Wood feeding and social living: Draft genome of the subterranean termite *Reticulitermes lucifugus* (Blattodea; Termitoidae)

Jacopo Martelossi[1], Giobbe Forni[1,2], Mariangela Iannello[1], Castrense Savojardo[3], Pier Luigi Martelli[3], Rita Casadio[3], Barbara Mantovani[1], Andrea Luchetti[1]*, Omar Rota-Stabelli[4]

[1] Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy
[2] Dipartimento di Scienze Agrarie e Ambientali, Università degli Studi di Milano, Milano, Italy
[3] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy
[4] Center Agriculture Food Environment C3A, University of Trento/Fondazione Edmund Mach, Trento, Italy

*Correspondence: Andrea Luchetti, Department of Biological, Geological and Environmental Sciences, University of Bologna, via Selmi 3, Bologna 40126, Italy. Email: andrea.luchetti@unibo.it

Funding information
Canziani funding, Grant/Award Number: Canziani-Mantovani; COFUND Marie Curie - Provincia Autonoma di Trento

Abstract
Termites (Insecta, Blattodea, Termitoidae) are a widespread and diverse group of eusocial insects known for their ability to digest wood matter. Herein, we report the draft genome of the subterranean termite Reticulitermes lucifugus, an economically important species and among the most studied taxa with respect to eusocial organization and mating system. The final assembly (813 Mb) covered up to 88% of the estimated genome size and, in agreement with the Asexual Queen Succession Mating System, it was found completely homozygous. We predicted 16,349 highly supported gene models and 42% of repetitive DNA content. Transposable elements of R. lucifugus show similar evolutionary dynamics compared to that of other termites, with two main peaks of activity localized at 25% and 8% of Kimura divergence driven by DNA, LINE and SINE elements. Gene family turnover analyses identified multiple instances of gene duplication associated with R. lucifugus diversification, with significant lineage-specific gene family expansions related to development, perception and nutrient metabolism pathways. Finally, we analysed P450 and odourant receptor gene repertoires in detail, highlighting the large diversity and dynamical evolutionary history of these proteins in the R. lucifugus genome. This newly assembled genome will provide a valuable resource for further understanding the molecular basis of termites biology as well as for pest control.

Keywords: cytochrome P450, eusociality, gene families, genome evolution, odourant receptors (ORs), termites

INTRODUCTION

Termites (Insecta, Blattodea, Termitoidae) are mainly wood-feeding, diplodiploid eusocial insects. They are essential components of soil communities in terrestrial ecosystems, where they have positive effects on soil structure and productivity because of their ability to efficiently degrade wood matter (Govorushko, 2019). In fact, although using different modalities, most termites feed on wood, which plays a major role in the degradation of lignocellulosic material. This ability, which is uncommon in animals, is achieved by both endogenous and symbiont-provided enzymes (Cragg et al., 2015), as has been shown by the North American subterranean termite *Reticulitermes flavipes* (Scharf & Tartar, 2008; Tartar et al., 2009).

There are approximately 3000 termite species distributed in 10 families (Wang et al., 2022). All termite species, although to different degrees, exhibit some form of eusociality (Krishna et al., 2013). This condition describes a complex social organization characterized by overlapping generations, cooperative brood care and reproductive division of labour (Wilson & Hölldobler, 2005). In addition to termites, eusociality has been observed in ambrosia beetles (Biedermann & Taborsky, 2011), snapping shrimps (Crustacea) and naked mole-rats (Mammalia) (Wilson & Hölldobler, 2005), but is mostly present in Hymenoptera (bees, ants, wasps). In termites and hymenopterans, the two most studied social insect lineages, eusociality evolved independently, as suggested by the different conditions in which it originated. Hymenopteran species are haplo-diploid, with members of castes being all adults and only female workers, whereas termite taxa are all diplo-diploid, with only the primary reproductives being adults and workers of both sexes (Korb & Hartfelder, 2008). Although eusociality has been the subject of several evolutionary studies, recent advances in genomics have provided a deeper insight into the molecular processes that underlie this trait and its evolution. For example, some studies have correlated the evolution of eusociality with extensive gene family expansions (Harrison et al., 2018; Kapheim et al., 2015; Simola et al., 2013) while others have indicated its role in transposable element accumulation (Chak et al., 2021; Harrison et al., 2018; Korb et al., 2015).

Regarding Hymenoptera, which presently includes more than 50 sequenced genomes (Walsh et al., 2022), termite genomics is still in its infancy. To date, the genomes of only five species have been fully sequenced: *Zootermopsis nevadensis* (Terrapon et al., 2014), *Cryptotermes secundus* (Harrison et al., 2018), *Coptotermes formosanus* (Itakura et al., 2020), *Reticulitermes speratus* (Shigenobu et al., 2022) and *Macrotermes natalensis* (Poulsen et al., 2014).

The termite genus *Reticulitermes* Holmgren (Rhinotermitidae) includes subterranean termite species distributed all over the Holarctic, with clades that are clearly recognizable on a biogeographic basis (Dedeine et al., 2016). They are among the most studied taxa with respect to breeding system, colony composition and development (Matsuura, 2010; Vargo & Husseneder, 2010). Primary reproductives are winged adults that swarmed from their natal nest to found a new colony (Figure 1a). Workers constitute the bulk of the colony and carry out most of the work to maintain the nest, including taking care of royals' offspring (Figure 1b), and soldiers differentiate from workers to deal with colony defence (Figure 1c). Upon primary reproductives' senescence or death, neotenic secondary reproductives differentiate from nymphs to help or replace the former during offspring production (Figure 1d). In three *Reticulitermes* species, including R. lucifugus, secondary queens are produced through thelytokous parthenogenesis by the primary queen, eventually replacing the latter during colony maturation (Luchetti, Velonà, et al., 2013; Matsuura et al., 2009; Vargo et al., 2012). The main advantage of this reproductive system, also called "Asexual Queen Succession" (AQS; Matsuura et al., 2009), is the increase in colony growth rate with the maintenance of the genetic diversity of the colony, avoiding inbreeding caused by the mating of the king with related secondary queens (Matsuura, 2017). Herein, we report the genome sequence of the Italian species *R. lucifugus* (Luchetti, Scicchitano, & Mantovani, 2013) and a comparative analysis with previously published Blattodea genomes. In line with other studies involving termites, our findings indicate the expansion of several gene families, most of which are associated with detoxification and metabolism. Moreover, the *R. lucifugus* genome exhibits a larger repertoire of odourant receptor proteins than any other termite sequenced to date.

RESULTS

Genome sequencing, genome survey and assembly Overall, a total of 69.36 Gb was generated during *R. lucifugus* genome sequencing. Trimmomatic tool and Blobtools were used to trim approximately 3% of read pairs and identify 201,770 reads mapped to contaminant contigs (0.0007% of the total reads obtained), respectively (Table S1). The final assembly contained 56,321 scaffolds, covering 813 Mb with an N50 of 42.6 kb (Table 1). The genome length was estimated to be 921.36 Mb by Genomescope analysis with a best-fit k-mer size of 30 bp; thus, the assembly covered up to 88% of the total genome length (Figure 2a; Table S2).

Ninety-eight percent of short reads were successfully remapped on the assembly, highlighting the successful assembly of most of the sequenced genome. In addition, assembly spectra copy-number plot revealed a k-mer completeness of 98% without artefactual duplications (Table 1; Figure S1). Indeed, for an expected haploid or completely homozygous assembly, the majority of k-mers present in the reads were also present in the assembly. In line with this, more than 99.9% of the genome was found to be homozygous using Genomescope (Figure 2a).

BUSCO revealed high completeness of the genome assembly, with 987 (96.7%) Arthropoda core genes as complete and single copies, 0.7% as complete and duplicated, 2.2% as fragmented and only 0.4% as missing. Merqury highlighted a

highly accurate assembly with a consensus quality value (QV) of 61.7 and an associated error rate of 0.0000007 (Table 1).

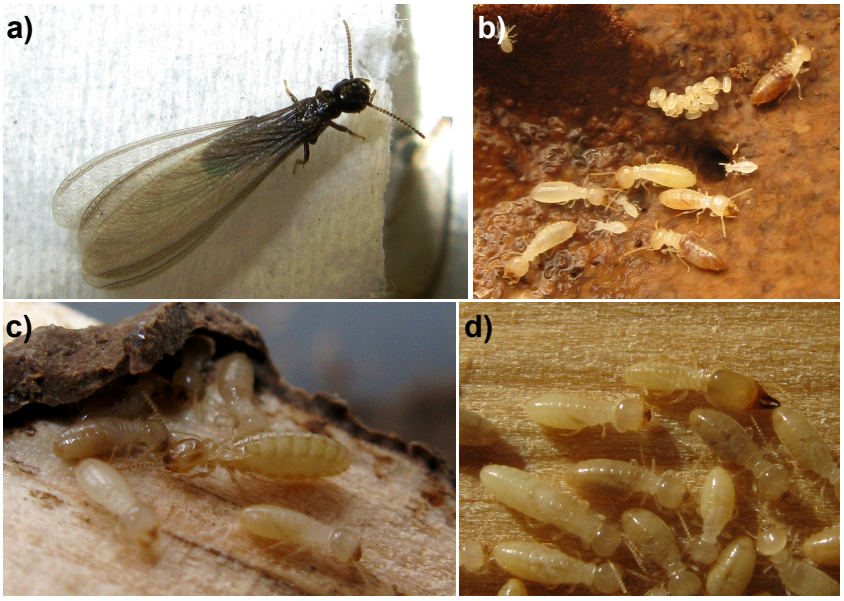

FIGURE 1
Reticulitermes lucifugus termites. (a) Swarming winged reproductives (primary royal); (b) workers, early instar nymphs and eggs; (c) workers and a soldier; and (d) workers with neotenic reproductives (secondary royal) exhibiting undeveloped wings. (photos created and provided by Silvia Ghesini)

Genome annotation

Overall, repetitive elements masked 41.7% of the genome. The repeat annotation pipeline identified 3,821 consensus sequences of transposable elements (TE); of these, 1,781 showed no significant similarity to RepBase, while the 73% were successfully classified using TEClass. The predominant interspersed repeats were LINEs (11.49%), followed by DNA elements (10.97%), SINEs (8.78%) and LTRs (3.69%), whereas MITE elements covered only 3.26% of the genome (Table 1). The activity of TE was measured using a repeated landscape plot. Briefly, the CpG-corrected Kimura distance of each TE copy from its relative consensus sequence was used as a proxy for the time from insertion, and the younger copies showed lower Kimura distance values (Figure 2b). We identified two main bursts of activity: an older one at 25% divergence, mainly driven by DNA and LINE elements and a younger one at 8% of divergence, which mainly involves DNA and SINE elements (Figure 2b).

Gene prediction resulted in 17,298 protein-coding genes with a mean length of 12,362 bp and an average of six exons per gene (Table 1). Of these gene models, 16,349 (95%) were highly supported by an Annotation Edit Distance (AED, a measure of the agreement of an annotation to external evidence such as proteins and transcripts, with 0 indicating maximum agreement and 1 indicating no support) of <0.5 and were, therefore, used for downstream analyses (Table 1; Figure S2). InterProScan successfully annotated 15,115 proteins with at least one hit in one or more of the queried databases. AAI-profiler analyses based on BLASTP results from the Uniprot database were annotated as Blattodea-related for 88% of the sequences, whereas less than 0.05% were assigned to non-eukaryotic organisms (Figure S3).

Ortholog search, phylogenomics and evolutionary analyses of gene families

An ortholog search was run on the proteome datasets of *Locusta migratoria*, *Blattella germanica*, *Z. nevadensis*, *Cr. secundus*, *M. natalensis*, *Co. formosanus*, *R. speratus* and *R. lucifugus*. When considering the entire *R. lucifugus* reference gene set (N = 16,349), Orthofinder identified 15,206 orthogroups (OGs), embodying 87.5% of the total analysed genes. Most of *R. lucifugus* genes (96.6%) were placed in an OG, with 1.3% in species-specific OGs; 4.7% of its genes were identified as species-specific (comprising species-specific OGs and unplaced genes; Figure 2c, right panel) and 1.6% were shared with *R. speratus* but not with other termites (i.e., they are Reticulitermesspecific). The ortholog search retrieved 3,034 1:1 OGs among the genomes that were used for phylogenetic inference. The maximum likelihood tree identified a monophyletic termite clade built by *Z. nevadensis*, *Cr. secundus*, *Co. formosanus*, *M.*

natalensis, *R. speratus* and *R. lucifugus* in a sister relationship with *B. germanica* (Figure 2c). All nodes were fully supported by bootstrap values (100%).

TABLE 1
Reticulitermes lucifugus genome assembly and annotation statistics (QV, consensus quality value calculated by Merqury; AED, annotation edit distance

Genome feature	Statistics
Assembly size	812,549,426 bp
G+C content	39.6%
Number of scaffold (N50)	56,321 (42,619 bp)
Number of contigs (N50)	135,150 (10,077 bp)
Mapping rate	98.47%
k-mer completeness	97.88%
BUSCO (N=1,013)	C:97.4%[S:96.7%,D:0.7%],F:2.2%,M:0.4%
QV (error rate)	61.7 (0.0000007)
N. of predicted genes	17,298
N. of highly supported genes (AED < 0.5)	16,349
Proportion of repeats coverage	41.71%
LINE	11.49%
SINE	8.78%
LTR	3.69%
DNA	10.97%
MITE	3.26%

The FULL-CAFE analysis, which included all OGs found using all *R. lucifugus* supported gene models (N = 16,349), revealed that 8501 OGs were present at the root of the species tree, calculating an associated error of 3% and a λ of 0.00089. In the *R. lucifugus* branch, the overall number of expanded OGs was higher than that in the other Isoptera branches, and the number of significantly expanded/ contracted OGs was 352/187 (Figure 2c). The same analysis was carried out including only *R. lucifugus* complete gene models—that is, spanning from start to stop codon—in order to consider possible effects of genome fragmentation (RED-CAFE analysis; N = 9,720; 59% of the whole reference gene set). Orthofinder search inferred 14,654 OGs; 98.1% of *R. lucifugus* genes were placed in an OG and 0.4% in species-specific OGs. The RED-CAFE analysis evaluated 8,480 OGs, calculating an error rate of 4% and a λ of 0.00083. As expected, the exclusion of the 41% of *R. lucifugus* genes greatly increased the number of inferred species-specific contraction events compared to that of the FULL-CAFE analyses (2,242 vs. 366) and decreased the number of expansions (170 vs. 744) and significant expansion events (93 vs. 352) (Figure 2c). Notably, we did not find any clear cases of false duplications through reciprocal BLASTP analyses.

Functional annotation of significantly expanded OGs from both FULL-CAFE and RED-CAFE analyses was carried out using Gene Ontology (GO) terms (Figure 3a–c), KEGG pathways (Figure 3b–d), and KEGG Orthology (KO) (Table 2). Most of the GO terms common between the two analyses were related to development and cellular growth (e.g., “larval somatic muscle development”,” regulation of multicellular organism growth”, “positive regulation of developmental process”, “positive regulation of cell growth”), metabolism (e.g., “fatty acid derivative metabolic process”, “xenobiotic metabolic process”, “lipid metabolic process”), transcription regulation/DNA replication (e.g., “heterochromatin assembly”, “DNAtemplated transcription, initiation”, “nucleosome assembly”, “positive regulation of DNA-binding transcription factor activity”) and perception (e.g., “ionotropic glutamate receptor signalling pathway”, “response to stimulus”, “sensory perception of taste”; Figure 3a–c).

The KEGG pathway annotations revealed that most of the expanded OGs were related to different metabolic pathways, such as “drug metabolism-cytochrome P450”, “metabolism of xenobiotics by cytochrome P450”, “steroid hormone biosynthesis”, “drug metabolismother enzymes” and “retinol metabolism” (Figure 3b–d). The significantly enriched KO terms identified in both analyses and related to these pathways were cytochrome P450 family 6 (K14999), transient receptor potential cation channel subfamily A member 1 (K04984), alcohol-forming fatty acyl-CoA reductase (K13356),

multiple histones (K11251, K11254 and K11252), gonadotropin-releasing hormone receptor (K04280), prostaglandin-H2 D-isomerase/glutathione transferase (K04097) and SOX transcription factor (K09267).

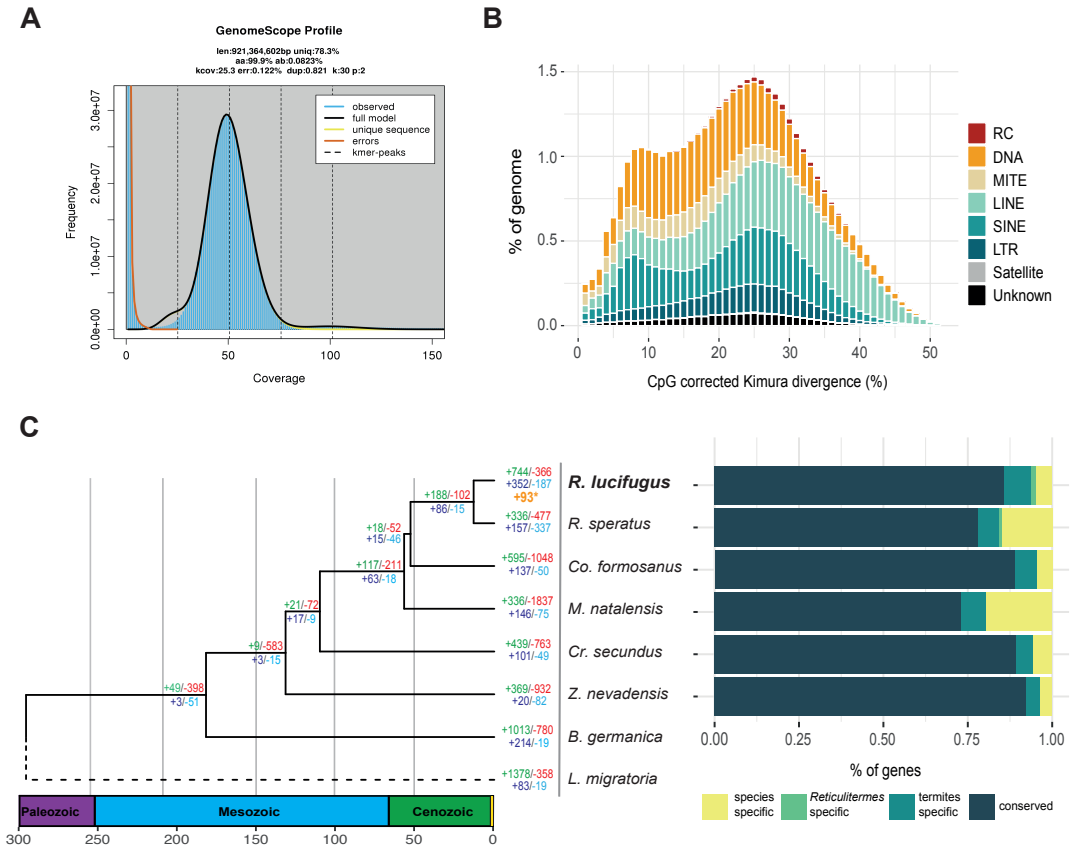

FIGURE 2
Genomic analyses of *Reticulitermes lucifugus*. (a) Genomescope results with best-fit k-mer size = 31 (model fit: Min = 78.24%; max = 96.95%). (b) Transposable elements' genomic landscape. (c) Time calibrated phylogenomic tree built on 3,034 orthogroups. All nodes received 100% UltraFast bootstrap nodal support. Numbers at nodes and tips represent results of the FULL-CAFE analyses (considering all the *R. lucifugus* reference gene set). The overall number of expanded and contracted orthogroups is indicated by green and red colours respectively, while the number of statistically significantly expanded or contracted orthogroups is indicated by blue and light blue colours, respectively. When analysing only complete gene models (i.e., spanning from start to stop codon, RED-CAFE analyses), the number of lineage-specific significantly expanded gene families is indicated by orange and an asterisk. In the right panel is reported a comparison of orthologous genes found between analysed genomes considering the whole *R. lucifugus* reference gene set

Evolutionary analyses of P450 and odourant receptors (ORs) gene families

The genome of *R. lucifugus* seemed to exhibit a large gene family expansion in the cytochrome P450 family, counting a total of 184 genes, followed by *B. germanica* (145) and *L. migratoria* (132), while the other termites ranged from 78 copies for *Z. nevadensis* to 106 copies for the sister species *R. speratus* (Figure 4a; Table S3). However, of the 184 *R. lucifugus* P450 genes, only 62 were annotated as complete, whereas 33 incomplete gene models possessed a protein product with a P450 Pfam alignment length equal to or higher than 60% of the expected size (203 amino acids), accounting for a total of 95 confident gene models. The analysis of the reduced dataset (N = 955 protein sequences) placed Blattodea and non-Blattodea sequences in all four P450 clans: CYP2, CYP4, CYP3 and MITO (Dermauw et al., 2020), even with an unbalanced distribution (Figure 4b). The *R. lucifugus* CYPome consists of members of all four clans; however, a lineage-specific bloom is localized in the CYP3 clan, specifically in the CYP6 family (Figure 4b; Supplementary Data S1). In the OR family, *R. lucifugus* appears to have a gene count marginally higher than that of *B. germanica* (129 vs. 128 genes; although Robertson et al., 2018, reported 123 genes in the German cockroach), but lower than the 154 genes reported for the cockroach *Periplaneta americana* (Li, 2018). However, only 38 out of these were annotated as complete, while the other 42 spanned more than the 60% of the expected Pfam domain size (153 amino acids), accounting for a total of 80 confident OR genes and making R. lucifugus the most gene-rich species among

analysed Isoptera. *Zootermopsis. nevadensis*, *Co. formosanus* and *Cr. secundus* had OR gene numbers that were considerably similar to each other (65, 64 and 63 genes, respectively). In contrast, *M. natalensis* has only 14 family members, representing the OR genes poorest species. Interestingly, the *R. lucifugus* congeneric species R. speratus showed a great reduction in the OR repertoire (only 22 genes) (Figure 4a; Table S4). In the final OR dataset (N = 891 protein sequences), the majority of the Blattodea and non-Blattodea sequences were grouped into distinct clusters (Figure 4c; Supplementary Data S2). The only exception was the olfactory receptor coreceptor (Orco) cluster that was composed of ubiquitous 1:1 orthologs from all considered taxa, with only *L. migratoria* showing two members. Blattodea protein sequences were clustered into two main clades, BClade1 and BClade2, which showed important differences in sequence composition, with BClade1 harbouring the majority of *B. germanica* sequences (74% of genes); for termites, we found an almost even distribution of genes between the two clades.

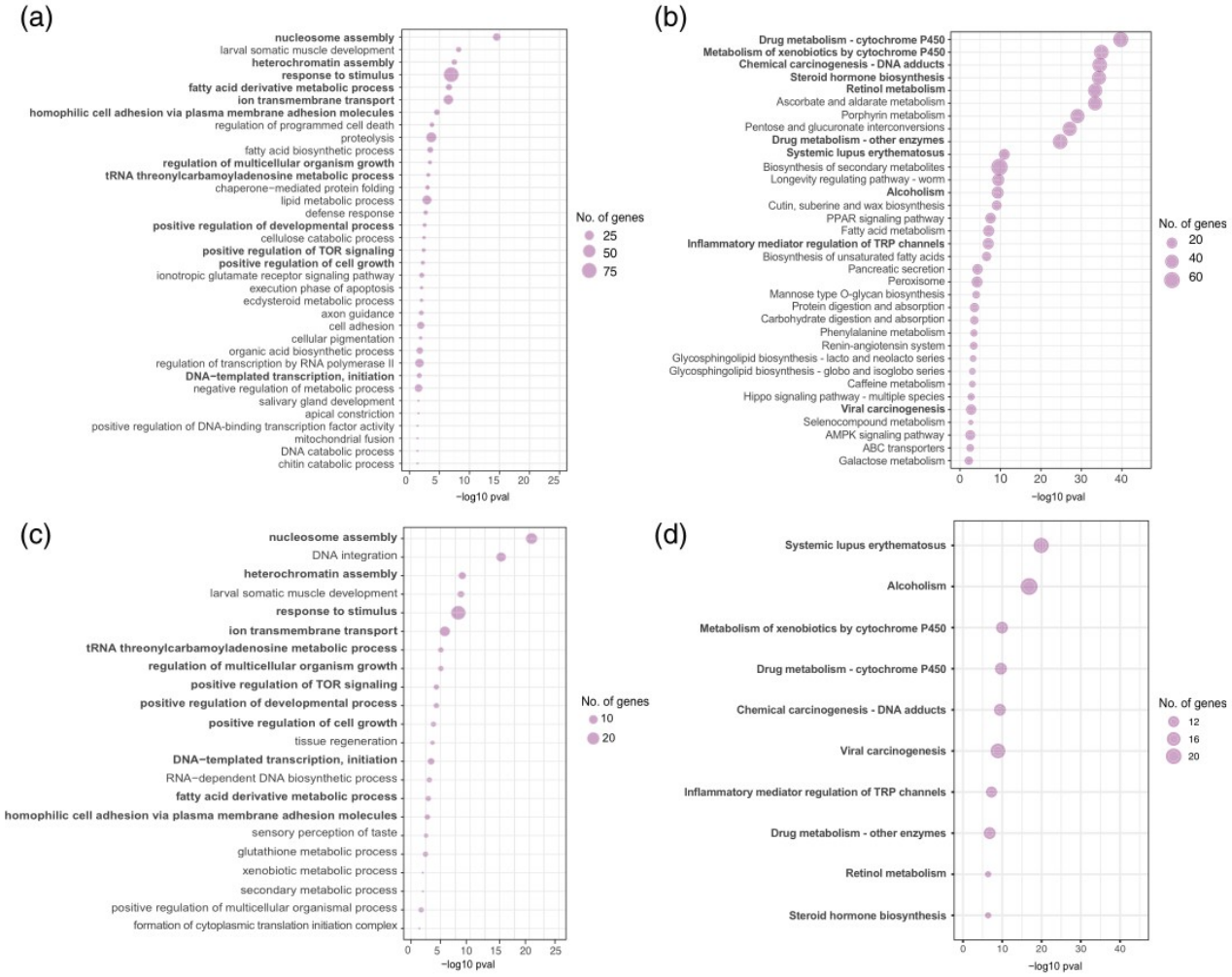

FIGURE 3

Functional enrichment analyses for FULL-CAFE and RE (a) And (c) GO enrichment analyses for *Reticulitermes lucifugus* significantly expanded orthogroups for FULL-CAFE (considering all the reference gene set) and RED-CAFE analyses (considering only complete gene models spanning from start to stop codon), respectively. (B) and (d) KEGG pathway enrichment for FULL-CAFE and RED-CAFE analyses. -log10 p-values of the enrichment analysis are on the x-axis, and the bubble size is proportional to the number of genes

TABLE 2
Significantly overrepresented Kegg Orthology (KO) term along with KO ID and description identified in *Reticulitermes lucifugus* significantly expanded gene families in both FULL-CAFE and RED-CAFE analyses

KO term	KO id	KO description	adjusted p-value
K14999	CYP6	cytochrome P450 family 6	<0.01 (<0.01)
K04984	TRPA1,ANKTM1	transient receptor potential cation channel subfamily A member 1	<0.01 (<0.01)
K13356	FAR	alcohol-forming fatty acyl-CoA reductase	<0.01 (<0.05)
K11251	H2A	histone H2A	<0.01 (<0.01)
K11254	H4	histone H4	<0.01 (<0.01)
K11252	H2B	histone H2B	<0.01 (<0.01)
K13111	SMU1	WD40 repeat-containing protein SMU1	<0.01 (<0.01)
K16507	DCHS1_2,PCDH16_23	protocadherin-16/23	<0.01 (<0.01)
K05125	DDR2,TKT,CD167b	discoidin domain receptor family member 2	<0.01 (<0.01)
K08851	TP53RK,PRPK,BUD32	TP53 regulating kinase and related kinases	<0.01 (<0.01)
K09381	PAX3_7	paired box protein 3/7	<0.01 (<0.01)
K01183	E3.2.1.14	chitinase	<0.01 (<0.05)
K04097	HPGDS	prostaglandin-H2 D-isomerase/glutathione transferase	<0.01 (<0.01)
K07827	KRAS,KRAS2	GTPase KRas	<0.01 (<0.01)
K13218	PTBP1,PTB	polypyrimidine tract-binding protein 1	<0.01 (<0.01)
K21754	KCTD1_15	BTB/POZ domain-containing protein KCTD1/15	<0.01 (<0.01)
K15407	QTRT2,QTRTD1	queuine tRNA-ribosyltransferase accessory subunit	<0.01 (<0.01)
K09373	LHX2_9	LIM homeobox protein 2/9	<0.01 (<0.01)
K09267	SOX1S	transcription factor SOX1/3/14/21 (SOX group B)	<0.05 (<0.01)
K04280	GNRHR	gonadotropin-releasing hormone receptor	<0.05 (<0.01)

DISCUSSION

Herein, we present a draft genome assembly and annotation of the lower termite, *R. lucifugus*. This species exhibits the AQS mating system (Luchetti, Velonà, et al., 2013; Matsuura et al., 2009), in which secondary queens are produced through automatic parthenogenesis, the cytological mechanism of which is terminal fusion. Hence, the genetic outcome resulted in completely homozygous individuals (Matsuura et al., 2009). Therefore, in order to minimize assembly errors derived from heterozygosity, we sequenced a parthenogenetically produced secondary queen, and in line with this expectation and previous data based on microsatellite data (Luchetti, Velonà, et al., 2013), the assembled genome showed a homozygosity of 99.9%.

Based on k-mer analysis, the length of the sequenced genome was expected to be 920 Mb, an estimate that is close to the cytofluorimetric evaluation of the genome size of the related species *R. flavipes* (1.14 Gb) and *R. speratus* (1.07 Gb) (Koshikawa et al., 2008). Therefore, the sequenced 813 Mb covers up to the 88% of the expected genome size, which is in line with the recently published genomes of the congeneric *R. speratus* (881 Mb; Shigenobu et al., 2022) and its sister *Co. formosanus* (875.84 Mb; Itakura et al., 2020). Overall, sequenced Rhinotermitidae exhibited an intermediate genome size among those observed in *Z. nevadensis* (562 Mb; Terrapon et al., 2014), *Cr. secundus* (1.30 Gb; Harrison et al., 2018) and *M. natalensis* (1.31 Gb; Poulsen et al., 2014). However, owing to the relatively high repetitive content and the use of only small insert size libraries, our assembly was largely fragmented compared to that of the previously mentioned termite genomes, particularly at the scaffold level (scaffold N50: *R. speratus* = 1.97 Mb; *Co. formosanus* = 1.43 Mb; *Z. nevadensis* = 751,105 bp; *Cr. secundus* = 1.19 Mb; *M. natalensis* = 2.0 Mb). Nevertheless, we obtained a good level of completeness, as reported by BUSCO results for read mapping (98%) and KAT k-mer completeness (97%) analyses. Moreover, the number of predicted genes validated by AED statistics (N = 16,349) is in line with the gene annotations performed in other termite genomes, ranging from 12,381 in *Z. nevadensis* (Terrapon et al., 2014) to 16,310 in *M. natalensis* (Poulsen et al., 2014) and 15,591 in *R. speratus (*Shigenobu et al., 2022). Despite we could not exclude that the slightly higher gene count of *R. lucifugus* compared to *R. speratus*, and, thus, the high number of lineage-specific expansions of gene families, could be due partly because of the fragmentation of one gene in two different scaffolds, thus resulting in two different gene models, our results indicated that the obtained assembly can be used as a reliable source for genomic information of *R. lucifugus*.

Phylogenomics and divergence time analyses using 3,034 ubiquitous, single copy orthologs shared between *L. migratoria* and Blattodea resulted in a clustering pattern that was in complete agreement with the current knowledge of Blattodea and termite phylogeny (Bourguignon et al., 2015; Bucek et al., 2019; Dedeine et al., 2016).

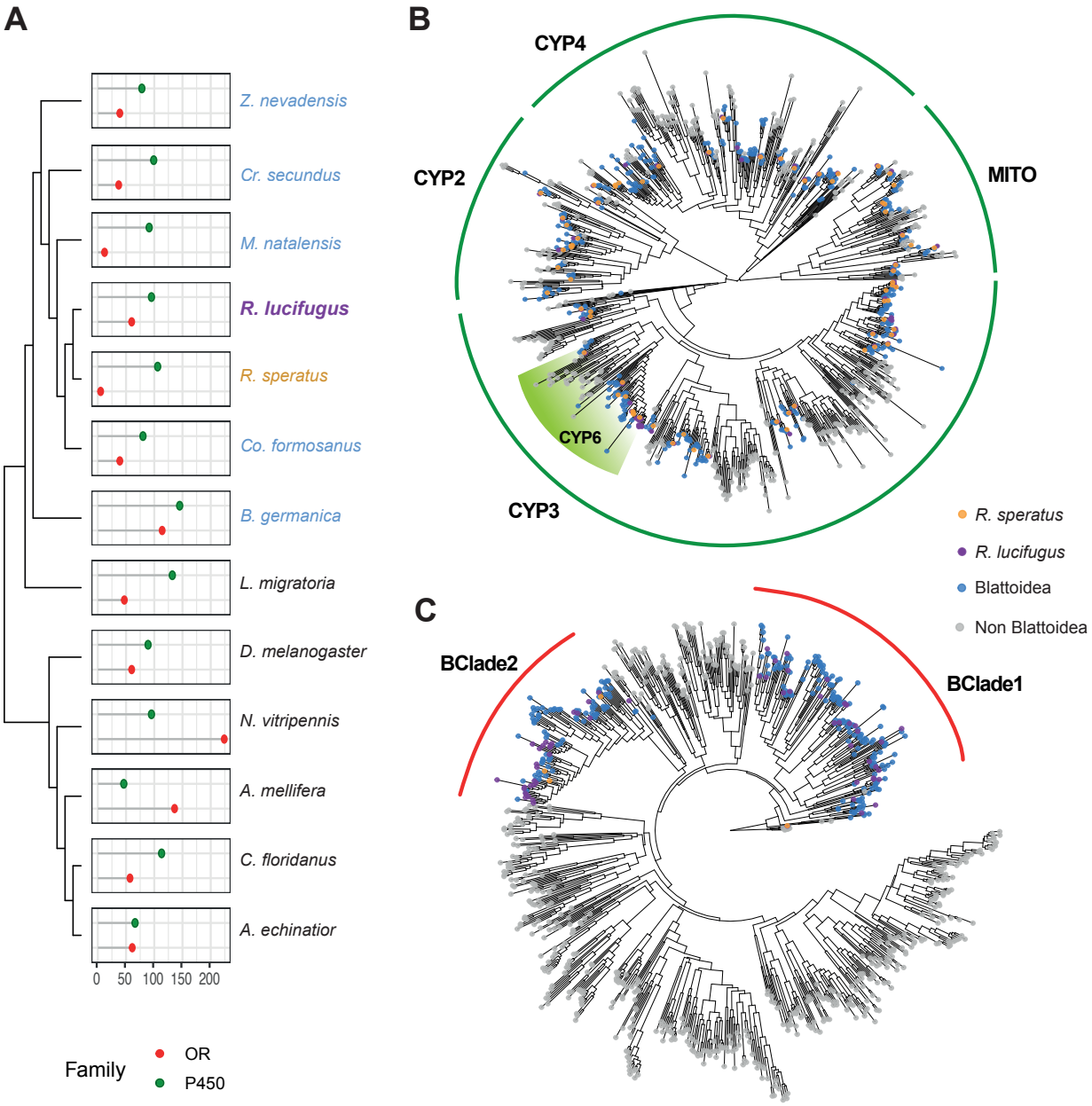

FIGURE 4
(a) Number of odourant receptors (OR) and cytochrome P450 genes annotated in Blattodea genomes and a set of other selected insects for comparative purpose. Note that only highly supported gene models are counted (i.e., spanning from start to stop codons or possessing an alignment covering at least the 60% of the expected Pfam signature). (b) and (c) are reconciled phylogenetic trees of P450 and OR genes, respectively. Clade annotation is discussed in the main text

In previous analyses of termite genomes, a link has been proposed between TE activity, gene duplications and the evolution of eusociality during the early phases of termite evolution (Harrison et al., 2018; Korb et al., 2015). *Reticulitermes lucifugus*, *R. speratus* and *M. natalensis* showed similar TE contents of respectively 41.7%, 40.4%

(Shigenobu et al., 2022) and 45.9% (Korb et al., 2015), placing them between the TE-rich genome of *Cr. secundus* (55%, Harrison et al., 2018) and the TE-poor *Z. nevadensis* (28% Korb et al., 2015). The TEs landscape of *R. lucifugus* shares a relatively old peak of transposition activity with other termites, corresponding to 25% of Kimura divergence, which involved mainly LINEs and DNA elements (Korb et al., 2015). Moreover, as observed in *M. natalensis* but not in *Z. nevadensis* (Korb et al., 2015), a second, lower, but more recent burst of transposition activity is observable at 7% to 8% of Kimura divergence, mainly driven by DNA and SINE elements. This suggests a shared genome architecture in the Rhinotermitidae+Termitidae lineage. Although we could not directly test a possible correlation between gene family amplification and TE activity because of the fragmentation of the obtained genome, we found multiple instances of species-specific significant gene family expansions in both the FULL- and RED-CAFE analyses.

When analysing the functional annotation of these expanded gene families, we identified functions related to the metabolism of xenobiotics and nutrients, perception, steroid biosynthesis and regulation of transcriptional activity. *Reticulitermes lucifugus*, similar to other Isoptera, can feed on a wide variety of plants with the ability to counter different plant toxins, such as pine terpenes (Nagnan & Clement, 1990), and to digest a broad range of plant compounds, including lignocellulose (Scharf & Tartar, 2008). Accordingly, KEGG analysis identified several significantly expanded gene families whose activity is linked to detoxification, such as “drug metabolisms by cytochrome P450”, “xenobiotic biodegradation and metabolism” and “drug metabolism-other enzymes”. In *R. lucifugus*, we identified one and two significantly expanded OGs of the cytochrome P450 (CYP) family in the RED-CAFE and FULL-CAFE analyses, respectively. This large, old and diverse group of heme-containing mono-oxygenase proteins possesses high inter- and intraspecific variability, and is related to detoxification processes (Lu et al., 2021), ecdysteroid metabolism (Dermauw et al., 2020), and caste differentiation affecting juvenile hormone signalling and synthesis (Helvig et al., 2004; Tarver et al., 2012; Terrapon et al., 2014). Our phylogenetic analyses recovered the monophyly of the four major P450 clans present in insects (CYP2, CYP3, CYP4 and MITO), with >80% of the genes falling inside CYP3 and CYP4 clusters, thus confirming a greater number of insect blooms inside these clades (Dermauw et al., 2020; Rane et al., 2019). When considering the whole gene set, *R. lucifugus* appeared to have the largest CYPosome among the presently analysed insects. However, if considering only highly supported gene models (i.e., complete and/or with a Pfam hit that covers at least 60% of the expected length), we found a more homogenous gene count across all Isoptera, with 95 members for *R. lucifugus*, 106 for *R. speratus*, 80 for *Co. formosanus*, 91 for *M. natalensis*, 99 for *Cr. secundus* and 98 for *Z. nevadensis*. Nevertheless, based on both KO term enrichment and gene phylogenetic analyses, we found a species-specific expansion of genes related to the CYP6 family, which is included in the CYP3 clan, whose proteins are known to be involved in phytochemical detoxification (Lu et al., 2021). A Blattodea-specific clade was placed in this family; in addition to 11 *R. lucifugus* genes, multiple paralogs from other termites, such as seven copies in the congeneric *R. speratus*, nine in *M. natalensis*, seven in *Z. nevadensis*, four in *Cr. secundus* and only one in *Co. formosanus*, were found. The molecular diversification of this family could be due to their different geographical distributions and, therefore, differences in dietary habits.

Chemosensory protein (CSP) gene families are among the most well-studied proteins of both eusocial and solitary insects that act as detectors for toxins, pathogens, food and pheromones, such as cuticular hydrocarbons (Joseph & Carlson, 2015). Among CSPs, OR genes have been particularly well analysed in eusocial insects. In addition to the Orco cluster (which contains 1:1 orthologs) (Brand et al., 2018), whose members encode coreceptors necessary for odourant detection, the hymenopteran OR gene family is greatly expanded, whereas Blattodea has fewer members. Within Hymenoptera, both eusocial and non-eusocial taxa appear to be characterized by an expanded OR gene repertoire, with the solitary wasp *Nasonia vitripennis* encoding more than 220 OR genes. These rich but highly variable gene repertoires are probably the result of independent amplification and loss of gene families, with the retention of duplicate copies in different OR lineages. Previous phylogenetic analyses have placed gene family amplifications of different species in different clusters, implying different potentially adaptive roles of these proteins in eusocial and non-eusocial taxa (Robertson et al., 2010). Indeed, OR genes in eusocial hymenopterans are mainly involved in colony communication, nestmate recognition and regulation of reproduction, whereas the solitary wasp N. vitripennis may benefit from a broad set of OR to find potential mates and hosts (Brand & Ramírez, 2017; Robertson et al., 2010; Zhou et al., 2015). Among Isoptera, the subterranean termite *R. lucifugus* showed the largest gene count when considering only highly supported gene models and even higher than that of *B. germanic*a when considering the whole reference gene set. Moreover, lineage-specific amplification of gene families in *R. lucifugus* and *B. germanica* tended to cluster separately, with that observed in the termite genome to be mainly localized in the highly divergent clade BClade2, reflecting a similar scenario observed in Hymenoptera. Interestingly, the extremely low gene count of the congeneric species *R. speratus* also deviates from *Reticulitermes*’ sister taxon *Co. formosanus*, implying a previously underestimated high turnover rate of the OR gene family inside Isoptera. This was already observed by Shigenobu et al. (2022), and despite the difficulties in assembling and annotating OR genes, both *R. lucifugus* and *R. speratus* genomes were automatically annotated with similar pipelines, making it difficult to explain such variation only by misannotation. Interestingly, GO term analysis of expanded gene families also indicated biological processes linked to ion transport, including Ca+, into the cytosol, which is in line with the finding that ORs seem to have the characteristics of a nonselective ligand-gated ion channel (odour or pheromone) (Sato et al., 2008). However, it is anticipated that CAFE analyses will not report any significantly expanded OR OGs. In fact, the high degree of Blattodea lineage-specific

duplications and losses, as resulting from the phylogenetic analysis of OR genes, can obscure orthology inferences, artificially splitting OGs (Emms & Kelly, 2020). If an OG is split, it cannot be reconstructed as present at the root of the tree and, thus, remains invisible to gene family turnover analyses. Evidence for this hypothesis comes from OG00011, obtained by analysing only complete gene models whose sequences are nested inside BClade2. It contains 12 *R. lucifugus* ORs, together with only one *M. natalensis* sequence. These OGs clearly showed a large copy number variation in favour of *R. lucifugus*; however, they were not reconstructed as present at the root and were excluded from analyses.

The biological significance of the P450 and OR gene family expansions is yet to be defined, particularly considering the similar social complexity and feeding biology with other presently analysed Rhinotermitidae (*C. formosanus* and *R. speratus*), which appears to suggest local adaptation. Further genomic analyses and functional experiments are necessary to better characterize the possible impacts of these differences on species-specific phenotypic traits. However, it is important to note that gene family expansions and contractions do not necessarily correlate with adaptive processes. In fact, birth-death processes do not require selection, allowing for wide copy number variations between species to be explained only by drift. For example, the link between CYPosome size evolution and plant–animal chemical warfare is a long and ongoing debate with apparently multiple contrasting evidence (e.g., Calla et al., 2017; Gonzalez & Nebert, 1990; Good et al., 2014; Seppey et al., 2019; Sezutsu et al., 2013). However, gene family blooms may constitute a pre-adaptive pool of gene and protein diversity that could eventually be subjected to selective pressures, as observed for the P450 gene family in *Drosophila* lineages (Good et al., 2014) and in dietary shifts during the diversification of beetles (Seppey et al., 2019) or the OR gene family in pea aphids (Smadja et al., 2009) and corbiculate bees (Brand & Ramírez, 2017).

EXPERIMENTAL PROCEDURES

Sample collection and DNA extraction and sequencing

Using a Stratec DNA Isolation Kit (Invisorb), total DNA was isolated from the body of a parthenogenetically produced secondary queen collected from Massa Marittima, after careful dissection for gut removal (Italy; Luchetti, Velonà, et al., 2013). Next-generation sequencing (Illumina HiSeq2000 platform) was performed at the Danish National High-Throughput DNA Sequencing Centre. Two libraries with 350 bp and 550 bp insert sizes were sequenced to produce 125 bp pair ends (PE). Raw reads were submitted to the NCBI Sequence Read Archive under acc. nos. SRR17267499 and SRR17267498.

Transcriptome assembly, K-mer based genome survey and genome assembly

Single-end transcriptomic reads from two neotenic samples sequenced by Dedeine et al. (2015) were downloaded from NCBI (SRR1325112, SRR1325111), filtered and cleaned (adapter sequences were trimmed) using Trimmomatic v. 0.35 (Bolger et al., 2014) with a minimum quality of 30 and a minimum residual length of 20. Reads with low abundance k-mers were removed using the ErrorCorrectReads Perl script from AllPaths-LG v.4.4.3 (Butler et al., 2008). Trinity v. 2.4.0 (Grabherr et al., 2011) was used with default parameters to assemble reads into transcripts. Contaminants were filtered from the de novo transcriptome assembly using a BLASTX DIAMOND search (e-value <0.005; Buchfink et al., 2021) against the NCBI non-redundant database (nr). Only transcripts with the best hits against the Arthropoda species were retained.

Genomic PE reads were deprived of adapters, trimmed and quality-checked using Trimmomatic; the reads were trimmed using a minimum quality set to 20 and a minimum residual length of 36 (Table S1). All trimmed PE reads were used for k-mer-based heterozygosity and genome size estimations. KMC v 3.1.1 (Kokot et al., 2017) was used to produce k-mer histograms (k-mer sizes of 21, 24, 27, 30), which were uploaded to the online implementation of Genomescope2 (Vurture et al., 2017). The best k-mer size was chosen based on the fit of the resulting Genomescope model.

Ab initio genome assembly was performed using the de Brujin graph ABySS assembler v.1.5.2 (Simpson et al., 2009). Both PE libraries (with 350 bp and 550 bp insert sizes) were adopted to perform both primary contig assembly and scaffolding using the dedicated ABySS modules. The size of the k-mer, used by the assembly algorithm to build the de Brujin graph, was optimized by evaluating different values in the range of 32 to 96. The optimal value was then determined by comparing the resulting assemblies according to N50 statistics. The final optimal assembly leading to higher contiguity statistics was obtained using a k-mer length of 64. Possible bacterial contamination was assessed using Blobtools v. 2 (Laetsch & Blaxter, 2017). Briefly, all contigs were blasted against the NCBI nr database, and their coverage was retrieved from minimap analyses. All reads that mapped to contaminant contigs were removed, and the genome was reassembled with the best-fit previously found k-mer length. An additional scaffolding step was performed using L_RNA_scaffolder (Xue et al., 2013) together with the assembled transcriptome. The completeness of the assembly was evaluated as follows: (1) mapping back all trimmed PE reads to the assembly with minimap2 (Li, 2018); (2) using K-mer Analysis Toolkit (KAT) v. 2.4.2 (Mapleson et al., 2017) to identify artefactual duplications and/or missing k-mers from the genome; (3) using the Arthropoda Orthodb set v.10 with Busco v. 5 on the online platform gVolante (https://gvolante.riken.jp/; Nishimura et al., 2019; last accessed in September 2021); (4) calculating the QV (i.e., a k-mer-based estimation of the frequency of consensus errors in the assembly, where higher values indicate a

more accurate consensus) using Merqury. The final assembly is available on NCBI GenBank, under BioProject PRJNA778916 (accession number JAKCWW000000000).

Genome annotation

The annotation of TEs was performed using a combination of de novo and homology-based approaches. RepeatModeler v. 2.0.1 (Flynn et al., 2020) with the LTR pipeline extension, MITE Tracker (Crescente et al., 2018), SINE_Scan v. 1.1.1 (Mao & Wang, 2017) and HelitronScanner v. 1.1 (Xiong et al., 2014) were run, with default parameters, to create de-novo consensus libraries. Non-TE-related genes were removed from each of these using BLASTX (e-value <1 e10) and ProtExcluder v. 1.1 (Campbell et al., 2014) with the predicted proteomes of *B. germanica, Z. nevadensis, Co. formosanus, Cr. secundus, M. natalensis, L. migratoria, Drosophila melanogaster, Danaus plexippus, Ephemera danica* and *Solenopsis invicta* (Table S5). Consensus sequences without similarity to RepBase were classified based on the mechanism of transposition (i.e., SINE, LINE, LTR, Helitron and DNA) using TEClass (Abrusán et al., 2009). All de novo libraries were merged, and redundancy was removed using CD-HIT (Fu et al., 2012; 95% similarity and coverage threshold). Finally, tandem repeats were cleaned from the resulting non-redundant library using the cleanup_tandem.pl script from the EDTA pipeline (Ou et al., 2019) to produce the final set of TEs consensus sequences. Genome annotation of repeats was achieved using RepeatMasker v. 4.1.0 (Tarailo-Graovac & Chen, 2009) in sensitive mode (s parameter) with both Arthropoda RepBase (Bao et al., 2015) and our de novo non-redundant library. The repeat landscape was calculated using the parseRM.pl script (available at https://github.com/4ureliek/Parsing-RepeatMaskerOutputs; last accessed in December 2021) on the CpG-corrected Kimura divergence distance of each TE copy from its consensus.

For gene annotation, we used Maker v. 2 (Holt & Yandell, 2011) on the masked genome using proteins, RNA-seq evidence and ab initio gene predictors. For homology-based annotation, the transcriptome of *R. lucifugus* was aligned to the genome, together with the insect proteomes used above. A second Maker run was followed, providing maker gene models obtained with GeneMark-EP v. 3.63 (Brůna et al., 2020), automatically trained with the same publicly available proteomes used in previous steps. The resulting prediction was filtered, keeping only gene models that fully satisfied the following criteria: (i) spanning from start to stop codon; (ii) with an Annotation Edit Distance <0.2 (AED, a measure of the agreement of an annotation to external evidence, such as proteins and transcripts, with 0 indicating maximum agreement and 1 no support; Holt & Yandell, 2011); (iii) spaced by more than 1,000 bp and (iv) encoding protein products with no more than 80% sequence similarity between each other. The 3,598 resulting gene models were then used for SNAP (Korf, 2004) training. Augustus v. 3.3.3 (Stanke & Waack, 2003) was trained using Busco Run in the long mode, along with the Insecta odb10 database. Finally, trained Augustus, SNAP and Genemark HMM profiles were used in conjunction for the last comprehensive gene prediction round, excluding genes encoding proteins shorter than 50 amino acids. Proteins encoded by genes supported by an AED score of <0.5 were retained for subsequent analyses, forming our R. lucifigus reference gene set.

A final check for possible contaminants was performed at the protein level using the online implementation of the AAI-profiler with default parameters (Medlar et al., 2018). Finally, the resulting proteome was functionally annotated with InterProScan v. 5 (Jones et al., 2014) against the SUPERFAMILY, Pfam, PRINTS, PROSITE, ProDom, Gene3D, PANTHER and SMART databases.

Orthology prediction, phylogenomics and orthogroups evolutionary analyses

The filtered *R. lucifugus* proteins were used for phylogenetic inference and comparative genomic analyses, along with other six publicly available Blattodea proteomes (*B. germanica, Co formosanus and Cr. secundus, M. natalensis, R. speratus* and *Z. nevadensis*) and *L. migratoria* was added as the outgroup (Table S5). Predicted proteins were clustered in OGs using Orthofinder v. 2.5.2 (Emms & Kelly, 2019) with the ultra-sensitive DIAMOND option (Buchfink et al., 2015). Subsequently, phylogenomic inference was performed using a super-matrix approach. Single-copy ortholog genes were aligned with MAFFT v. 7.475 (Katoh & Standley, 2013) and concatenated, and the corresponding best-fit evolutionary model was identified with ModelFinder2 (Kalyaanamoorthy et al., 2017). IQ-TREE v. 2.0.3 (Minh et al., 2020) was used for species tree inference with nodal support values assessed with 1,000 ultra-fast bootstrap replicates. We inferred a time-calibrated phylogenetic tree using the least squares dating method (To et al., 2016), implementing a lognormal relaxed clock, with nodes constrained by secondary calibrations obtained from the timetree.org database (last accessed in September 2021; Table S6).

Gene family turnover evolutionary analyses were performed with CAFE v.5 (Mendes et al., 2020) using the inferred time tree, the OGs inferred by CAFE as being present at the root of the tree, an error model and a single lambda value for the whole tree. To avoid possible biases in the identification of expansion events due to the fragmentation of the assembly and, therefore, of the gene models (e.g., if a single gene is split into two different scaffolds, it is annotated as two separate genes), two different CAFE analyses were performed: (1) one with the complete *R. lucifugus* reference gene set using the same OGs previously inferred (FULL-CAFE analyses) and (2) one with only *R. lucifugus* complete gene models (i.e., spanning from start to stop codon; RED-CAFE analyses). For the latter analysis, Orthofinder was re-

run with the same parameters used above but included only complete gene models. In both analyses, significantly fast-evolving OGs in each branch of the tree were considered when they had a Viterbi $p < 0.05$, for the respective branch (i.e., OG with more copy number variation than expected by a null birth-death model; Hahn et al., 2007). As a further check for false positives in species-specific significantly expanded gene families identified by the RED-CAFE analyses, the first 30 largest significantly expanded gene families were separately subjected to an all-vs-all BLASTP search (e-value<1 e10, max_target_seqs = 5). After removing self-hits, the number of times two or more *R. lucifugus* homologues were aligned to two different regions of the same protein from other species was counted. In fact, if R. lucifugus gene counts in significantly expanded gene families are overestimated by gene fragmentation, we may expect to see two or more different *R. lucifugus* proteins aligned to two different, possibly non-overlapping, regions of the same homologue protein from other species. To check this, we extracted the query coordinates of the BLASTP results and used bedtools merge to remove redundant intervals, requiring a minimum overlap of 50 amino acids. The resulting bed files were manually checked to identify any false duplications.
We then tested for enrichment of GO terms, KO and KEGG pathways in the set of proteins clustered in significantly expanded gene families for both the FULL and RED-CAFE analyses. GO terms were assigned to all proteins using PANNZER2 with standard parameters, and only the GO terms having non-IEA (Inferred from Electronic Annotation) annotations in Arthropoda (Medlar et al., 2018) were considered. Enrichment analyses were performed for "biological processes" ontologies with the TopGO package in Bioconductor: the "classic" algorithm were used (Alexa & Rahnenführer, 2009), and GO-terms were considered to be significantly enriched with the Fisher exact test: $p < 0.01$. For KEGGS, functional annotation was performed using the EggNOG-mapper web tool (http://eggnog-mapper.embl.de/; last accessed December 2021; Cantalapiedra et al., 2021) with default parameters. For the enrichment test, we used the "enricher" function of the "Clusterprofiles" Bioconductor package (Yu et al., 2012).

Evolutionary analyses of P450 and OR gene families

Proteins belonging to the cytochrome P450 and OR gene families were extracted from the InterProScan results (IPR001128 and IPR004117 interpro codes, respectively) and selected for deeper evolutionary analyses. Proteins from the same families belonging to the model species *Drosophila melanogaster* and the eusocial hymenoptera *Apis mellifera, Camponotus floridanus, Nasonia vitripennis* and *Acromyrmex echinatior* were used for comparison (Table S5). Protein sequences from incomplete gene models and with a Pfam hit on the corresponding domain shorter than 60% of the expected length (341 and 255 amino acids, respectively, as reported in the Pfam database) were considered highly fragmented and not considered for phylogenetic analyses. Remaining sequences were aligned using MAFFT (e-INSi strategy), and spurious sequences and ambiguous positions (i.e., positions with gaps in more than 50% of the sequences) were removed using TrimAl (CapellaGutiérrez et al., 2009). Maximum likelihood gene tree inference was performed as described in the previous section.
The resulting gene family trees were automatically refined by removing outlier long branches and rogue taxa using Treeshrink v. 1.3.9 (Mai & Mirarab, 2018) and RogueNaRok v. 1 (Aberer et al., 2013), respectively; finally, they were corrected and reconciled to the underlying species tree using GeneRax v. 2.0.1 (Morel et al., 2020) with a duplication-loss model and a radius of 5. Hence, our species tree topology, which fully agrees with previous phylogenomics and mitogenomics studies among termites, was merged with the one obtained by Harrison et al. (2018) owing to their overlapping taxon sampling.

Author contributions

Jacopo Martelossi: Data curation (equal); formal analysis (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). Giobbe Forni: Formal analysis (equal); methodology (equal). Mariangela Iannello: Formal analysis (equal); methodology (equal). Castrense Savojardo: Data curation (equal); formal analysis (equal); methodology (equal). Pier Luigi Martelli: Supervision (equal). Rita Casadio: Supervision (equal). Barbara Mantovani: Funding acquisition (equal); supervision (equal); writing – review and editing (equal). Andrea Luchetti: Conceptualization (equal); formal analysis (equal); funding acquisition (equal); supervision (equal); writing – original draft (equal). Omar Rota-Stabelli: Conceptualization (equal); funding acquisition (equal); supervision (equal); writing – review and editing (equal).

Acknowledgement

Authors thank Silvia Ghesini for providing photos of termites. The work has been supported by Canziani Funding to AL and BM, and by COFUND Marie Curie—Provincia Autonoma di Trento post doc 2010-Incoming to ORS. Authors declare no conflict of interests.

Data Availability Statement

Raw reads were submitted to the NCBI Sequence Read Archive under acc. nos. SRR17267499 and SRR17267498. The final assembly is available on NCBI GenBank, under BioProject PRJNA778916 (accession number JAKCWW000000000). Supplementary data S1 and S2, the genome assembly, and gene annotations are available in Figshare under DOI: https://doi.org/10.6084/m9.figshare. 19115654.v2.

ORCID

Jacopo Martelossi https://orcid.org/0000-0003-4227-0669
Giobbe Forni https://orcid.org/0000-0003-3669-8693
Mariangela Iannello https://orcid.org/0000-0001-5736-1024
Castrense Savojardo https://orcid.org/0000-0002-7359-0633
Pier Luigi Martelli https://orcid.org/0000-0002-0274-5669
Rita Casadio https://orcid.org/0000-0002-7462-7039
Barbara Mantovani https://orcid.org/0000-0001-5175-7458

Andrea Luchetti https://orcid.org/0000-0002-2986-721X
Omar Rota-Stabelli https://orcid.org/0000-0002-0030-7788

REFERENCES

Aberer, A.J., Krompass, D. & Stamatakis, A. (2013) Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Systematic Biology, 62, 162–166.
Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. (2009) TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics, 25, 1329–1330.
Alexa, A. & Rahnenführer, J. (2009) Gene set enrichment analysis with topGO. Bioconductor Improvements, 27, 1–26.
Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6, 11.
Biedermann, P.H.W. & Taborsky, M. (2011) Larval helpers and age polyethism in ambrosia beetles. Proceedings of the National Academy of Sciences of the United States of America, 108, 17064–17069.
Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–2120.
Bourguignon, T., Lo, N., Cameron, S.L., Sobotník, J., Hayashi, Y., Shigenobu, S. et al. (2015) The evolutionary history of termites as inferred from 66 mitochondrial genomes. Molecular Biology and Evolution, 32, 406–421.
Brand, P. & Ramírez, S.R. (2017) The evolutionary dynamics of the odorant receptor gene family in corbiculate bees. Genome Biology and Evolution, 9, 2023–2036.
Brand, P., Robertson, H.M., Lin, W., Pothula, R., Klingeman, W.E., JuratFuentes, J.L. et al. (2018) The origin of the odorant receptor gene family in insects. eLife, 7, e38340.
Brůna, T., Lomsadze, A. & Borodovsky, M. (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics and Bioinformatics, 2, lqaa026.
Bucek, A., Sobotník, J., He, S., Shi, M., McMahon, D.P., Holmes, E.C. et al. ˇ (2019) Evolution of termite symbiosis informed by transcriptomebased phylogenies. Current Biology, 29, 3728–3734.e4.
Buchfink, B., Reuter, K. & Drost, H.G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods, 18, 366–368. 128
Buchfink, B., Xie, C. & Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12, 59–60.
Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S. et al. (2008) ALLPATHS: de novo assembly of wholegenome shotgun microreads. Genome Research, 18, 810–820.
Calla, B., Noble, K., Johnson, R.M., Walden, K.K.O., Schuler, M.A., Robertson, H.M. et al. (2017) Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths: birth, death and adaptation. Molecular Ecology, 26, 6021–6035.
Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E. et al. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiology, 164, 513–524.
Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. & HuertaCepas, J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular Biology and Evolution, 38, 5825–5829.
Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25, 1972–1973.
Chak, S.T.C., Harris, S.E., Hultgren, K.M., Jeffery, N.W. & Rubenstein, D.R. (2021) Eusociality in snapping shrimps is associated with larger genomes and an accumulation of transposable elements. Proceedings of the National Academy of Sciences of the United States of America, 118, e2025051118.
Cragg, S.M., Beckham, G.T., Bruce, N.C., Bugg, T.D.H., Distel, D.L., Dupree, P. et al. (2015) Lignocellulose degradation mechanisms across the tree of life. Current Opinion in Chemical Biology, 29, 108–119.
Crescente, J.M., Zavallo, D., Helguera, M. & Vanzetti, L.S. (2018) MITE tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics, 19, 348.
Dedeine, F., Dupont, S., Guyot, S., Matsuura, K., Wang, C., Habibpour, B. et al. (2016) Historical biogeography of Reticulitermes termites (Isoptera: Rhinotermitidae) inferred from analyses of mitochondrial and nuclear loci. Molecular Phylogenetics and Evolution, 94, 778–790.
Dedeine, F., Weinert, L.A., Bigot, D., Josse, T., Ballenghien, M., Cahais, V. et al. (2015) Comparative analysis of transcriptomes from secondary reproductives of three Reticulitermes termite species. PLoS One, 10, e0145596.
Dermauw, W., Van Leeuwen, T. & Feyereisen, R. (2020) Diversity and evolution of the P450 family in arthropods. Insect Biochemistry and Molecular Biology, 127, 103490.
Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology, 20, 238.
Emms, D.M. & Kelly, S. (2020) Benchmarking Orthogroup inference accuracy: revisiting Orthobench. Genome Biology and Evolution, 12, 2258–2266.
Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C. et al. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America, 117, 9451–9457.
Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28, 3150–3152.
Gonzalez, F.J. & Nebert, D.W. (1990) Evolution of the P450 gene superfamily: animal-plant ‘warfare’, molecular drive and human genetic differences in drug oxidation. Trends in Genetics, 6, 182–186.
Good, R.T., Gramzow, L., Battlay, P., Sztal, T., Batterham, P. & Robin, C. (2014) The molecular evolution of cytochrome P450 genes within and between drosophila species. Genome Biology and Evolution, 6, 1118–1134.
Govorushko, S. (2019) Economic and ecological importance of termites: a global review. Entomological Science, 22, 21–35.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology, 29, 644–652.
Hahn, M.W., Demuth, J.P. & Han, S.G. (2007) Accelerated rate of gene gain and loss in primates. Genetics, 177, 1941–1949.
Harrison, M.C., Jongepier, E., Robertson, H.M., Arning, N., BitardFeildel, T., Chao, H. et al. (2018) Hemimetabolous genomes reveal molecular basis of termite eusociality. Nature Ecology and Evolution, 2, 557–566.
Helvig, C., Koener, J.F., Unnithan, G.C. & Feyereisen, R. (2004) CYP15A1, the cytochrome P450 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora allata. Proceedings of the National Academy of Sciences of the United States of America, 101, 4024–4029.
Holt, C. & Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics, 12, 491.
Itakura, S., Yoshikawa, Y., Togami, Y. & Umezawa, K. (2020) Draft genome sequence of the termite, Coptotermes formosanus: genetic insights into the pyruvate dehydrogenase complex of the termite. Journal of Asia-Pacific Entomology, 23, 666–674.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C. et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236–1240.
Joseph, R.M. & Carlson, J.R. (2015) Drosophila chemoreceptors: a molecular interface between the chemical world and the brain. Trends in Genetics, 31, 683–695.
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.
Kapheim, K.M., Pan, H., Li, C., Salzberg, S.L., Puiu, D., Magoc, T. et al. (2015) Social evolution. Genomic signatures of evolutionary transitions from solitary to group living. Science, 348, 1139– 1143.
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.
Kokot, M., Długosz, M. & Deorowicz, S. (2017) KMC 3: counting and manipulating k-mer statistics. Bioinformatics, 33, 2759–2761.
Korb, J. & Hartfelder, K. (2008) Life history and development - a framework for understanding developmental plasticity in lower termites. Biological Reviews of the Cambridge Philosophical Society, 83, 295–313.
Korb, J., Poulsen, M., Hu, H., Li, C., Boomsma, J.J., Zhang, G. et al. (2015) A genomic comparison of two termites with different social complexity. Frontiers in Genetics, 6, 9. Korf, I. (2004) Gene finding in novel genomes. BMC Bioinformatics, 5, 59.
Koshikawa, S., Miyazaki, S., Cornette, R., Matsumoto, T. & Miura, T. (2008) Genome size of termites (Insecta, Dictyoptera, Isoptera) and wood roaches (Insecta, Dictyoptera, Cryptocercidae). Naturwissenschaften, 95, 859–867.
Krishna, K., Grimaldi, D.A., Krishna, V. & Engel, M.S. (2013) Treatise on the Isoptera of the world. Bulletin of the American Museum of Natural History, 377(2704), 1–200. Laetsch, D.R. & Blaxter, M.L. (2017) BlobTools: interrogation of genome assemblies. F1000Research, 6, 1287.
Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–3100.
Lu, K., Song, Y. & Zeng, R. (2021) The role of cytochrome P450-mediated detoxification in insect adaptation to xenobiotics. Current Opinion in Insect Science, 43, 103–107.
Luchetti, A., Scicchitano, V. & Mantovani, B. (2013) Origin and evolution of the Italian subterranean termite Reticulitermes lucifugus (Blattodea, Termitoidae, Rhinotermitidae). Bulletin of Entomological Research, 103, 734–741.
Luchetti, A., Velonà, A., Mueller, M. & Mantovani, B. (2013) Breeding systems and reproductive strategies in Italian Reticulitermes colonies (Isoptera: Rhinotermitidae). Insectes Sociaux, 60, 203–211.
Mai, U. & Mirarab, S. (2018) TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics, 19, 272.
Mao, H. & Wang, H. (2017) SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics, 33, 743–745.
Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B.J. (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics, 33, 574–576.
Matsuura, K. (2010) Sexual and asexual reproduction in termites. In: Bignell, D., Roisin, Y. & Lo, N. (Eds.) Biology of termites: a modern synthesis. Dordrecht: Springer.
Matsuura, K. (2017) Evolution of the asexual queen succession system and its underlying mechanisms in termites. Journal of Experimental Biology, 220, 63–72.
Matsuura, K., Vargo, E.L., Kawatsu, K., Labadie, P.E., Nakano, H., Yashiro, T. et al. (2009) Queen succession through asexual reproduction in termites. Science, 323, 1687.
Medlar, A.J., Törönen, P. & Holm, L. (2018) AAI-profiler: fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. Nucleic Acids Research, 46, W479–W485.
Mendes, F.K., Vanderpool, D., Fulton, B. & Hahn, M.W. (2020) CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics, 36, 5516–5518.
Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. et al. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution, 37, 1530–1534.
Morel, B., Kozlov, A.M., Stamatakis, A. & Szöllosi, G.J. (2020) GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Molecular Biology and Evolution, 37, 2763–2774.
Nagnan, P. & Clement, J.L. (1990) Terpenes from the maritime pine Pinus pinaster: toxins for subterranean termites of the genus Reticulitermes (Isoptera: Rhinotermitidae)? Biochemical Systematics and Ecology, 18, 13–16.
Nishimura, O., Hara, Y. & Kuraku, S. (2019) Evaluating genome assemblies and gene models using gVolante. In: Kollmar, M. (Ed.) Methods in molecular biology, Vol. 1962. New York, NY: Springer, pp. 247–256.
Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J. et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology, 20, 275.
Poulsen, M., Hu, H., Li, C., Boomsma, J.J. & Zhang, G. (2014) Macrotermes natalensis genome assembly data. GigaScience Database. Available from: https://doi.org/10.5524/100057
Rane, R.V., Ghodke, A.B., Hoffmann, A.A., Edwards, O.R., Walsh, T.K. & Oakeshott, J.G. (2019) Detoxifying enzyme complements and host use phenotypes in 160 insect species. Current Opinion in Insect Science, 31, 131–138.
Robertson, H.M., Baits, R.L., Walden, K.K.O., Wada-Katsumata, A. & Schal, C. (2018) Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach Blattella germanica. Journal of Experimental Zoology Part B, Molecular and Developmental Evolution, 330, 265–278.
Robertson, H.M., Gadau, J. & Wanner, K.W. (2010) The insect chemoreceptor superfamily of the parasitoid jewel wasp Nasonia vitripennis. Insect Molecular Biology, 19(Suppl 1), 121–136.
Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L.B. & Touhara, K. (2008) Insect olfactory receptors are heteromeric ligandgated ion channels. Nature, 452, 1002–1006.
Scharf, M.E. & Tartar, A. (2008) Termite digestomes as sources for novel lignocellulases. Biofuels, Bioproducts and Biorefining, 2, 540–552.
Seppey, M., Ioannidis, P., Emerson, B.C., Pitteloud, C., RobinsonRechavi, M., Roux, J. et al. (2019) Genomic signatures accompanying the dietary shift to phytophagy in polyphagan beetles. Genome Biology, 20, 98.
Sezutsu, H., Le Goff, G. & Feyereisen, R. (2013) Origins of P450 diversity. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 368, 20120428.
Shigenobu, S., Hayashi, Y., Watanabe, D., Tokuda, G., Hojo, M.Y., Toga, K. et al. (2022) Genomic and transcriptomic analyses of the subterranean termite Reticulitermes speratus: gene duplication facilitates social evolution. Proceedings of the National Academy of Sciences of the United States of America, 119, e2110361119.
Simola, D.F., Wissler, L., Donahue, G., Waterhouse, R.M., Helmkampf, M., Roux, J. et al. (2013) Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Research, 23, 1235–1247.
Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. & Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. Genome Research, 19, 1117–1123.

Smadja, C., Shi, P., Butlin, R.K. & Robertson, H.M. (2009) Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Molecular Biology and Evolution, 26, 2073–2086.
Stanke, M. & Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 19(Suppl 2), ii215– ii225.
Tarailo-Graovac, M. & Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics Chapter, 25, 4.10.1–4.10.14.
Tartar, A., Wheeler, M.M., Zhou, X., Coy, M.R., Boucias, D.G. & Scharf, M. E. (2009) Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite R. flavipes. Biotechnology for Biofuels and Bioproducts, 2, 25.
Tarver, M.R., Coy, M.R. & Scharf, M.E. (2012) Cyp15F1: a novel cytochrome P450 gene linked to juvenile hormone-dependent caste differention in the termite Reticulitermes flavipes. Archives of Insect Biochemistry and Physiology, 80, 92–108.
Terrapon, N., Li, C., Robertson, H.M., Ji, L., Meng, X., Booth, W. et al. (2014) Molecular traces of alternative social organization in a termite genome. Nature Communications, 5, 3636.
To, T.H, Jung, M., Lycett, S. & Gascuel, O. (2016) Fast dating using least-squares criteria and algorithms. Systematic Biology, 65, 82–97.
Vargo, E.L. & Husseneder, C. (2010) Genetic structure of termite colonies and populations. In: Bignell, D., Roisin, Y. & Lo, N. (Eds.) Biology of termites: a modern synthesis. Dordrecht: Springer.
Vargo, E.L., Labadie, P.E. & Matsuura, K. (2012) Asexual queen succession in the subterranean termite Reticulitermes virginicus. Proceedings of the Royal Society B. Biological Sciences, 279, 813–819.
Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J. et al. (2017) GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 33, 2202–2204.
Walsh, A.T., Triant, D.A., Le Tourneau, J.J., Shamimuzzaman, M. & Elsik, C. G. (2022) Hymenoptera genome database: new genomes and annotation datasets for improved go enrichment and orthologue analyses. Nucleic Acids Research, 50, D1032–D1039.
Wang, M., Hellemans, S., Sobotník, J., Arora, J., Bucek, A., Sillam-Dussès, D. et al. (2022) Phylogeny, biogeography and classification of Teletisoptera (Blattaria: Isoptera). Systematic Entomology, 47, 581–590.
Wilson, E.O. & Hölldobler, B. (2005) Eusociality: origin and consequences. Proceedings of the National Academy of Sciences of the United States of America, 102, 13367–13371.
Xiong, W., He, L., Lai, J., Dooner, H.K. & Du, C. (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proceedings of the National Academy of Sciences of the United States of America, 111, 10263–10268.
Xue, W., Li, J.T., Zhu, Y.P., Hou, G.Y., Kong, X.F., Kuang, Y.Y. et al. (2013) L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics, 14, 604.
Yu, G., Wang, L.G., Han, Y. & He, Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS, 16, 284–287.
Zhou, X., Rokas, A., Berger, S.L., Liebig, J., Ray, A. & Zwiebel, L.J. (2015) Chemoreceptor evolution in Hymenoptera and its implications for the evolution of eusociality. Genome Biology and Evolution, 7, 2407–2416.

Supporting Information

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1: Assembly spectra copy number plots showing a nearly complete and homozygous assembly. K-mers present in the reads but missing from the assembly are represented in black, whereas k-mers present in the reads and just once in the assembly are represented in red

Figure S2: Cumulative distribution of the AED values of annotated protein-coding genes. All genes with an AED value of <0.5 (95% of total genes) were retained in the final gene set

Figure S3: Taxonomic annotation of *Reticulitermes lucifugus* proteome

Table S1: Sequencing results, number of reads retained after trimming and number of reads classified as possible bacterial contaminations

Table S2: Genome size and heterozygosity relative to different k-mer sizes

Table S3: Details of the number of sequences used for P450 gene family analysis. The final number of sequences refers to the total number of sequences used after filtering employing Trimal, RogueNaRok and Treeshrink for the final phylogenetic inference

Table S4: Details of the number of sequences used for OR gene family analysis. The final number of sequences refers to the total number of sequences used after filtering employing Trimal, RogueNaRok and Treeshrink for the final phylogenetic inference

Table S5: Source and accession numbers of publicly available proteomes used in this study Table S6: Calibration points used for divergence time estimation

Data S1: P450 gene tree

Data S2: Odourant receptor gene tree

How to cite this article: Martelossi, J., Forni, G., Iannello, M., Savojardo, C., Martelli, P.L., Casadio, R. et al. (2023) Wood feeding and social living: Draft genome of the subterranean termite Reticulitermes lucifugus (Blattodea; Termitoidae). Insect Molecular Biology, 32(2), 118–131. Available from: https://doi. org/10.1111/imb.12818