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

researchers awarded to SH (19F19819), and 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 1-9).

**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 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

123 baiting conserved elements and their phylogenetically-informative flanking regions from
124 fragmented genome assemblies (Blaimer *et al.*, 2016; Faircloth, 2017; Derkarabetian *et al.*,
125 2019). No UCE bait set has been designed for termites so far. We filled this gap as follows: (*i*)
126 we designed *in silico* baits to capture UCEs; (*ii*) we compared phylogenetic trees reconstructed
127 using all possible combinations of mitochondrial genomes, nuclear ribosomal RNA genes, and
128 UCEs; and (*iii*) we showed that UCEs obtained from low-coverage shotgun genome assemblies
129 are an alternative to transcriptome-based phylogenies for the reconstruction of multi-gene
130 phylogenies. Finally, we set up a Termite UCE Database, thereby ensuring a long-term re-
131 usability of published data.

132

133 **2. Material and Methods**

134 *Biological samples and sequencing*

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

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

148

149 *UCE loci identification and* in silico *bait design*

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

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

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

181 A preliminary set of 120 bp baits was designed from the base genome of *M. natalensis*
182 using phyluce_probe_get_tiled_probes. Baits targeted a region of 180 bp and overlapped in its
183 center by 60 bp (at 2X tiling density). UCEs with ambiguous base calls and GC-content above
184 70% or below 30% were discarded from the bait set. Duplicates, defined as sequences having
185 50% identity over half of their length, were also removed from the bait set using LASTZ (Harris,
186 2007) implemented in the programs phyluce_probe_easy_lastz and
187 phyluce_probe_remove_duplicate_hits_from_probes_using_lastz. In order to further identify
188 and remove non-specific baits, we aligned the bait set (Supplementary Data 3) to the four
189 genomes with phyluce_probe_run_multiple_lastzs_sqlite using a minimum identity threshold
190 of 80% and minimum coverage of 83%. Sequences shorter than 180 bp were buffered to 180
191 bp by including 5' and 3' flanking regions in equal amounts and extracted from the alignments
192 using phyluce_probe_slice_sequence_from_genomes. The loci shared by the four termite
193 genomes were identified using phyluce_probe_get_multi_fasta_table and
194 phyluce_probe_query_multi_fasta_table (Supplementary Data 4). The final UCE bait set was

195 designed with phyluce_probe_get_tiled_probe_from_multiple_inputs, and duplicates were
196 removed using LASTZ as described above (397,910 baits targeting 50,616 loci; Supplementary
197 Data 5). This final set of loci was tentatively annotated using the GFF file (NCBI Annotation
198 Release 100) from the *Z. nevadensis* genome assembly (GCF_000696155).

199

200 *Genome assembling and mining of phylogenetic markers*

201 Adapters and low-quality bases were trimmed from raw reads using fastp *v*0.20.1 (Chen *et al.*,
202 2018), resulting in a total of 4.55 to 448.64 millions of paired-end reads per sample (for details,
203 see Table S1). Trimmed reads were assembled using metaSPAdes *v*3.13 (Nurk *et al.*, 2017).
204 The quality and completeness of assemblies were assessed with QUAST and BUSCO (Table
205 S1). Mitochondrial genome scaffolds were identified in metaSPAdes assemblies and annotated
206 using MitoFinder *v*1.4 (Allio *et al.*, 2020). Nuclear ribosomal RNA genes (5S, 5.8S, 18S, and
207 28S) were extracted from metaSPAdes assemblies using barrnap *v*0.9
208 (https://github.com/tseemann/barrnap). UCE loci were extracted from metaSPAdes assemblies
209 using the final set of termite baits we designed and the PHYLUCE suite of programs with
210 parameter values set as recommended in the tutorial and previously published studies (Faircloth
211 *et al.*, 2015; Faircloth, 2017; Quattrini *et al.*, 2018). Briefly, baits were aligned to the
212 metaSPAdes assemblies at a minimum similarity threshold of 50% with
213 phyluce_probe_run_multiple_lastzs_sqlite. Sequences of the metaSPAdes assemblies
214 matching baits were extracted with the flanking 200 bp situated at both the 5' and 3' ends using
215 phyluce_probe_slice_sequence_from_genomes. Extracted sequences were mapped back to the
216 baits using phyluce_assembly_match_contigs_to_probes with a minimum identity of 80% over
217 67% of bait length to remove duplicates and sequences matching multiple UCE loci
218 (Supplementary Data 6; Contribution #1 to the Termite UCE Database available at:

219 https://github.com/oist/TER-UCE-DB/). The average coverage of UCE loci per sample was
220 obtained using the mapping workflow of PHYLUCE *v*1.7.1.
221

222 *Sequence alignment*

223 The 13 mitochondrial protein-coding genes, two mitochondrial rRNA genes, 22 mitochondrial
224 tRNA genes, four nuclear rRNA genes, and UCEs were aligned using MAFFT *v*7.305 (Katoh
225 & Standley, 2013). For mitochondrial protein-coding genes, we translated DNA sequences into
226 the corresponding amino acid sequences using the transeq function from EMBOSS *v*6.6.0 (Rice
227 *et al.*, 2000) and aligned protein sequences with MAFFT. Protein alignments were back-
228 translated into codon alignments using PAL2NAL *v*14 (Suyama *et al.*, 2006). The other four
229 types of genes, the mitochondrial rRNA and tRNA genes, nuclear rRNA genes, and UCEs,
230 were aligned as DNA sequences. UCE loci were aligned using MAFFT implemented in
231 phyluce_align_seqcap_align, and internal trimming was performed under default parameters
232 with Gblocks (Castresana, 2000; Talavera & Castresana, 2007) implemented in
233 phyluce_align_get_gblocks_trimmed_alignments_from_untrimmed. Loci absent in more than
234 25% of taxa were filtered out with phyluce_align_get_only_loci_with_min_taxa. The final
235 UCE supermatrix was exported using phyluce_align_format_nexus_files_for_raxml
236 (Supplementary Data 7: alignments; Supplementary Data 8: corresponding reduced bait set).
237 Mitochondrial and nuclear gene alignments were concatenated using FASconCAT-G_v1.04.pl
238 (Kück & Longo, 2014).

239

240 *Phylogenetic analyses*

241 We ran one separate phylogenetic analysis for the mitochondrial genome alignment, the nuclear
242 rRNA alignment, and the UCE alignment. We also ran one phylogenetic analysis for the

243 combined UCE and mitochondrial genome alignments. The mitochondrial genome alignment
244 was separated into five distinct partitions: combined rRNAs, combined tRNAs, and combined
245 first, second, and third codon positions of protein-coding genes. Nuclear rRNA gene and UCE
246 alignments were given a single partition each. Phylogenetic trees were reconstructed in a
247 maximum likelihood (ML) framework using IQ-TREE *v*1.6.12 with 1,000 ultrafast bootstrap
248 replicates (UFB) to assess branch supports (Nguyen *et al.*, 2015; Chernomor *et al.*, 2016; Hoang
249 *et al.*, 2018). The best-fit nucleotide substitution model was selected for each partition with the
250 Bayesian Information Criterion using ModelFinder implemented in IQ-TREE
251 (Kalyaanamoorthy *et al.*, 2017). We calculated a global bootstrap support (GBS) value for each
252 tree by averaging bootstrap values of all nodes. To assess concordance among UCEs, we carried
253 out a multi-gene coalescence analysis with ASTRAL-III *v*5.7.7 (Zhang *et al.*, 2018) using
254 individual gene trees obtained with IQ-TREE as described above. We allowed polytomies to
255 reduce gene tree biases. Branch supports calculated with ASTRAL represent local posterior
256 probabilities (LPP), which are based on gene tree quartet frequencies (Sayyari & Mirarab, 2016).
257 Topological conflicts between individual gene trees and the ASTRAL species tree were
258 assessed with PhyParts (Smith *et al.*, 2015) and visualized with PhyPartsPieCharts
259 (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).
260

261 **3. Results**

262 In silico *data mining*

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

265 Our termite UCE bait set targeted a total of 50,616 loci distributed across 1,094 scaffolds
266 of the *Z. nevadensis* genome assembly (GCF_000696155). Of these 50,616 loci, 6,325 (12.5 %)

267 were found in the non-coding regions of 787 scaffolds (Supplementary Data 9). The remaining
268 44,291 loci were found in the coding regions (33,932 in exons) of 8,636 annotated genes
269 distributed across 886 scaffolds. Of the 8,636 genes, 6,538 (75.7 %) contained more than one
270 ultraconserved loci.

271 From the 50,616 targeted loci, we extracted between 3,426 and 42,860 non-duplicated
272 UCE loci from 42 termite metaSPAdes assemblies (Table S1). The number of non-duplicated
273 UCE loci extracted from the assemblies of *Cryptocercus* roaches varied between 13,480 and
274 16,331. The average coverage of UCE loci per sample was between 8.38 to 134.71x (Table S1).
275 The final supermatrix, complete at 75% and containing loci present in at least 33 of the 45 taxa,
276 was composed of 5,934 loci spanning over 1,677,394 nucleotide positions, 591,343 of which
277 were parsimony-informative. The 45 taxa were represented by 939 to 5,928 loci.

278

279 *Phylogenetic reconstructions*

280 Many deep and shallow relationships within termites were poorly resolved by the nuclear rRNA
281 phylogenetic tree (GBS = 72.12) (Figure S1). Because of its poor performance, we excluded
282 the rRNA alignment from the analysis performed on combined marker classes. The
283 phylogenetic reconstruction based on mitochondrial genomes resolved most relationships (GBS
284 = 87), except for several nodes within the Serritermitidae, the Rhinotermitidae, and the
285 Termitinae (Figure S2), as previously reported (Bourguignon *et al.*, 2015). The phylogenetic
286 analysis performed exclusively on UCEs provided the most robust phylogenetic tree among the
287 analyses performed on separate marker classes (Figure S3; GBS = 98.59). Combining UCEs
288 and mitochondrial genomes marginally improved the resolution of the phylogenetic
289 reconstruction (Figure 1; Figure S4; GBS = 99.02). The combined phylogenetic reconstruction
290 resolved all nodes with high supports, except for the position of the rhinotermitid *Termitogeton*

291 *planus* (UFB = 52 and 65, respectively). The phylogenetic analysis with ASTRAL revealed
292 minimal discordance among the 5,934 UCE markers (Figure S5; final normalized quartet score
293 of 0.89) (LPP = 1), except for five of the 42 nodes that presented conflicts among UCE markers.
294 Within the Rhinotermitidae, the nodes corresponding to the split of *T. planus* and
295 *Prorhinotermes simplex* displayed moderate concordance among UCE markers (LPP of 0.89
296 and 0.83, respectively). Within the Termitidae, the nodes corresponding to the split of
297 *Neocapritermes utiariti*, *Pericapritermes* sp. 4, and *Nitiditermes* + *Cavitermes* showed
298 moderate to high levels of discordance (LPP of 0.66, 0.98, and 0.39, respectively). PhyParts
299 analyses on a subset of 1,000 gene trees revealed some levels of topological discordances
300 (Figure S6). Nodes with discordance were mostly dominated by a plethora of topologies rather
301 than by a single alternative and uninformative gene trees.

302

303 **4. Discussion**

304 We reconstructed phylogenetic trees for 42 species of termites and three species of
305 *Cryptocercus* using three classes of markers: nuclear rRNA genes, mitochondrial genomes, and
306 UCEs. The performance of the three types of phylogenetic markers decreased along the
307 sequence: UCEs, mitochondrial genomes, and nuclear rRNA genes. The phylogenetic tree
308 inferred from the latter class of markers, the nuclear rRNA genes, was poorly resolved and did
309 not recover well-established relationships, such as the sister position of *Mastotermes* in respect
310 to all other termites. The phylogenetic tree inferred from mitochondrial genomes was robust
311 but failed to retrieve Sphaerotermitinae as sister to Macrotermitinae, as previously reported
312 (Bourguignon *et al.*, 2015; Bucek *et al.*, 2019). The best phylogenetic tree was that
313 reconstructed using the 75%-occupancy matrix comprised of 5,934 UCE loci (Figure 1). This
314 phylogenetic tree was almost fully resolved and largely congruent with the phylogenetic trees

315 inferred from transcriptomic data (Bucek *et al.*, 2019). Therefore, our results indicate that the
316 termite UCE bait set we designed performs very well when reconstructing phylogenetic
317 relationships among termite species. The addition of mitochondrial genome data, which, as
318 UCEs, can be recovered from shotgun genome assemblies, slightly improved the global
319 bootstrap support of the termite phylogenetic tree.

320 The analysis with ASTRAL revealed a few cases of discordance among UCE markers
321 for lineages of Rhinotermitidae and Termitidae whose phylogenetic position was also
322 unresolved with transcriptomic data (Bucek *et al.*, 2019). We used 5,934 UCE loci, a large
323 number of markers that inevitably led to topological discordances between individual UCE
324 trees and the species tree. These discordances are possibly caused by the lack of phylogenetic
325 signal present in a single UCE marker and by population-level processes, such as incomplete
326 lineage sorting and introgression, which frequently occurs during the emergence of new
327 lineages (Degnan & Rosenberg, 2009; Blom *et al.*, 2017; Parins-Fukuchi *et al.*, 2021). While
328 recent studies indicated that the phylogenetic resolution can be improved by removing loci
329 localized in the exons of a single gene (Hedin *et al.*, 2019; Van Dam *et al.*, 2021), our unfiltered
330 analysis performed as well as analyses performed on transcriptome-based data. The actual
331 relationships among termite lineages with unresolved positions remain unclear, possibly
332 reflecting intricate evolutionary history that cannot be satisfactorily resolved by molecular
333 phylogenetic techniques.

334 We ran our analyses on samples for which we generated low coverage genome
335 assemblies. The final bait set targeting a total of 50,616 orthologous loci was obtained from
336 four termite genomes, belonging to four families. This high number of UCE loci is comparable
337 to that found in other groups of arthropods (e.g., Buenaventura *et al.*, 2021). Overall, 67% of
338 loci in the termite bait set were found in exonic regions, further indicating that the UCEs of
339 arthropods are mostly found in exons (Hedin *et al.*, 2019; Van Dam *et al.*, 2021). We retrieved

340 numerous UCE sequences for all samples, including many that produced highly fragmented
341 assemblies with low BUSCO scores (for details, see Table S1). All samples were accurately
342 placed on the phylogenetic tree reconstructed with the 5,934 loci present in the 75%-occupancy
343 supermatrix. Therefore, our UCE bait set has the potential to be used for mining
344 phylogenetically informative genetic data from assemblies obtained from shotgun sequencing
345 experiments. We established a centralized termite UCE database (https://github.com/oist/TER-
346 UCE-DB/), which we plan to use to reference all UCE data extracted with the presently
347 designed bait set, thereby ensuring the long-term re-usability of the available data.

348 While producing low-coverage genomes is more costly than targeting UCEs through
349 synthetized baits, low-coverage genome data can be used to investigate a broad range of
350 questions in addition to phylogenetic reconstruction. Used in combination with non-destructive
351 DNA extraction protocols, our UCE baits could also be used to obtain sequence data from
352 material that cannot be damaged, such as specimens from type series. This approach was
353 successfully applied to centuries-old museum specimens of Opiliones, carpenter bees, and
354 weevils (Blaimer *et al.*, 2016; Van Dam *et al.*, 2017; Derkarabetian *et al.*, 2019). We recently
355 obtained the full mitogenome of a Syntype of the termite *Archotermopsis wroughtoni* collected
356 at the end of the 19$^{th}$ century using shotgun sequencing data (Wang *et al.*, 2021). Termite UCEs
357 could be extracted using the same procedures. Termite taxonomy, which is led by a shrinking
358 pool of experts and is largely based on soldier and worker gut morphology, could benefit from
359 the use of the many UCE markers designed in this study (Eggleton, 1999; Korb *et al.*, 2019).
360 UCE baiting from whole-genome shotgun sequencing is the perfect tool to carry out a global
361 taxonomic revision of termites.

362 **References**

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

366 Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D.
367 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–1325.

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

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

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

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

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

384 Buenaventura, E., Lloyd, M.W., Perilla López, J.M., González, V.L., Thomas-Cabianca, A.,
385 Dikow, T. 2021. Protein-encoding ultraconserved elements provide a new phylogenomic
386 perspective of Oestroidea flies (Diptera: Calyptratae). *Systematic Entomology* **46**: 5–27.

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

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

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

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

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

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

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

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

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

408 Engel, M.S., Grimaldi, D.A., Nascimbene, P.C., Singh, H. 2011. The termites of Early Eocene
409 Cambay amber, with the earliest record of the Termitidae (Isoptera). *ZooKeys* **148**: 105–
410 123.

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

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

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

418 Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C.
419 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple
420 evolutionary timescales. *Systematic Biology* **61**: 717–726.

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

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

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

427 Hedin, M., Derkarabetian, S., Alfaro, A., Ramírez, M.J., Bond, J.E. 2019. Phylogenomic

428 analysis and revised classification of atypoid mygalomorph spiders (Araneae,
429 Mygalomorphae), with notes on arachnid ultraconserved element loci. *PeerJ* **7**: e6864.

430 Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. 2018. UFBoot2:
431 improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**:
432 518–522.

433 Huang, W., Li, L., Myers, J.R., Marth, G.T. 2012. ART: a next-generation sequencing read
434 simulator. *Bioinformatics* **28**: 593–594.

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

438 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S. 2017.
439 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*
440 **14**: 587–589.

441 Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7:
442 improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–
443 780.

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

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

449 Kück, P., Longo, G.C. 2014. FASconCAT-G: extensive functions for multiple sequence

450 alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* **11**: 81.

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

455 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
456 Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:
457 2078–2079.

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

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

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

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

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

470 Quattrini, A.M., Faircloth, B.C., Dueñas, L.F., Bridge, T.C.L., Brugler, M.R., Calixto-Botía,
471 I.F., DeLeo, D.M., Forêt, S., Herrera, S., Lee, S.M.Y., Miller, D.J., Prada, C., Rádis-

472 Baptista, G., Ramírez-Portilla, C., Sánchez, J.A., Rodríguez, E., McFadden, C.S. 2018.
473 Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: new
474 approaches to long-standing problems. *Molecular Ecology Resources* **18**: 281–295.

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

477 Rice, P., Longden, L., Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open
478 Software Suite. *Trends in Genetics* **16**: 276–277.

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

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

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

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

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

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

493 Suyama, M., Torrents, D., Bork, P. 2006. PAL2NAL: robust conversion of protein sequence

494 alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**: W609–
495 W612.

496 Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and
497 ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**:
498 564–577.

499 Thorne, B.L., Grimaldi, D.A., Krishna, K. 2000. Early fossil history of the termites. In:
500 *Termites: Evolution, Sociality, Symbioses, Ecology* (T. Abe, D. E. Bignell, & M. Higashi,
501 eds), pp. 77–93. Kluwer Academic Publishers, Dordrecht, The Netherlands.

502 Van Dam, M.H., Henderson, J.B., Esposito, L., Trautwein, M. 2021. Genomic characterization
503 and curation of UCEs improves species tree reconstruction. *Systematic Biology* **70**: 307–
504 321.

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

508 Wang, M., Hellemans, S., Šobotník, J., Arora, J., Buček, A., Sillam-Dussès, D., Clitheroe, C.,
509 Lu, T., Lo, N., Engel, M.S., Roisin, Y., Evans, T.A., Bourguignon, T. 2021. Historical
510 biogeography of early diverging termite lineages (Isoptera: Teletisoptera). *bioRxiv*
511 2021.12.02.471008.

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

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

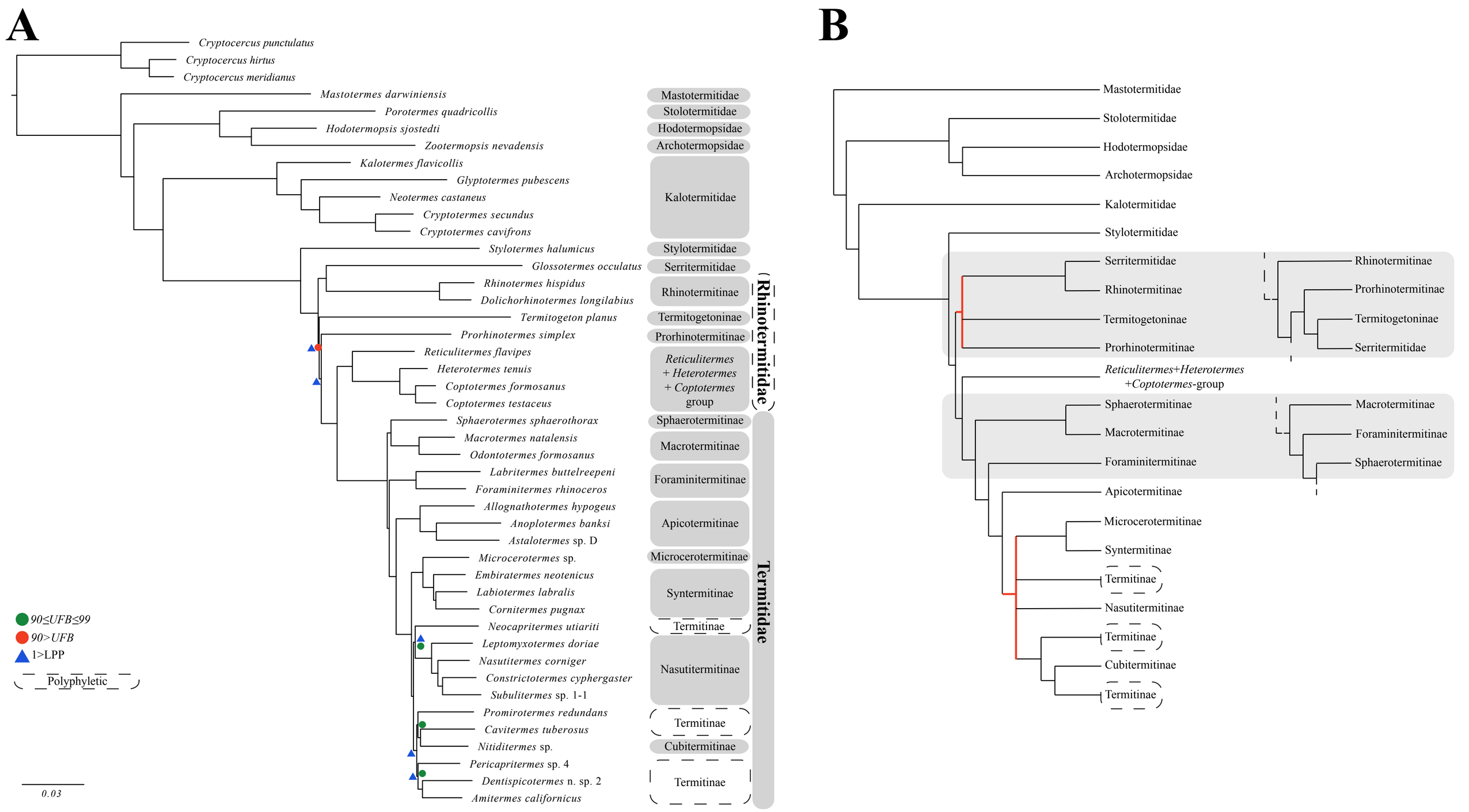
517 Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. 2018. ASTRAL-III: polynomial time species
518 tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.

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

522

523 **Figure captions**

524 **Figure 1**: (**A**) Maximum likelihood phylogenetic tree of termites reconstructed with IQ-TREE
525 using 5,934 UCE loci and complete mitochondrial genomes. Only UCE loci present in more
526 than 75% of species were used. Support values are indicated for non-fully resolved nodes:
527 ultrafast bootstrap (UFB; summarized from the phylogenetic trees reconstructed using UCE
528 only and UCE + mitochondrial DNA displayed in Figures S3 and S4, respectively) and
529 ASTRAL-III local posterior probabilities (LPP; phylogenetic tree displayed in Figure S5)
530 values. (**B**) Family-level summary topology of termites supported by both UCEs (this study)
531 and transcriptomic data (Bucek *et al*., 2019), with the indication of alternative topologies
532 inferred from mitochondrial genome data alone (Bourguignon *et al*., 2015, 2017). Unsupported
533 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
Rhinotermitinae
Prorhinotermitinae
Termitogetoninae
Serritermitidae
Reticulitermes+Heterotermes +Coptotermes-group
Sphaerotermitinae
Macrotermitinae
Foraminitermitinae
Macrotermitinae
Foraminitermitinae
Sphaerotermitinae
Apicotermitinae
Microcerotermitinae
Syntermitinae
Termitinae
Nasutitermitinae
Termitinae
Cubitermitinae
Termitinae