA Score (log₁₀) for each category per genus.
Fig. 11. Hierarchical clustering of samples. A) Euclidean distance. B) Hierarchical clustering represented by means of a heatmap.

Fig. 12. Volcano plot for the DPGs in the comparisons between microbial nests metagenomic samples, when using COG annotated genes as a reference. A) Cubitermes vs. Anoplotermes. B) Nasutitermes vs. Cubitermes. C) Nasutitermes vs Anoplotermes. Cub.: Cubitermes; Ban.: Anoplotermes; Nas.: Nasutitermes. Blue: genes with higher presence on the first component of the comparison (up-“regulated”); Red: genes with lower presence on the first component of the comparison (down-“regulated”).