Using ultraconserved elements to reconstruct the termite tree of life

Running title: UCEs in termites

Simon Hellemans[1,*], Menglin Wang[1], Nonno Hasegawa[1], Jan Šobotník[2], Rudolf H. Scheffrahn[3], Thomas Bourguignon[1,2,*]

[1]Okinawa Institute of Science & Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa 904-0495, Japan.

[2]Faculty of Tropical AgriScience, Czech University of Life Sciences, Kamýcka 129, 16521 Prague, Czech Republic.

[3]Fort Lauderdale Research and Education Center, Institute for Food and Agricultural Sciences, 3205 College Avenue, Davie, Florida 33314 USA.

*Corresponding authors: simon.hellemans@gmail.com (SH); Thomas.Bourguignon@oist.jp (TB)

ORCID references of authors

Simon Hellemans: 0000-0003-1266-9134

Menglin Wang: 0000-0003-2206-9503

Nonno Hasegawa: 0000-0001-5149-9129

Jan Šobotník: 0000-0002-8581-637X

Rudolf H. Scheffrahn: 0000-0002-6191-5963

Thomas Bourguignon: 0000-0002-4035-8977

Author contributions

SH and TB conceptualized the experiments. JS and RS collected the samples. MW performed lab experiments and generated data. SH, MW and NH analyzed the data. SH and TB wrote the manuscript.

Competing interests

We declare we have no competing interests.

Funding

This work was supported by Czech Science Foundation (project No. 15-07015Y), by the Internal Grant Agency of the Faculty of Tropical AgriSciences, CULS (20213112), by the Japan Society for the Promotion of Science JSPS through a postdoctoral fellowship for foreign

researchers awarded to SH (19F19819), and by the Okinawa Institute of Science and Technology through core unit funding.

Acknowledgments

We thank Yves Roisin, Saran Traoré and Guy Josens for providing specimens as well as Crystal Clitheroe for helping with DNA extraction and library preparation. We also thank the Scientific Computation and Data Analysis Section (SCDA) of the Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan, for providing access to the OIST computing cluster.

Electronic Supplementary Materials

Additional information is available for this manuscript.

Data Archiving

The data that support the findings of this study will be deposited in GenBank (see Supplementary Table S1 for details). Bait sequences, as well as the extracted UCE sequences, will be deposited on Dryad (Supplementary Data 1-7).

Abstract

The phylogenetic history of termites has been investigated using mitochondrial genomes and transcriptomes. However, both sets of markers have specific limitations. Mitochondrial genomes represent a single genetic marker likely to yield phylogenetic trees presenting incongruences with species trees, and transcriptomes can only be obtained from well-preserved samples. In contrast, ultraconserved elements (UCEs) include a great many independent markers that can be retrieved from poorly preserved samples. Here, we designed termite-specific baits targeting 50,616 UCE loci. We tested our UCE bait set on 42 samples of termites and three samples of *Cryptocercus*, for which we generated low-coverage highly-fragmented genome assemblies and successfully extracted *in silico* between 3,426 to 42,860 non-duplicated UCEs per sample. Our maximum likelihood phylogenetic tree, reconstructed using the 5,934 UCE loci retrieved from upward of 75% of samples, was congruent with transcriptome-based phylogenies, demonstrating that our UCE bait set is reliable and phylogenetically informative. Combined with non-destructive DNA extraction protocols, our UCE bait set provides the tool needed to carry out a global taxonomic revision of termites based on poorly preserved specimens such as old museum samples. The Termite UCE database is maintained at: https://github.com/oist/TER-UCE-DB/.

Key words: Data Mining; Isoptera; Phylogenomics; Mitochondrial genome; Nuclear genome; Taxonomy.

1. Introduction

Termites are the most ancient lineage of social insects, with a fossil record dating back to the Early Cretaceous, ~135 million years ago (Mya) (Thorne *et al.*, 2000; Grimaldi & Engel, 2005; Engel *et al.*, 2009). All modern termites share a common ancestor estimated at 140-150 Mya by time-calibrated phylogenetic trees (Bourguignon *et al.*, 2015; Bucek *et al.*, 2019). However, the bulk of the modern termite species diversity belongs to the Termitidae, a lineage that originated during the early Eocene, ~50 Mya, as indicated by time-calibrated phylogenetic trees (Bourguignon *et al.*, 2015, 2017; Bucek *et al.*, 2019) and by the fossil record (Engel *et al.*, 2011). While the backbone of the phylogenetic tree of termites is now largely resolved, most termite species are still awaiting to be placed on the tree of life.

Our understanding of the phylogenetic history of termites was mostly based on mitochondrial markers and fossils to calibrate estimated times of divergence until Bucek *et al.* (2019) published a phylogenetic tree of termites based on transcriptome data. The first phylogenetic trees of termites were based on a couple of PCR-amplified mitochondrial markers, sometimes combined with nuclear 18S or 28S sequences and/or morphological characters, which hardly contributed to any phylogenetic signal (e.g., Lo *et al.*, 2004; Inward *et al.*, 2007; Legendre *et al.*, 2008). These phylogenies provided a good overview of the relationships among the main termite lineages but lacked the robustness of phylogenetic trees inferred from full mitochondrial genomes (e.g., Cameron *et al.*, 2012; Bourguignon *et al.*, 2015, 2017). Full mitochondrial genomes, which became easy to sequence with the rise of second-generation sequencing technologies, resolve both shallow and deep divergences in the evolutionary history of termites and other insect lineages (Cameron, 2014), making them a marker of choice for phylogenetic reconstructions. However, mitochondrial genomes form a single marker, as all mitochondrial genes are linked and maternally inherited as a single package. Consequently, mitochondrial phylogenies are sometimes discordant with species phylogenies, especially for

closely related species and short internal branches that diverged in periods of time too brief for alleles to coalesce (Whitfield & Lockhart, 2007; Degnan & Rosenberg, 2009). One example of such discordance is provided by Sphaerotermitinae, the unambiguous sister group of Macrotermitinae according to transcriptomic data; but supported as sister to non-macrotermitine non-foraminitermitine Termitidae by mitochondrial genome phylogenies (Figure 1; Bucek *et al.*, 2019). Phylogenies based on multiple independent nuclear markers are needed to resolve the evolutionary history of organisms accurately.

Transcriptomes, the snapshot of genes expressed by an organism during tissue sampling, include many independent nuclear markers that can be used to build robust phylogenetic trees. Transcriptome-based phylogenies, reconstructed using up to ~4,000 single-copy orthologous nuclear genes spanning over 7.7 million nucleotide positions, have provided a robust picture of the ancient evolutionary history of termites (Bucek *et al.*, 2019). The sequencing of transcriptomes is now affordable, but, unfortunately, RNA is unstable and can only be extracted from samples that have been adequately preserved and stored, preventing the use of most samples collected before the genomic era began and making the approach impractical for large-scale studies. One alternative is to mine the conserved genetic markers present in whole-genome shotgun sequencing datasets, such as some datasets generated to sequence mitochondrial genomes.

Ultraconserved Elements (UCEs) are highly conserved nuclear regions whose functions remain largely unknown (Bejerano *et al.*, 2004; Faircloth *et al.*, 2012). UCEs are found across all regions of animal genomes, including the exonic, intronic, and intergenic regions. Phylogenetic trees inferred from UCEs have contributed to our understanding of the evolutionary history of various animal lineages spanning across the animal tree of life (e.g., Faircloth *et al.*, 2012; Ryu *et al.*, 2012; Smith *et al.*, 2014; White & Braun, 2019; Zhang *et al.*, 2019). Unlike transcriptomes, UCEs can readily be obtained from museum samples through

baiting conserved elements and their phylogenetically-informative flanking regions from fragmented genome assemblies (Blaimer *et al.*, 2016; Faircloth, 2017; Derkarabetian *et al.*, 2019). No UCE bait set has been designed for termites so far. We filled this gap as follows: (*i*) we designed *in silico* baits to capture UCEs; (*ii*) we compared phylogenetic trees reconstructed using all possible combinations of mitochondrial genomes, nuclear ribosomal RNA genes, and UCEs; and (*iii*) we showed that UCEs obtained from low-coverage shotgun genome assemblies are an alternative to transcriptome-based phylogenies for the reconstruction of multi-gene phylogenies. Finally, we set up a Termite UCE Database, thereby ensuring a long-term re-usability of published data.

2. Material and Methods

Biological samples and sequencing

We used sequence data from 42 samples of termites and three samples of *Cryptocercus*, the wood-feeding subsocial cockroach genus forming the sister group of termites. The sequencing data of 14 species were retrieved from previous studies (for details, see Table S1). The sequencing data from the remaining 31 species were obtained from samples preserved in 80% ethanol stored at room temperature or from samples preserved in RNA-later® and stored at temperatures ranging between -20°C and -80°C until DNA extraction. DNA was extracted using the DNeasy Blood & Tissue extraction kit (Qiagen). Libraries were prepared using the NEBNext® Ultra™ II FS DNA Library Preparation Kit (New England Biolabs) and the Unique Dual Indexing Kit (New England Biolabs), with reagent volumes reduced to one-fifteenth of recommended volumes. For samples preserved in 80% ethanol, libraries were prepared without the enzymatic fragmentation step as the DNA of these samples is typically highly fragmented.

Libraries were pooled in equimolar concentration and paired-end sequenced using the Illumina HiSeq X or Novaseq platforms at a read length of 150 bp.

UCE loci identification and in silico *bait design*

The identification of UCE loci was carried out using PHYLUCE *v*1.6.6 (Faircloth, 2016) following the recommendations of the tutorial (https://phyluce.readthedocs.io/en/stable/) and outlined workflow (Faircloth, 2017). Four publicly available genomes belonging to distantly related termite species were used to design baits: *Zootermopsis nevadensis* (Archotermopsidae), *Cryptotermes secundus* (Kalotermitidae), *Coptotermes formosanus* (Rhinotermitidae), and *Macrotermes natalensis* (Termitidae). Genome completeness was assessed using BUSCO *v*4.1.2 (Simão *et al.*, 2015) and QUAST *v*5.0.2 (Gurevich *et al.*, 2013). The genome of *M. natalensis* was chosen as the base genome for bait design due to its comparatively higher quality (for details, see Table S1).

Repetitive elements, retroelements, transposons, and small RNAs were masked from genome assemblies using RepeatMasker *v*4.1.1 (Smit *et al.*, 2015) with the command line "-species eukaryota -div 50". Assemblies were converted in the 2-bit format using the faToTwoBit tool of the BLAT suite of programs (Kent, 2002). We simulated 100 bp error-free paired-end sequencing reads from the three genome assemblies other than that of *M. natalensis* using art_illumina Q *v*2.5.8 (Huang *et al.*, 2012) with the command line "--fcov 2 --mflen 200 --sdev 150". In order to identify orthologous loci representing putative UCEs, the reads simulated from the three termite genome assemblies were mapped independently on the genome assembly of *M. natalensis* with a 0.05 substitution rate onto the base genome using the permissive raw-read aligner Stampy *v*1.0.32 (Lunter & Goodson, 2011). The three alignment maps were handled with SAMtools *v*1.9 (Li *et al.*, 2009) and converted into BED files with

bedtools *v*2.29.2 (Quinlan & Hall, 2010). In each BED file, putative conserved regions overlapping by at least one nucleotide were merged using bedtools. Conserved sequences shorter than 80 bp or containing over 25% of masked nucleotides were discarded using the phyluce program phyluce_probe_strip_masked_loci_from_set. The putative orthologous loci found across the four termite genomes were combined into a database using phyluce_probe_get_multi_merge_table (Supplementary Data 1). A total of 175,535 loci shared by the four termite genomes were identified and extracted using phyluce_probe_query_multi_merge_table and phyluce_probe_get_genome_sequences_from_bed, respectively. Extracted UCE sequences shorter than 180 bp were buffered to 180 bp by including 5' and 3' flanking regions in equal amounts with phyluce_probe_get_genome_sequences_from_bed (Supplementary Data 2).

A preliminary set of 120 bp baits was designed from the base genome of *M. natalensis* using phyluce_probe_get_tiled_probes. Baits targeted a region of 180 bp and overlapped in its center by 60 bp (at 2X tiling density). UCEs with ambiguous base calls and GC-content above 70% or below 30% were discarded from the bait set. Duplicates, defined as sequences having 50% identity over half of their length, were also removed from the bait set using LASTZ (Harris, 2007) implemented in the programs phyluce_probe_easy_lastz and phyluce_probe_remove_duplicate_hits_from_probes_using_lastz. In order to further identify and remove non-specific baits, we aligned the bait set (Supplementary Data 3) to the four genomes with phyluce_probe_run_multiple_lastzs_sqlite using a minimum identity threshold of 80% and minimum coverage of 83%. Sequences shorter than 180 bp were buffered to 180 bp by including 5' and 3' flanking regions in equal amounts and extracted from the alignments using phyluce_probe_slice_sequence_from_genomes. The loci shared by the four termite genomes were identified using phyluce_probe_get_multi_fasta_table and phyluce_probe_query_multi_fasta_table (Supplementary Data 4). The final UCE bait set was

designed with phyluce_probe_get_tiled_probe_from_multiple_inputs, and duplicates were removed using LASTZ as described above (397,910 baits targeting 50,616 loci; Supplementary Data 5).

Genome assembling and mining of phylogenetic markers

Adapters and low-quality bases were trimmed from raw reads using fastp *v*0.20.1 (Chen *et al.*, 2018). Trimmed reads were assembled using metaSPAdes *v*3.13 (Nurk *et al.*, 2017). The quality and completeness of assemblies were assessed with QUAST and BUSCO (for details, see Table S1). Mitochondrial genome scaffolds were identified in metaSPAdes assemblies and annotated using MitoFinder *v*1.4 (Allio *et al.*, 2020). Nuclear ribosomal RNA genes (5S, 5.8S, 18S, and 28S) were extracted from metaSPAdes assemblies using barrnap *v*0.9 (https://github.com/tseemann/barrnap). UCE loci were extracted from metaSPAdes assemblies using the final set of termite baits we designed and the PHYLUCE suite of programs with parameter values set as recommended in the tutorial and previously published studies (Faircloth *et al.*, 2015; Faircloth, 2017; Quattrini *et al.*, 2018). Briefly, baits were aligned to the metaSPAdes assemblies at a minimum similarity threshold of 50% with phyluce_probe_run_multiple_lastzs_sqlite. Sequences of the metaSPAdes assemblies matching baits were extracted with the flanking 200 bp situated at both the 5' and 3' ends using phyluce_probe_slice_sequence_from_genomes. Extracted sequences were mapped back to the baits using phyluce_assembly_match_contigs_to_probes with a minimum identity of 80% over 67% of bait length to remove duplicates and sequences matching multiple UCE loci (Supplementary Data 6; Contribution #1 to the Termite UCE Database available at: https://github.com/oist/TER-UCE-DB/).

Sequence alignment

The 13 mitochondrial protein-coding genes, two mitochondrial rRNAs genes, 22 mitochondrial tRNA genes, four nuclear rRNA genes, and UCEs were aligned using MAFFT *v*7.305 (Katoh & Standley, 2013). For mitochondrial protein-coding genes, we translated DNA sequences into the corresponding amino acid sequences using the transeq function from EMBOSS *v*6.6.0 (Rice *et al.*, 2000) and aligned protein sequences with MAFFT. Protein alignments were back-translated into codon alignments using PAL2NAL *v*14 (Suyama *et al.*, 2006). The other four types of genes, the mitochondrial rRNAs and tRNA genes, nuclear rRNA genes, and UCEs, were aligned as DNA sequences. UCE loci were aligned using MAFFT implemented in phyluce_align_seqcap_align, and internal trimming was performed under default parameters with Gblocks (Castresana, 2000; Talavera & Castresana, 2007) implemented in phyluce_align_get_gblocks_trimmed_alignments_from_untrimmed. Loci absent in more than 25% of taxa were filtered out with phyluce_align_get_only_loci_with_min_taxa. The final UCE supermatrix was exported using phyluce_align_format_nexus_files_for_raxml (Supplementary Data 7). Mitochondrial and nuclear gene alignments were concatenated using FASconCAT-G_v1.04.pl (Kück & Longo, 2014).

Phylogenetic analyses

We ran one separate phylogenetic analysis for the mitochondrial genome alignment, the nuclear rRNA alignment, and the UCE alignment. We also ran one phylogenetic analysis for the combined UCE and mitochondrial genome alignments. The mitochondrial genome alignment was separated into five distinct partitions: combined rRNAs, combined tRNAs, and combined first, second, and third codon positions of protein-coding genes. Nuclear rRNA gene and UCE alignments were given a single partition each. Phylogenetic trees were reconstructed in a

maximum likelihood (ML) framework using IQ-TREE *v*1.6.12 with 1,000 ultrafast bootstrap replicates (UFB) to assess branch supports (Nguyen *et al.*, 2015; Chernomor *et al.*, 2016; Hoang *et al.*, 2018). The best-fit nucleotide substitution model was selected for each partition with the Bayesian Information Criterion using ModelFinder implemented in IQ-TREE (Kalyaanamoorthy *et al.*, 2017). We calculated a global bootstrap support (GBS) value for each tree by averaging bootstrap values of all nodes. To assess concordance among UCEs, we carried out a multi-gene coalescence analysis with ASTRAL-III *v*5.7.7 (Zhang *et al.*, 2018) using individual gene trees obtained with IQ-TREE as described above. We allowed polytomies to reduce gene tree biases. Branch supports calculated with ASTRAL represent local posterior probabilities (LPP), which are based on gene tree quartet frequencies (Sayyari & Mirarab, 2016). Topological conflicts between individual gene trees and the ASTRAL species tree were assessed with PhyParts (Smith *et al.*, 2015) and visualized with PhyPartsPieCharts (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).

3. Results

In silico *data mining*

The mitochondrial genomes were retrieved from all 42 termite metaSPAdes assemblies. We also retrieved the four nuclear rRNA genes from 84% of the samples (see Table S1).

Our termite UCE bait set targeted a total of 50,616 loci. We extracted between 3,426 and 42,860 non-duplicated UCE loci from 42 termite metaSPAdes assemblies (Table S1). The number of non-duplicated UCE loci extracted from the assemblies of *Cryptocercus* roaches varied between 13,480 and 16,331. The final supermatrix, complete at 75% and containing loci present in at least 33 of the 45 taxa, was composed of 5,934 loci (Table S1) spanning over

1,677,394 nucleotide positions, 591,343 of which were parsimony-informative. The 45 taxa were represented by 939 to 5,928 loci.

Phylogenetic reconstructions

Many deep and shallow relationships within termites were poorly resolved by the nuclear rRNA phylogenetic tree (GBS = 72.12) (Figure S1). Because of its poor performance, we excluded the rRNA alignment from the analysis on combined marker classes. The phylogenetic reconstruction based on mitochondrial genomes resolved most relationships (GBS = 87), except for several nodes within the Serritermitidae, the Rhinotermitidae, and the Termitinae (Figure S2), as previously reported (Bourguignon *et al.*, 2015). The phylogenetic analysis performed exclusively on UCEs provided the most robust phylogenetic tree among the analyses performed on separate marker classes (Figure S3; GBS = 98.59). Combining UCEs and mitochondrial genomes marginally improved the resolution of the phylogenetic reconstruction (Figure 1; Figure S4; GBS = 99.02). The combined phylogenetic reconstruction resolved all nodes with high supports, except for the position of the rhinotermitid *Termitogeton planus* (UFB = 52 and 65, respectively). The phylogenetic analysis with ASTRAL revealed minimal discordance among the 5,934 UCE markers (Figure S5; final normalized quartet score of 0.89) (LPP = 1), except for five of the 42 nodes that presented conflicts among UCE markers. Within the Rhinotermitidae, the nodes corresponding to the split of *T. planus* and *Prorhinotermes simplex* displayed moderate concordance among UCE markers (LPP of 0.89 and 0.83, respectively). Within the Termitidae, the nodes corresponding to the split of *Neocapritermes utiariti*, *Pericapritermes* sp. 4, and of *Nitiditermes* + *Cavitermes* showed moderate to high levels of discordance (LPP of 0.66, 0.98, and 0.39, respectively). PhyParts analyses on a subset of 1,000 gene trees revealed some levels of topological discordances (Figure S6). Nodes with

discordance were mostly dominated by a plethora of topologies rather than by a single alternative and uninformative gene trees.

4. Discussion

We reconstructed phylogenetic trees for 42 species of termites and three species of *Cryptocercus* using three classes of markers: nuclear rRNA genes, mitochondrial genomes, and UCEs. The performance of the three types of phylogenetic markers decreased along the sequence: UCEs, mitochondrial genomes, and nuclear rRNA genes. The phylogenetic tree inferred from the latter class of markers, the nuclear rRNA genes, was poorly resolved and did not recover well-established relationships, such as the sister position of *Mastotermes* in respect to all other termites. The phylogenetic tree inferred from mitochondrial genomes was robust but failed to retrieve Sphaerotermitinae as sister to Macrotermitinae, as previously reported (Bourguignon *et al.*, 2015; Bucek *et al.*, 2019). The best phylogenetic tree was that reconstructed using UCEs (Figure 1). This phylogenetic tree was almost fully resolved and largely congruent with the phylogenetic trees inferred from transcriptomic data (Bucek *et al.*, 2019). Therefore, our results indicate that the termite UCE bait set we designed performs very well when reconstructing phylogenetic relationships among termite species. The addition of mitochondrial genome data, which, as UCEs, can be recovered from shotgun genome assemblies, slightly improved the global bootstrap support of the termite phylogenetic tree.

The analysis with ASTRAL revealed a few cases of discordance among UCE markers for lineages of Rhinotermitidae and Termitidae whose phylogenetic position was also unresolved with transcriptomic data (Bucek *et al.*, 2019). We used 5,934 UCE loci, a large number of markers that inevitably led to topological discordances between individual UCE trees and the species tree. These discordances are possibly caused by the lack of phylogenetic signal present in a single UCE marker and by population-level processes, such as incomplete

lineage sorting and introgression, which frequently occurs during the emergence of new lineages (Degnan & Rosenberg, 2009; Blom *et al.*, 2017; Parins-Fukuchi *et al.*, 2021). The actual relationships among termite lineages with unresolved positions remain unclear, possibly reflecting intricate evolutionary history that cannot be satisfactorily resolved by molecular phylogenetic techniques.

We ran our analyses on samples for which we generated low coverage genome assemblies. Notably, we retrieved numerous UCE sequences for all samples, including many that produced highly fragmented assemblies with low BUSCO scores (for details, see Table S1). All samples were accurately placed on the phylogenetic tree. Therefore, our UCE bait set has the potential to be used for mining phylogenetically informative genetic data from assemblies obtained from shotgun sequencing experiments. Used in combination with non-destructive DNA extraction protocols, our UCE baits could also be used to obtain sequence data from material that cannot be damaged, such as specimens from type series. This approach was successfully applied to centuries-old museum specimens such as Opiliones, carpenter bees, and weevils (Blaimer *et al.*, 2016; Van Dam *et al.*, 2017; Derkarabetian *et al.*, 2019). We recently sequenced the mitochondrial genome of a Syntype of the termite *Archotermopsis wroughtoni* collected at the end of the 19th century (Wang *et al.*, 2021). Termite UCEs could be extracted using the same procedures. Termite taxonomy, which is led by a shrinking pool of experts and is largely based on soldier and worker gut morphology, could benefit from the use of the many UCE markers designed in this study (Eggleton, 1999; Korb *et al.*, 2019). UCE baiting from whole-genome shotgun sequencing is the perfect tool to carry out a global taxonomic revision of termites.

References

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., Delsuc, F. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources* **20**: 892–905.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X., Brady, S.G. 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE* **11**: e0161531.

Blom, M.P.K., Bragg, J.G., Potter, S., Moritz, C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Systematic Biology* **66**: 352–366.

Bourguignon, T., Lo, N., Cameron, S.L., Šobotník, J., Hayashi, Y., Shigenobu, S., Watanabe, D., Roisin, Y., Miura, T., Evans, T.A. 2015. The evolutionary history of termites as inferred from 66 mitochondrial genomes. *Molecular Biology and Evolution* **32**: 406–421.

Bourguignon, T., Lo, N., Šobotník, J., Ho, S.Y.W., Iqbal, N., Coissac, É., Lee, M., Jendryka, M.M., Sillam-Dussès, D., Křížková, B., Roisin, Y., Evans, T.A. 2017. Mitochondrial phylogenomics resolves the global spread of higher termites, ecosystem engineers of the tropics. *Molecular Biology and Evolution* **34**: 589–597.

Bucek, A., Šobotník, J., He, S., Shi, M., McMahon, D.P., Holmes, E.C., Roisin, Y., Lo, N., Bourguignon, T. 2019. Evolution of termite symbiosis informed by transcriptome-based phylogenies. *Current Biology* **29**: 3728-3734.e4.

Cameron, S.L. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. *Annual Review of Entomology* **59**: 95–117.

Cameron, S.L., Lo, N., Bourguignon, T., Svenson, G.J., Evans, T.A. 2012. A mitochondrial genome phylogeny of termites (Blattodea: Termitoidae): robust support for interfamilial relationships and molecular synapomorphies define major clades. *Molecular Phylogenetics and Evolution* **65**: 163–173.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* **17**: 540–552.

Chen, S., Zhou, Y., Chen, Y., Gu, J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Chernomor, O., Von Haeseler, A., Minh, B.Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology* **65**: 997–1008.

Degnan, J.H., Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* **24**: 332–340.

Derkarabetian, S., Benavides, L.R., Giribet, G. 2019. Sequence capture phylogenomics of historical ethanol-preserved museum specimens: unlocking the rest of the vault. *Molecular Ecology Resources* **19**: 1531–1544.

Eggleton, P. 1999. Termite species description rates and the state of termite taxonomy. *Insectes Sociaux* **46**: 1–5.

Engel, M.S., Grimaldi, D.A., Krishna, K. 2009. Termites (Isoptera): their phylogeny, classification, and rise to ecological dominance. *American Museum Novitates* **3650**: 1–27.

Engel, M.S., Grimaldi, D.A., Nascimbene, P.C., Singh, H. 2011. The termites of Early Eocene

Cambay amber, with the earliest record of the Termitidae (Isoptera). *ZooKeys* **148**: 105–123.

Faircloth, B.C. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. *Methods in Ecology and Evolution* **8**: 1103–1112.

Faircloth, B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* **32**: 786–788.

Faircloth, B.C., Branstetter, M.G., White, N.D., Brady, S.G. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources* **15**: 489–501.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61**: 717–726.

Grimaldi, D.A., Engel, M.S. 2005. *Evolution of the Insects*. Cambridge University Press, Cambridge.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Harris, R.S. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**: 518–522.

Huang, W., Li, L., Myers, J.R., Marth, G.T. 2012. ART: a next-generation sequencing read

simulator. *Bioinformatics* **28**: 593–594.

Inward, D.J.G., Vogler, A.P., Eggleton, P. 2007. A comprehensive phylogenetic analysis of termites (Isoptera) illuminates key aspects of their evolutionary biology. *Molecular Phylogenetics and Evolution* **44**: 953–967.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Korb, J., Kasseney, B.D., Cakpo, Y.T., Casalla Daza, R.H., Gbenyedji, J.N.K.B., Ilboudo, M.E., Josens, G., Koné, N.A., Meusemann, K., Ndiaye, A.B., Okweche, S.I., Poulsen, M., Roisin, Y., Sankara, F. 2019. Termite taxonomy, challenges and prospects: West Africa, a case example. *Insects* **10**: 32.

Kück, P., Longo, G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* **11**: 81.

Legendre, F., Whiting, M.F., Bordereau, C., Cancello, E.M., Evans, T. a., Grandcolas, P. 2008. The phylogeny of termites (Dictyoptera: Isoptera) based on mitochondrial and nuclear markers: Implications for the evolution of the worker and pseudergate castes, and foraging behaviors. *Molecular Phylogenetics and Evolution* **48**: 615–627.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,

Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lo, N., Kitade, O., Miura, T., Constantino, R., Matsumoto, T. 2004. Molecular phylogeny of the Rhinotermitidae. *Insectes Sociaux* **51**: 365–371.

Lunter, G., Goodson, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21**: 936–939.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**: 824–834.

Parins-Fukuchi, C., Stull, G.W., Smith, S.A. 2021. Phylogenomic conflict coincides with rapid morphological innovation. *Proceedings of the National Academy of Sciences of the United States of America* **118**: e2023058118.

Quattrini, A.M., Faircloth, B.C., Dueñas, L.F., Bridge, T.C.L., Brugler, M.R., Calixto-Botía, I.F., DeLeo, D.M., Forêt, S., Herrera, S., Lee, S.M.Y., Miller, D.J., Prada, C., Rádis-Baptista, G., Ramírez-Portilla, C., Sánchez, J.A., Rodríguez, E., McFadden, C.S. 2018. Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: new approaches to long-standing problems. *Molecular Ecology Resources* **18**: 281–295.

Quinlan, A.R., Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rice, P., Longden, L., Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open

Software Suite. *Trends in Genetics* **16**: 276–277.

Ryu, T., Seridi, L., Ravasi, T. 2012. The evolution of ultraconserved elements with different phylogenetic origins. *BMC Evolutionary Biology* **12**: 236.

Sayyari, E., Mirarab, S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* **33**: 1654–1668.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Smit, A.F.A., Hubley, R., Green, P. 2015. RepeatMasker Open-4.0 2013-2015.

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology* **63**: 83–95.

Smith, S.A., Moore, M.J., Brown, J.W., Yang, Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15**: 150. BMC Evolutionary Biology.

Suyama, M., Torrents, D., Bork, P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**: W609–W612.

Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**: 564–577.

Thorne, B.L., Grimaldi, D.A., Krishna, K. 2000. Early fossil history of the termites. In:

Termites: Evolution, Sociality, Symbioses, Ecology (T. Abe, D. E. Bignell, & M. Higashi, eds), pp. 77–93. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Van Dam, M.H., Lam, A.W., Sagata, K., Gewa, B., Laufa, R., Balke, M., Faircloth, B.C., Riedel, A. 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. *PLoS ONE* **12**: e0188044.

Wang, M., Hellemans, S., Šobotník, J., Arora, J., Buček, A., Sillam-Dussès, D., Clitheroe, C., Lu, T., Lo, N., Engel, M.S., Roisin, Y., Evans, T.A., Bourguignon, T. 2021. Historical biogeography of early diverging termite lineages (Isoptera: Teletisoptera). *bioRxiv* 2021.12.02.471008.

White, N.D., Braun, M.J. 2019. Extracting phylogenetic signal from phylogenomic data: higher-level relationships of the nightbirds (Strisores). *Molecular Phylogenetics and Evolution* **141**: 106611.

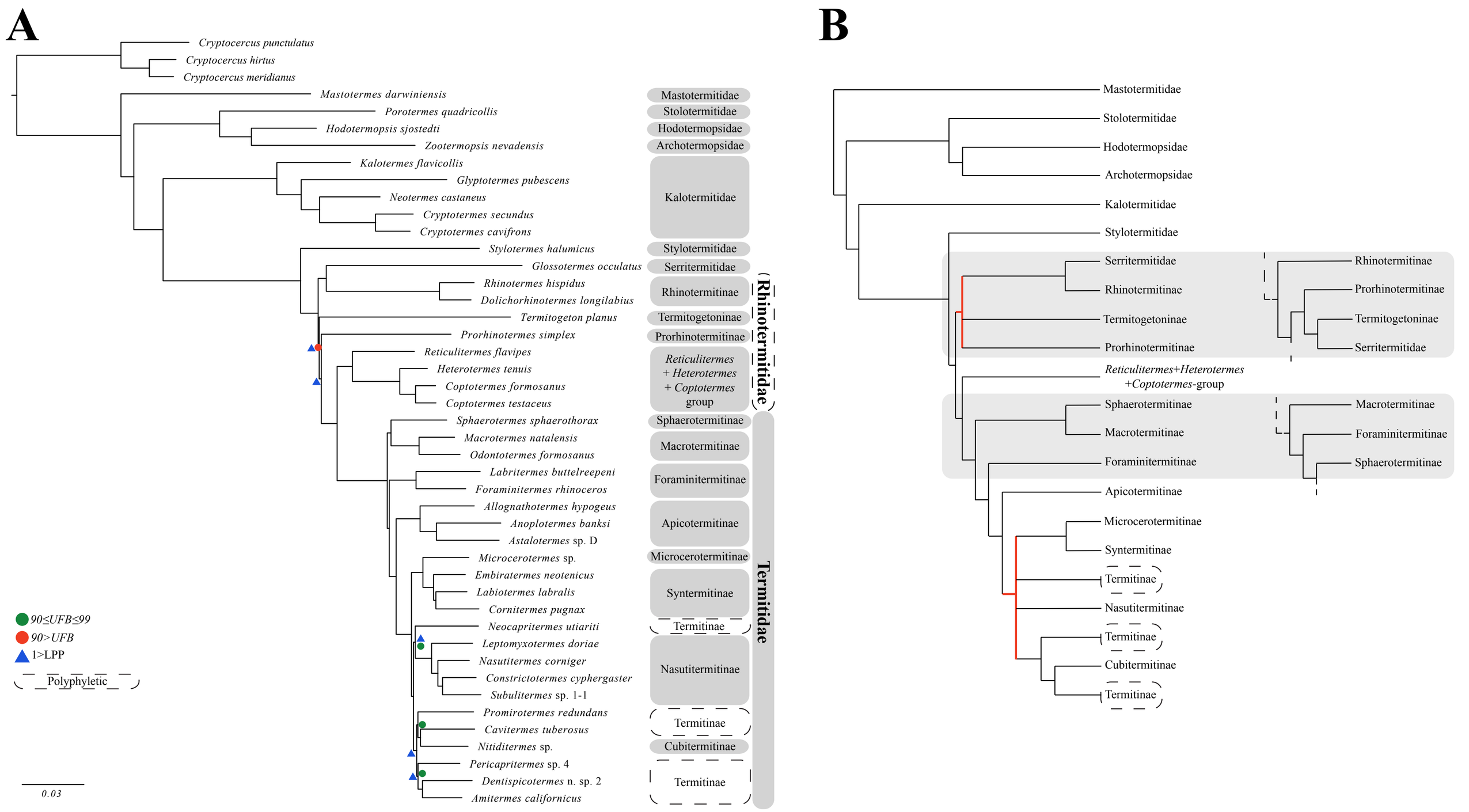
Whitfield, J.B., Lockhart, P.J. 2007. Deciphering ancient rapid radiations. *Trends in Ecology and Evolution* **22**: 258–265.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.

Zhang, Y.M., Williams, J.L., Lucky, A. 2019. Understanding UCEs: a comprehensive primer on using ultraconserved elements for arthropod phylogenomics. *Insect Systematics and Diversity* **3**: 3.

Figure captions

Figure 1: (**A**) Maximum likelihood phylogenetic tree of termites reconstructed with IQ-TREE using 5,934 UCE loci and complete mitochondrial genomes. Only UCE loci present in more than 75% of species were used. Support values are indicated for non-fully resolved nodes: ultrafast bootstrap (UFB; summarized from the phylogenetic trees reconstructed using UCE only and UCE + mitochondrial DNA displayed in Figures S3 and S4, respectively) and ASTRAL-III local posterior probabilities (LPP; phylogenetic tree displayed in Figure S5) values. (**B**) Family-level summary topology of termites supported by both UCEs (this study) and transcriptomic data (Bucek *et al*., 2019), with the indication of alternative topologies inferred from mitochondrial genome data alone (Bourguignon *et al*., 2015, 2017). Unsupported splits were summarized as polytomies (branches in red).


A
Cryptocercus punctulatus
Cryptocercus hirtus
Cryptocercus meridianus
Mastotermes darwiniensis
Porotermes quadricollis
Hodotermopsis sjostedti
Zootermopsis nevadensis
Kalotermes flavicollis
Glyptotermes pubescens
Neotermes castaneus
Cryptotermes secundus
Cryptotermes cavifrons
Stylotermes halumicus
Glossotermes occulatus
Rhinotermes hispidus
Dolichorhinotermes longilabius
Termitogeton planus
Prorhinotermes simplex
Reticulitermes flavipes
Heterotermes tenuis
Coptotermes formosanus
Coptotermes testaceus
Sphaerotermes sphaerothorax
Macrotermes natalensis
Odontotermes formosanus
Labritermes buttelreepeni
Foraminitermes rhinoceros
Allognathotermes hypogeus
Anoplotermes banksi
Astalotermes sp. D
Microcerotermes sp.
Embiratermes neotenicus
Labiotermes labralis
Cornitermes pugnax
Neocapritermes utiariti
Leptomyxotermes doriae
Nasutitermes corniger
Constrictotermes cyphergaster
Subulitermes sp. 1-1
Promirotermes redundans
Cavitermes tuberosus
Nitiditermes sp.
Pericapritermes sp. 4
Dentispicotermes n. sp. 2
Amitermes californicus
Mastotermitidae
Stolotermitidae
Hodotermopsidae
Archotermopsidae
Kalotermitidae
Stylotermitidae
Serritermitidae
Rhinotermitinae
Termitogetoninae
Prorhinotermitinae
Reticulitermes + Heterotermes + Coptotermes group
Sphaerotermitinae
Macrotermitinae
Foraminitermitinae
Apicotermitinae
Microcerotermitinae
Syntermitinae
Termitinae
Nasutitermitinae
Termitinae
Cubitermitinae
Termitinae
(Rhinotermitidae)
Termitidae
90≤UFB≤99
90>UFB
1>LPP
Polyphyletic
0.03
B
Mastotermitidae
Stolotermitidae
Hodotermopsidae
Archotermopsidae
Kalotermitidae
Stylotermitidae
Serritermitidae
Rhinotermitinae
Termitogetoninae
Prorhinotermitinae
Reticulitermes+Heterotermes +Coptotermes-group
Sphaerotermitinae
Macrotermitinae
Foraminitermitinae
Apicotermitinae
Microcerotermitinae
Syntermitinae
Termitinae
Nasutitermitinae
Termitinae
Cubitermitinae
Termitinae
Rhinotermitinae
Prorhinotermitinae
Termitogetoninae
Serritermitidae
Macrotermitinae
Foraminitermitinae
Sphaerotermitinae