Title: High-quality long-read genome assemblies reveal evolutionary patterns of transposable elements and DNA methylation in termites

Bitao Qiu[1,*], Daniel Elsner[1], Judith Korb[1.2.*]

[1]Evolutionary Biology & Ecology, University of Freiburg, Germany

[2]Research Institute for the Environment and Livelihoods, Charles Darwin University, Casuarina Campus, NT0909 Darwin, Australia

[*]corresponding authors

ORCID

BQ: 0000-0003-4588-7680

DE: 0000-0002-1397-7617

JK: 0000-0001-9577-9376

Summary

Termites are social cockroaches that evolved eusociality independently from social Hymenoptera (ants, bees, wasps). Termites' large genome sizes, generally composed of around 50% of transposable elements, make it challenging to obtain high quality genome assemblies with short-read sequencing technology. Correspondingly, there are only six published termite genome assemblies. Using PacBio HiFi sequencing, we generated high quality genome assemblies using single individuals for three additional termite species (*Trinervitermes* sp.1, *Odontotermes* sp.2, *Macrotermes bellicosus*) and a highly improved assembly for a fourth species (*Cryptotermes secundus*). We simultaneously profiled TE abundance and CpG DNA methylation from the same individuals. We show that TE abundance and DNA methylation levels across species reflect their phylogenetic relationships. We identified lineage- and species-specific TE- and DNA-methylation patterns, and revealed a negative correlation between TE abundance and methylation level. These results imply evolutionary arms-races between TE and DNA methylation during termite evolution.

Key-words: DNA methylation, genomes, HiFi sequencing, social insect, termites, transposable elements

Introduction

Termites (Infraorder Isoptera) play an important role in many ecosystems where they are the main decomposers and function as true ecosystem engineers that are essential for maintenance of biodiversity (Bignell & Eggleton 1997, Pringle et al. 2010). Yet, a few species are also known as obnoxious pests that cause damage in the range of billions of US$ annually (Su 2019). The termites' success can be attributed to their social life. In fact, termites are the oldest eusocial insects. They evolved eusociality independently from social Hymenoptera (ants, bees, wasps) more than 130 million years ago as a monophyletic clade in the Blattodean (cockroach) lineage (Inward et al. 2007). All termites belong to the eusocial insects due to the presence of sterile soldiers. However, the degree of social complexity varies widely, associated with the workers' reproductive capacity (ranging from totipotent to sterile) and phylogenetic position (Korb & Heinze 2016).

A challenge to provide high quality genome assemblies of termites with short-read sequences can be their size of around 1.3 GB and their high repeat element content (including transposable elements, TEs) that accounts for roughly half of the genome size (e.g., Koshikawa et al. 2008, Poulsen et al. 2014, Korb et al. 2015, Shigenobu et al. 2022; exception: *Zootermopsis nevadenesis*: Terrapon et al. 2014). In comparison, the better studied social Hymenoptera typically have genomes about one third of this size and much fewer TEs. Accordingly, among the ca. 3,000 termite species, only six termite genome assemblies have been published so far, all based on short-read sequencing technologies: *Zootermopsis nevadensis* (Archotermopsidae) (Terrapon et al. 2014), *Macrotermes natalensis* (Termitidae: Macrotermitinae) (Poulsen et al. 2014), *Cryptotermes secundus* (Kalotermitidae) (Harrison et al. 2018), *Coptotermes formosanus* (Rhinotermitidae) (Itakura et al. 2020), *Reticulitermes speratus* (Rhinotermitidae) (Shigenobu et al. 2022) and *Reticulitermes lucifugus*, (Rhinotermitidae) (Martelossi et al. 2023).

Importance of TEs, in general and especially in termites

A major portion of the repeat sequences in termites and elsewhere are TEs. TEs are selfish genetic elements that can replicate independently within the genome. Their presence can be regarded as double-edged swords. They can cause costs for the individuals bearing them, for instance, when active TEs 'jump' and translocate within the genome interfering with gene regulation or gene function, or when they cause ectopic recombination (i.e., crossing overs at homologous sites among non-homologous loci) (e.g., González & Petrov 2012). On the other side, they can be a source of innovation as ectopic recombination can lead to gene duplication that can result in neofunctionalization of gene copies (e.g., Zhang 2003, Innan & Kondrashov 2010, González & Petrov 2012, and references therein).

In termites, TEs have been associated with gene duplications during termite social evolution (Terrapon et al. 2014, Korb et al. 2015, Korb 2016, Harrison et al. 2018, Shigenobu et al. 2022). It has been proposed that TEs-facilitated gene duplications have been co-opted during social evolution for caste-specific gene expression and the evolution of caste phenotypes (Gadagkar 1997, Korb et al. 2015, Korb 2016, Shigenobu et al. 2022). On the other hand, TEs in termites also seem to have costs. Like in other animals (e.g. De Cecco et al. 2013; Wood &

Helfand 2013; Wood et al. 2016, Sturm et al. 2015), TE activity increases with age and it has been proposed as a major cause of ageing to explain the short lifespan of termite workers compared to queens and kings (Elsner et al. 2018, Post et al. 2023).

Role of DNA methylation in TE defence

A mechanism to control TEs activity is epigenetic silencing. DNA methylation, especially deposition of 5-methylcytosine (5mC), represses TEs in many higher eukaryotes (e.g. Deniz et al. 2019 and references therein). Yet, its role in insects is less clear. Comparative phylogenetic analyses revealed that TE methylation largely varies across insects. Some species show no methylation of TEs (like the fly *Drosophila melanogaster*) while the German cockroach, *Blattella germanica*, exhibits considerable levels of TE methylation (Lewis et al. 2020). Current analyses suggest that DNA methylation as a mechanism to control TE activity might have been lost in holometabolous insects, while it is common in other arthropods (Lewis et al. 2020). Previous studies found that DNA methylation is widespread in termite genomes (e.g., Terrapon et al. 2014, Glastad et al. 2016, Harrison et al. 2022) and may indicate an association between DNA methylation and caste difference (Gladstad et al. 2016, Harrison et al. 2022, reviewed in Okwaro & Korb 2023). However, the evolutionary pattern and the role of DNA methylation in TE defence in termites has never been studied.

To study the evolution of TE abundance and DNA methylation across the termite phylogeny, we used PacBio HiFi (high-fidelity) sequencing to generate high-quality genome assemblies of four termite species. These species reflect the full sociality spectrum of termites and cover a wide range of their phylogeny. The species comprise the Macrotermitinae *Macrotermes bellicosus* and *Odontotermes* sp.2, which are fungus-growing termites that live in an obligate symbiosis with *Termitomyces* fungi that they culture within their nests (Nobré et al. 2011); the Nasutitermitinae *Trinervitermes* sp.1, which also belongs to the highly eusocial Termitidae like the fungus-growers, and *Cryptotermes secundus*, a drywood termite (Kalotermitidae) of low social complexity (**Figure 1A**). All four species are ecological dominant in their respective habitats. For further information on the species, see also the Supplementary Material.

We simultaneously profiled the TE abundance and the CpG DNA methylation levels of the same individuals. We show that both TE abundance and DNA methylation exhibit a strong phylogenetic signal and that some TE families are highly methylated. Our study indicates an arm-race between TE and DNA methylation in termites and suggests an important role of DNA methylation in accommodating a high level of potentially harmful TEs in termite genomes.

Results

New high-quality long-read assemblies for four termite species

We used the PacBio HiFi long-read system to sequence high-molecular weight DNA extracted from single individual samples of four termite species (for details, see Methods): *C.*

secundus, *Trinervitermes* sp.1, *Macrotermes bellicosus* and *Odontotermes* sp.2 (**Figure 1A**). For *M. bellicosus*, we sequenced both a queen and a soldier sample to compare a potential caste effect on TE abundance and DNA methylation. For each individual, we generated 18 to 40 Gb HiFi reads, with mean read length from 7,139 to 16,384 bp and estimated sequencing depth from 16 to 37x (**Table 1**).

Assembly	***C. secundus***	***Trinervitermes* sp.1**	***Odontotermes* sp.2**	***M. bellicosus* (Q)**	***M. bellicosus* (S)**
Caste	Worker	Worker	Queen	Queen	Soldier
Library method	HiFi low input	HiFi low input	HiFi low input	HiFi standard	HiFi low input
Tissue	Gut-removed whole bodies	Gut-removed whole bodies	Fat body	Fat body	Gut-removed whole bodies
# contigs	2,091	3,287	441	187	3,339
Total length (bp)	1,121,125,343	1,462,145,465	1,325,203,851	1,273,012,709	1,254,704,789
GC (%)	41.17	40.47	40.05	40.25	40.22
Largest contig (bp)	7,090,375	6,339,778	22,652,318	50,731,358	3,583,072
N50 (bp)	1,213,394	805,561	7,092,793	16,378,181	643,405
L50	277	525	63	26	607
auN (bp)	1,539,826	1,066,953	7,844,277	17,577,793	777,220
Coverage	20	16	37	27	17
Total reads	2,483,158	1,994,195	4,418,500	1,773,689	3,613,653
Raw data (Gb)	17.73	19.15	39.84	29.06	18.17
mean Read length (bp)	7,139	9,598	9,010	16,384	5,028
Complete BUSCOs	1353 (99.0%)	1335 (97.6%)	1359 (99.3%)	1360 (99.5%)	1353 (99.0%)
Complete and single-copy BUSCOs (S)	1312 (96.0%)	1306 (95.5%)	1344 (98.3%)	1340 (98.0%)	1330 (97.3%)
Complete and duplicated BUSCOs (D)	41 (3.0%)	29 (2.1%)	14 (1.0%)	20 (1.5%)	23 (1.7%)
Fragmented BUSCOs (F)	6 (0.4%)	16 (1.2%)	4 (0.3%)	0 (0.0%)	5 (0.4%)
Missing BUSCOs (M)	8 (0.6%)	16 (1.2%)	5 (0.4%)	7 (0.5%)	9 (0.6%)

Table 1. Genome assembly information and statistics.

We assembled our termite HiFi reads with *Hifiasm* (Cheng et al. 2021) and removed potential false genome duplications with *purge dup* (Guan et al. 2020). We then obtained five genome assemblies with sizes of around 1.3 Gb, similar to previously published termite genomes (Poulsen et al. 2014, Harrison et al. 2018) (**Table 1**). Our genome assemblies have a genome completeness between 97.6% and 99.5% (**Table 1**) and median mapping rates of 93% and 94% with previous published RNAseq data in *C. secundus* and *M. bellicosus*, respectively (see Methods). In particular, both the genome completeness and the RNAseq mapping rate of our *C. secundus* long-read genome assembly were higher than the previous *C. secundus* short-read genome assembly (**Supplementary Figure 1**), which required DNA input from 1000 individuals from the same colony (Harrison et al. 2018). Our results show that the HiFi approach can generate high-quality genome assemblies efficiently from single individual insects.

The distribution of TE abundances in termite genomes

To examine the distribution of transposable elements (TEs) and other repeats across termites, we first constructed a non-redundant Blattodea repeat database that contains consensus sequences for homologous repeats. We next quantified the abundance of homologous repeats in our five termite genome assemblies (i.e., the queen or the worker of four different species plus the additional soldier of *M. bellicosus*) and the published *B. germanica* genome assembly (Harrison et al. 2018).

The cluster dendrogram of abundances of different repeat types reflects the phylogeny of the termite species (**Figure 1B**), and this pattern remains the same when we excluded repeats from unknown families. Accordingly, the repeat abundances of the two castes of *M. bellicosus* were most similar, followed by those of the other fungus-growing termite *Odontotermes* sp.2. Next came *Trinervitermes* sp.1, the third Termitidae. *C. secundus*, a representative of the drywood termites (Kalotermitidae), and the non-termite cockroach, *B. germanica*, had the most dissimilar repeat abundances.

The phylogenetic signal of repeat abundances was associated with lineage/species-specific TE expansions (**Figure 1C**). For example, genomes of Termitidae had the highest abundance of long terminal repeats (LTR, Class I TEs: retrotransposons), the majority of which were from unknown LTR super-families (**Figure 1C**). By contrast, the *C. secundus* genome had the highest abundance of Copia and Gypsy LTRs. Concerning DNA transposons (Class II TEs), genomes of Termitidae had a high abundance of Mariner and TC1 elements, while the *C. secundus* genome contained a high abundance of TC2 elements. Interestingly, the *M. bellicosus* soldier and queen genomes always had equal proportions of all major TE super-families, indicating TE expansion at the phylogenetic level is much stronger than at the colony and/or individual level. Compared to the termites, the *B. germanica* genome had very few LTRs. This could be because the published *B. germanica* genome assembly was

generated by short-read technology. It is hard to detect LTRs with short-read genome assemblies.

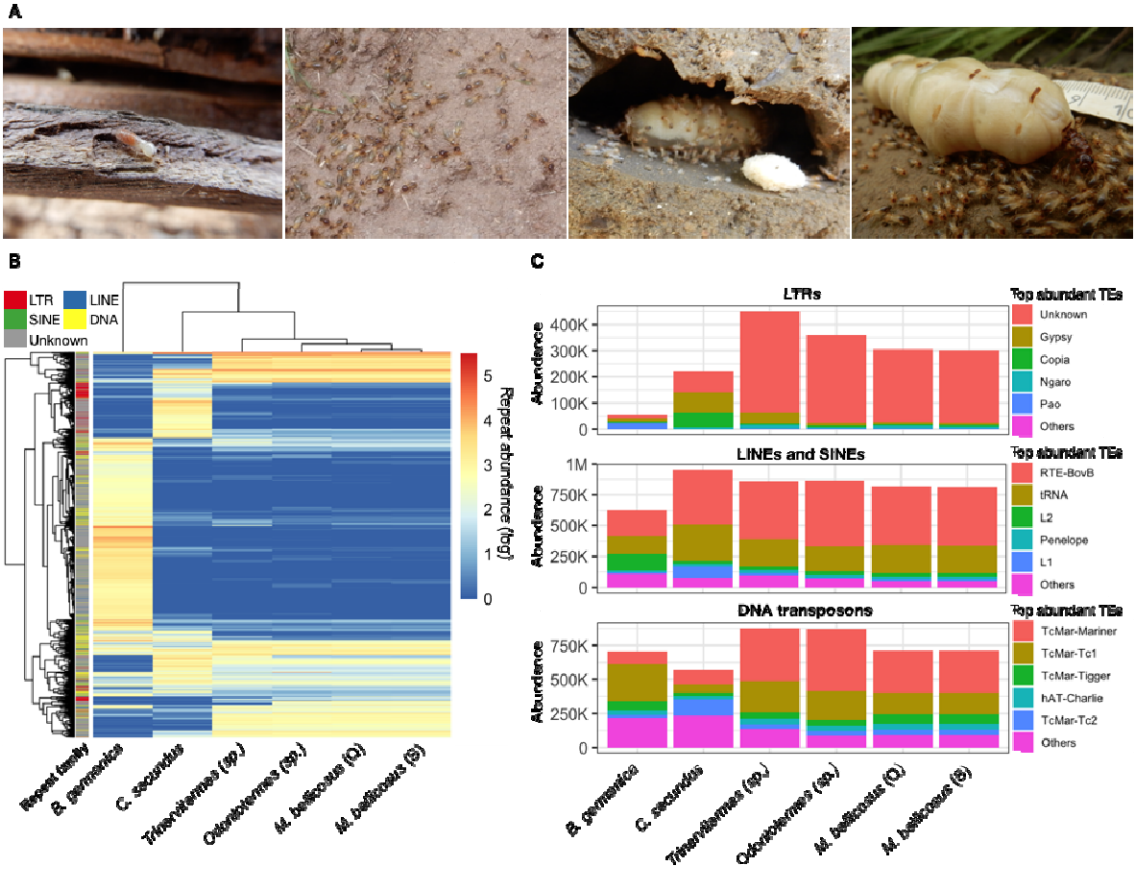

Figure 1. TEs in termite genomes. (A) The four sequenced termite species. From left to right: *Cryptotermes secundus*, *Trinervitermes* sp.1, *Odontotermes* sp.2. and *Macrotermes bellicosus*. (B) Heatmap of repeat abundances (log scale) for five Blattodean species, including a queen and a soldier for *M. bellicosus*. Each row represents one homologous repeat sequence (family), with its repeat category annotated on the left. (C) Bar plots for the abundance of major TE categories, Class I (retrotransposon, including LTR, LINE, and SINE) and Class II (DNA transposons), in five Blattodean species, including the two *M. bellicosus* castes. Only the top five abundant TE super-families for each TE category are shown. The remaining TE super-families are summarised as 'Others'.

The distribution of DNA CpG methylation in termite genomes

To quantify the genomic DNA methylation levels in termites, we used the kinetic information from the PacBio HiFi reads to generate the methylation (5mC) probabilities for CpG sites in our genome assemblies (see Methods). The percentages of methylated CpG sites in our five genome assemblies ranged from 6.0 % to 11.0%, with the lowest levels in *Odontotermes* sp.2 and the highest in *C. secundus*. For *M. bellicosus*, the percentage of methylated CpG sites in

the soldier genome assembly is slightly but significantly higher than in the queen genome assembly (9.6% vs. 8.9%, p < 10e-5, two-sided binomial test).

Next, we examined the percentage of methylated CpG sites across different genomic regions. Exonic regions had significantly higher percentages of methylated CpGs than the genomic background (p < 10e-5, two-sided binomial test), whereas intergenic regions were significantly depleted of methylated CpG sites (p < 10e-5, two-sided binomial test) (**Figure 2A**), consistent with previous findings for termite genomes and those of other insects (Terrapon et al. 2014, Glastad et al. 2016, Lewis et al. 2020, Harrison et al. 2022). In addition, CpG methylation percentages of Class II TEs (DNA transposons) were in all species significantly higher than those of the genomic backgrounds (p < 10e-5, two-sided binomial tests) (**Figure 2B**). CpG methylation levels of Class I TEs (retrotransposons) were more variable across species. The two Macrotermitinae, *M. bellicosus* and *Odontotermes* sp.2, were significantly less methylated than expected, whereas *Trinervitermes* sp.1 and *C. secundus* were more methylated (p < 10e-5, two-sided binomial tests). The low complexity repeats were significantly depleted of methylated CpG sizes (p < 10e-5, two-sided binomial tests).

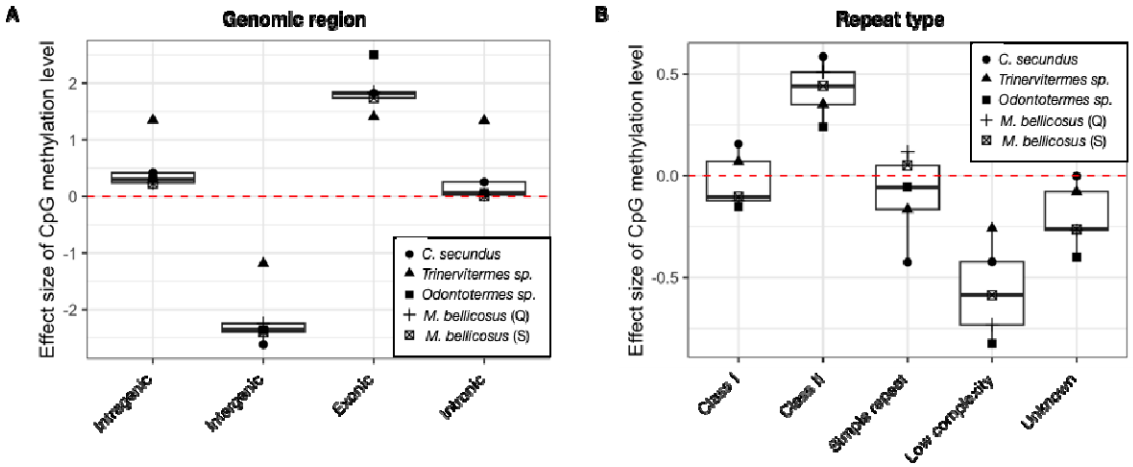

Figure 2. CpG DNA methylation patterns in termite genomes. (A) Percentages of methylated CpG sites across different genomic regions in the five termite genome assemblies. (B) Percentages of methylated CpG sites across major repeat classes in the five termite genome assemblies. Effect sizes of CpG methylation levels are calculated as log2 (Percentage of methylated CpG sites / Percentage of methylated CpG sites in the genome).

The association between DNA CpG methylation and TEs in termites

We further examined the CpG methylation patterns for each TE family across the four termite species. Principle component analysis (PCA) revealed a strong lineage/species effect for the CpG methylation pattern of repeat sequences (**Figure 3A**). PC1 separated the *C. secundus* genome assembly from the other genome assemblies that belong to the Termitidae and PC2 separated the Macrotermitinae genome assemblies from that of *Trinervitermes* sp.1 (see

Supplementary Table 1 for the associated TE families in each PC). Clustering analysis showed a similar pattern, with the dendrogram of repeat CpG methylation levels reflecting the phylogeny of the termite species (**Supplementary Figure 2**).

Comparing the methylated CpG pattern of TE super-families across termite species revealed both shared and lineage/species-specific patterns (**Figure 3B**). The Class I LINE retrotransposons R2 and L2 were highly methylated across species, as were PiggyBac- and P-DNA transposons (Class II). In addition, some species-specific signals seem to exist. For instance, *M. bellicosus* had the highest methylation of the DNA transposons Tc2 and MULE, which separated *M. bellicosus* from *Odontotermes* sp.2. *Trinervitermes* sp.1 had a high species-specific methylation of the DNA transposon haT-haT19 and the high methylation of the Kolobok-Hydra DNA transposons separated it from the other Termitidae. Compared to the Termitidae, *C. secundus* was characterized by a high methylation of DNA transposons, especially from the Pogo superfamily. Interestingly, as *Trinervitermes* sp.1, *C. secundus* also had high methylation levels of the Kolobok-Hydra DNA transposons, which could be the ancestral state of the Termitidae (but subsequently reduced in Macrotermitinae) or convergent evolution.

As both TE abundance and TE CpG methylation levels exhibit a strong phylogeny signal, we further examined the association between TE abundance and CpG methylation percentage for the 962 TE families (**Figure 3C**). For all genome assemblies, percentages of methylated CpGs correlated negatively with TE abundance, regardless of the TE categories (all $p < 10e-5$ except for SINE which was not significant, two-sided Spearman's rank test) (**Figure 3C**). A detailed analysis with a generalised linear model shows that this correlation pattern can be super-family dependent (**Supplementary Table 2**). For instance, within Jockey, a LINE TE super-family, TE-family-5448 was highly methylated and with low abundance in all five termite genome assemblies, whereas TE-family-2292 and TE-family-2449 were lowly methylated and with high abundance (**Supplementary Figure 3**). Because DNA methylation can suppress TE activity, the family difference within a super-family indicates the negative correlation between TE abundance and TE CpG methylation levels can be driven by the TE-family specific evolutionary history.

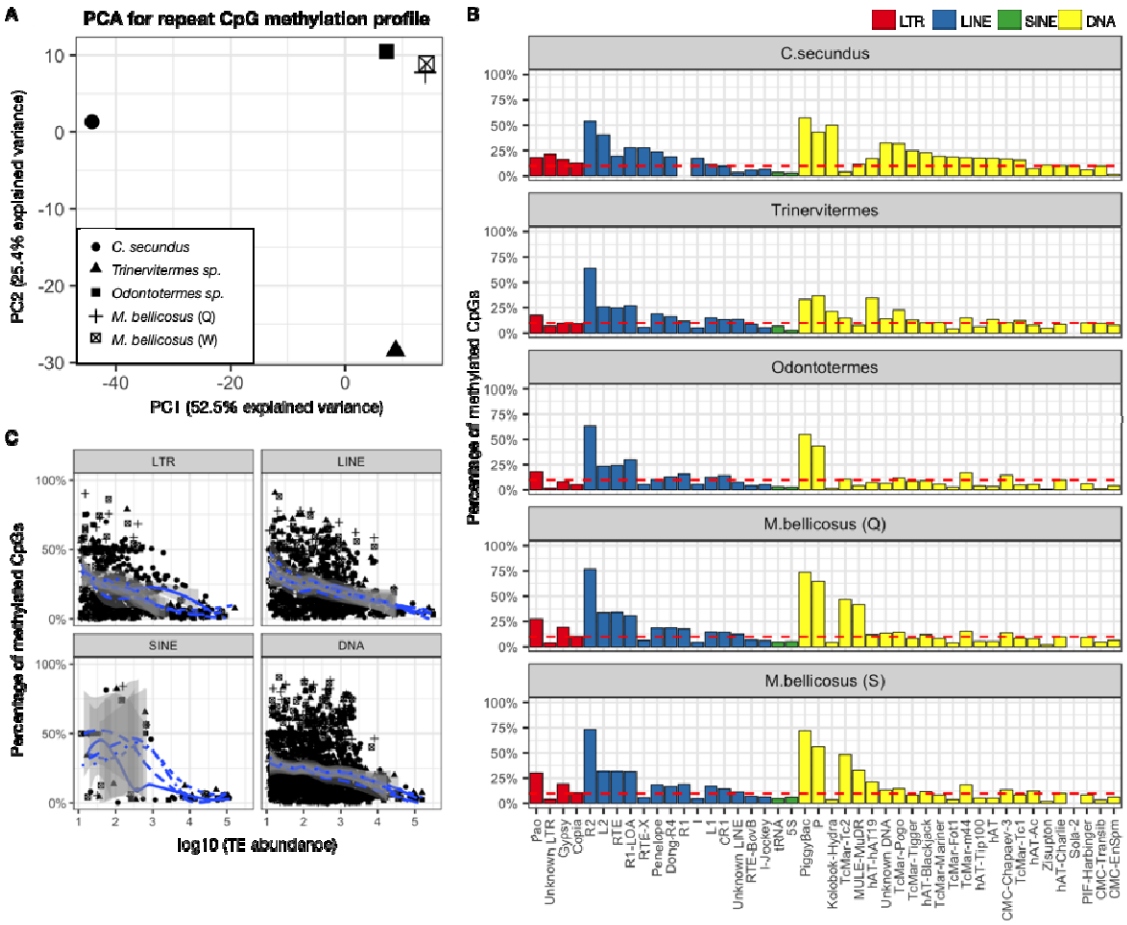

Figure 3. TE CpG DNA methylation level in termite genomes. (A) Results of a PCA of CpG methylation patterns across the five termite genome assemblies, using the methylation data of all 962 TE families. (B) Percentages of methylated CpGs across TE super-families, grouped and coloured by major TE categories and plotted separately for each genome assembly. *M. bellicosus* queen (Q) and soldier (S) genome assemblies have a similar pattern for repeat CpG methylation. Only TE super-families with > 30 CpG sites and > 5% methylated CpGs in at least one genome assembly are shown. The red dashed lines show the 10% methylation level that represents the genomic background in termites. (C) Plot of percentages of methylated CpGs against TE abundance for all 962 TE families, plotted separately for each TE category. The associations were shaped by species identity but the five termite genome assemblies shared similar patterns of declining CpG methylation with TE abundance, except for SINE elements.

Discussion

Our study provided high quality genome assemblies for three new termite species across the termite phylogeny and improved the quality of the fourth one, *C. secundus*, that is available from short-read sequencing only (Harrison et al. 2018). Our results show that PacBio HiFi sequencing can not only generate high quality genome assemblies from single individuals of small insect species but even more strikingly, CpG methylation can be inferred from the same samples. This approach is especially important for the assembly of large genomes with a high TE content, as is typical for termites (Koshikawa et al. 2008, Korb et al. 2015, Poulsen et al. 2014, Harrison et al 2018, Shigenobu et al. 2022).

In addition, our comparative analyses across the four termite species revealed that the similarity of repeat (including TEs) abundances reflects termite relatedness at the phylogenetic scale. In line with this result, we could identify lineage and species-specific TEs that seem to characterize certain taxa, implying a TE history mirroring that of their termite hosts. The higher-than-expected CpG DNA methylation of TEs compared to the background contrast with results for holometabolous insects in which TEs seem to be depleted of methylation (Lewis et al. 2020). It supports the notion that TE methylation might be restricted to non-holometabolous arthropods, in line with results for the German cockroach, *B. germanica*, and some other hemimetabolous insects, which also have methylated TEs (Lewis et al. 2020).

The lineage and species-specific methylation of certain TE categories supports a view of co-evolutionary events, in which certain TEs became silenced via TE methylation (**Supplementary Figure 4**) (Zhou et al. 2020). This view is backed by our results of a negative association between TE abundance and CpG methylation, which suggests that methylated TEs can no longer proliferate. Overall, our results support the hypothesis of an evolutionary scenario in which cycles of specific TE invasions, followed by their proliferation and increase in abundance, are stopped by DNA methylation. These co-evolutionary 'arms races' lead to TE abundance similarities reflecting species phylogeny. Under neutral evolution of TEs we would also expect TE abundance similarity to reflect species phylogeny. However, such a neutral view of TE evolution would imply no association between TE abundance and CpG methylation. This is not in line with our results of the negative association between TE abundance and CpG methylation. Finer scaled analyses with more samples per species and more detailed analyses of single striking TE families are necessary to test this hypothesis. The co-evolution of DNA methylation and TE allows termite genomes to accommodate abundant, potentially harmful, TEs and lead to new TE-derived genetic regulation, which could contribute to the evolution of caste phenotypes in termite sociality.

The strong phylogenetic signal of DNA methylation in termite genomes either implies that the re-establishment of DNA methylation marks during each generation is DNA-sequence-specific (Heard & Martienssen 2014), or, more likely, it supports the emerging hypothesis of

heritability of DNA methylation in insects (Harris et al. 2019, Yagound et al. 2020), in which DNA methylation reprogramming seems to be absent during development. Because DNA methylation might be functional in termites (Okwaro & Korb 2023), the heritability of DNA methylation suggests that genomic DNA methylation patterns could be under natural selection. Future studies will benefit from using long-read sequencing to profile the DNA methylation pattern with pedigree samples and conducting experimental manipulation to examine the adaptive value of DNA methylation.

Methods

Sample collection

A *M. bellicosus* and an *Odontotermes* sp.2 queen had been collected in the Comoé National Park, Côte d'Ivoire, in 2017 and 2019, respectively. The *M. bellicosus* queen was stored in 100% ethanol and the *Odontotermes* sp.2 queen in RNA later at -20 °C until DNA extraction. A *M. bellicosus* soldier (minor caste) had been collected in the Pendjari National Park, Benin, in 2017 and stored in 100% ethanol until DNA extraction. For the *C. secundus* and *Trinervitermes* sp.1, a worker was used for genome sequencing. The corresponding *C. secundus* colony was collected at Channel Island, Australia, in 2019, and kept in a climate chamber in Freiburg, Germany, until it was freshly killed to extract its DNA. The *Trinervitermes* sp.1 worker was collected in Lamto, Côte d'Ivoire, in 2021 and stored in 100% ethanol at -20 °C until DNA extraction. For all samples, the termite gut was excluded before DNA extraction to avoid contamination with gut symbionts.

DNA extraction and sequencing

Because body sizes of the four termite species differ by several orders of magnitude, genomic DNA was either extracted from dissected fat bodies for queen samples of *M. bellicosus* and *Odontotermes* sp.2 or from gut-removed whole bodies for soldier/worker samples of *M. bellicosus*, *Trinervitermes* sp.1, and *C. secundus*. DNA extractions were carried out using the cetyltrimethylammonium bromide (CTAB) protocol (Fuchs et al. 2003) with minor modifications, such as using cut pipette tips to obtain high molecular weight DNA and digesting proteins with over-night proteinase K incubation to increase DNA yield. DNA samples were tested with a NanoDrop.

Depending on the amount of extracted DNA material, low DNA input or standard DNA input libraries (Table 1) were subsequently constructed for single-molecule real-time (SMRT) sequencing to generate long high-fidelity (HiFi) reads with the Pacific Biosciences (PacBio) Sequel II system at the West German Genome Sequencing Center of HHU Düsseldorf.

Genome assembly

Genomes of the four termite species were assembled with *Hifiasm* (v.0.19.5) (Cheng et al. 2021), a *de novo* assembler for PacBio HiFi reads. We have tested multiple genome assembly methods and found that Hifiasm had the best performance: genomes assembled by Hifiasm had BUSCO genome completeness between 97.6% and 99.5% and had genome sizes of

around 1.3 Gb, similar to previously published termite genomes (Poulsen et al. 2014, Harrison et al. 2018) (Table 1). We noticed that the *C. secundus* and the *Trinervitermes* sp.1 genome assemblies both had > 10% duplicated BUSCOs, significantly higher than the *M. bellicosus* and the *Odontotermes* sp.2 genome assemblies (3.1% and 1.8% duplicated BUSCOs, respectively) (Table 1). We reason that the higher BUSCO duplication rates may be caused by the lower sequencing depth in *C. secundus* and *Trinervitermes* sp.1 because a lower sequencing depth can lead to false duplications in diploid genomes with high genomic heterozygosity (Ko et al. 2022). To remove false duplications, we employed a second round of haplotig-duplication purging on all the genome assemblies with *purge_dups* (v. 1.2.6) and with the *-e* option to only purge haplotypic duplications at the ends of the contigs (Guan et al. 2020). The purged genomes were then screened and contamination removed with FCS-GX (Astashyn A et al. 2023).

Genome annotation

Because *Odontotermes* sp.2, *M. bellicosus* and *C. secundus* have available RNAseq data (published or in-house: Elsner et al. 2018, Harrison et al. 2018, Monroy Kuhn et al. 2019, Lin et al. 2021, Rau et al. 2023), genome assemblies of these three species were annotated with the Braker3 (v. 3.0.3) pipeline (Gabriel et al. 2023). Genome assemblies were first soft-masked with RepeatMasker (Smit et al 2013-2015) (see *Repeat and transposable element annotation*). RNAseq data were then mapped to the target genome with *Hisat2* (v. 2.2.1) (Kim et al. 2019) with the *--dta* option to report alignments tailored for transcript assemblers. The masked genome was then annotated with Braker3, using both RNAseq data (from the Hisat2 mapped reads) and protein data (Arthropoda orthologs from OrthoDB 11, Kuznetsov et al. 2023). For *M. bellicosus*, the queen genome assembly was used as the representative genome for genome annotation because the genome assembly from the queen sample had a better quality than that of the soldier sample (based on N50 and BUSCO).

For *Trinervitermes* sp.1, which lacks RNAseq data, the genome assembly was annotated with GeMoMa (Keilwagen et al. 2019), a homology-based gene prediction method. Because *Trinervitermes* sp.1 is a Termitidae, we used the annotated *M. bellicosus* queen genome assembly as reference.

Repeat and transposable element annotation

A *de novo* Blattodea repeat database was built with *RepeatModeler2* (v. 2.0.4) (Flynn et al. 2020), a *de novo* TE family identification tool that integrates the prediction results of RECON, RepeatScout and LtrHarvest/Ltr_retriever. For the repeat database construction, we used the published short-read genome assembly of *B. germanica* (Harrison et al. 2018) and our newly assembled genomes of *C. secundus* and *M. bellicosus* (queen) as input. The database therefore contains non-redundant consensus repeat sequences from the full spectrum of social complexities in Blattodea. We next classified these repeat sequences with *RepeatModeler2* and the Dfam TE family database (v.3.7) (Storer et al. 2021). With the classified Blattodea repeat database, we used *RepeatMasker* (v. 4.1.5) to annotate the repeat (including TE) families in each genome assembly

With the classified Blattodea repeat database, we used *RepeatMasker* (v. 4.1.5) to annotate the repeat (including TE) families in each genome assembly, including *B. germanica* and the new genome assemblies of the four termite species. This annotation procedure provided us with homologous repeats that were comparable across our genome assemblies.

For classifying TEs into super-families/families we applied the classification recommended in Piégu et al. (2015).

Quantification of CpG DNA methylation levels

PacBio HiFi reads contain kinetic information that reflects CpG 5-Methylcytosine (5mC) modifications (Flusberg et al. 2010). We thus used *Jasmine* (v. 2.0.0) to predict the CpG 5mC for our HiFi read sequences. The HiFi reads with annotated 5mC information were then mapped to the corresponding genome assembly with *minimap2 (pbmm2,* v. 1.12.0*)* (Li 2018). To filter suspicious read alignments, we further used *samclip* to remove aligned reads with soft-clipped. With the aligned reads, we calculated the CpG-site methylation probabilities for each genome assembly using PacBio's *pb-CpG-tools* (*aligned_bam_to_cpg_scores* with the *model* pileup mode, v. 2.3.2), a machine-learning-based method for genomic CpG 5mC quantification.

For a robust determination of genomic CpG 5mC methylation, we only retained CpG sites with a read coverage between 10 to 50x, a CpG site was considered methylated when its CpG-site methylation probability was over 50%.

To quantify the CpG methylation level in different genomic regions and different repeat categories/super-families/families, we used the R package GenomicRanges (Lawrence et al. 2013) to determine the total number of CpG sites in target genomic regions/repeats. The percentage of methylated CpGs was then calculated as the number of methylated CpG sites divided by the total number of CpG sites.

The effect size of CpG methylation level was calculated as the log2-transformed-ratio between the percentage of methylated CpGs in target genomic regions/repeats and the percentage of methylated CpGs in the genome. Therefore, a positive value means a higher CpG methylation level than the genomic background and a negative value means a lower one.

Clustering analyses

For hierarchical clustering of repeat abundance across genome assemblies, repeat abundances were summarized at the family level (homologous repeats generated from RepeatModeler, see above). The family level repeat abundance was first $\log_{10}$ transformed (with one added pseudo count to avoid $\log_{10} 0$). The log transformed repeat abundance was then clustered with the R package pheatmap (Kolde 2019).

For hierarchical clustering of CpG methylation level across genome assemblies, the percentages of methylated CpGs (at the super-family level) were first standardized within each genome assembly. The standardization within a genome assembly is necessary because the sequencing depth differences across genome assemblies can influence the confidence of

modelling CpG methylation probability. The standardized CpG methylation levels were then clustered with the R package pheatmap (Kolde 2019).

Examination of the correlation between TE abundance and the percentage of methylated CpGs

The correlation was examined in R (R Core Team 2023) with a generalized linear model:

*Percentage of methylated CpG ~ log10(TE abundance + 1) * TE category + Species identity*

To further identify the correlations in different TE super-families, we used the following generalized linear model:

*Percentage of methylated CpG ~ log10(TE abundance + 1) * TE super-family + Species identity*

For both models, TE abundance was quantified at the family level.

Acknowledgement

We thank the West German Genome Sequencing Center at the University of Düsseldorf for support, N'golo Koné for logistic support and David Kouadio Kouamé for assistance in the field in Côte d'Ivoire. Charles Darwin University, and especially S. Garnett and the Horticulture and Aquaculture team, provided logistic support to collect the *C. secundus* in Australia. The Office Ivoirien des Parcs et Réserves provided sampling and export permits for the termites from Côte d'Ivoire. The Parks and Wildlife Commission, Northern Territory, the Department of the Environment, Water, Heritage and the Arts gave permission to collect and export *C. secundus*. This study was conducted according to the Nagoya protocol. This research was supported by the state of Baden-Württemberg through Baden-Württemberg High Performance Computing and the Deutsche Forschungsgemeinschaft (DFG) by a grant with the Sequencing Cost Initiative to JK (KO1895/26-1: 417980976).

Data availability

Sequence data are deposited at the National Center for Biotechnology Information (NCBI) (BioProject ID PRJNA1033592) and can be accessed via a reviewer link: http://www.ncbi.nlm.nih.gov/bioproject/1033592

Authors' contribution

BG designed the study, performed bioinformatic analyses and drafted the manuscript, DE prepared DNA samples and performed bioinformatic analyses, JK designed the study, obtained funding, sampled the specimens and drafted the manuscript. All authors read and approved the submission of this manuscript.

References

Astashyn A et al. (2023) Rapid and sensitive detection of genome contamination at scale with FCS-GX. bioRxiv. doi: https://doi.org/10.1101/2023.06.02.543519

Bignell DE, Eggleton P (1997) Termites in ecosystems. In: Termites: Evolution, Sociality, Symbioses, Ecology (Eds. Abe T, Bignell DE, Higashi M) Kluwer Academic Publishers, Dordrecht, 363-388.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18: 170-175.

De Cecco M et al. (2013) Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. Aging 5:867-883.

Deniz O, Frost JM & Branco MR (2019) Regulation of transposable elements by DNA modifications. Nat Rev Genet 20:417–431.

Elsner D, Meusemann K, Korb J (2018) Longevity and transposon defense, the case of termite reproductives. Proc Natl Acad Sci USA 115:5504-5509.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods 7:461-465.

Flynn JM, Hubley R, Goubert C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117: 9451-9457. https://doi.org/10.1073/pnas.1921046117

Fuchs A, Heinze J, Reber-Funk C, Korb J (2003) Isolation and characterisation of six microsatellite loci in the drywood termite *Cryptotermes secundus* (Kalotermitidae). Molecular Ecology Notes 3:355-357

Gabriel L, Brůna T, Hoff JK, Ebel M, Lomsadze A, Borodovsky M, Stanke M (2023) BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGSTUS and TSEBRA. bioRxiv doi: https://doi.org/10.1101/2023.06.10.544449

Gadagkar R (1997) The evolution of caste polymorphism in social insects: genetic release followed by diversifying evolution. J Genetics 76:167-179

Glastad KM, Gokhale K, Liebig J, Goodisman MAD (2016) The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. Sci Rep 6:1–14.

González J, Petrov DA (2012) Evolution of genome content: population dynamics of transposable elements in flies and humans. In: Anisimova M (ed) Evolutionary Genomics: Statistical and Computational Methods 1. Springer, Berlin

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R (2020) Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896-2898. doi: 10.1093/bioinformatics/btaa025

Harris KD, Lloyd JPB, Domb K, Zilberman D, Zemach A (2019) DNA methylation is maintained with high fidelity in the honey bee germline and exhibits global non-functional fluctuations during somatic development. Epigenetic & Chromatin 12:62

Harrison MC et al. (2018) Hemimetabolous genomes reveal molecular basis of termite eusociality. Nature Eco Evo 2:557-566

Harrison MC, Dohmen E, George S, Sillam-Dussès D, Séité S, Vasseur-Cognet M (2022) Complex regulatory role of DNA methylation in caste- and age-specific expression of a termite. Open Biol 12:220047.

Heard E, Martienssen RA (2014) Transgnerational epigenetic inheritance: myths and mechanisms. Cell 157:95-109. http://dx.doi.org/10.1016/j.cell.2014.02.045

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97-108

Inward D, Beccaloni G, Eggleton P (2007) Death of an order: a comprehensive molecular phylogenetic study confirms that termites are eusocial cockroaches. Biol Lett 3:331–335.doi:10.1098/rsbl.2007.0102

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. J Asia-Pacific Ento 23:666-674

Keilwagen J, Hartung F, Grau J (2019) GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. Methods Molecular Biology 1962:161-177. doi: 10.1007/978-1-4939-9173-0_9.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37:907-915.

Ko BJ, Le C, Kim J, Rhie A, Yoo DA, Howe K, Wood J, Cho S, Brown S, Formenti G, Jarvis ED, Kim H (2022) Widespread false gene gains caused by duplication errors in genome assemblies. Genome Biology 23:205

Kolde R (2019) *pheatmap: Pretty Heatmaps.* R package version 1.0.12, https://CRAN.R-project.org/package=pheatmap.

Korb J (2016) Genes underlying reproductive division of labor in termites with comparisons to social Hymenoptera. Front Ecol Evol 4:e45

Korb J, Heinze J (2016) Major hurdles for the evolution of sociality. Annual Reviews of Entomology 61:297-316

Korb J, Poulsen M, Hu H, Li C, Boomsma JJ Zhang G, Liebig J (2015) A genomic comparison of two termites with different social complexity. Front Genet 6:e9

Koshikawa S, Miyazaki S, Cornette R, Matsumoto T, Miura T (2008). Genome size of termites (Insecta, Dictyoptera, Isoptera) and woodroaches (Insecta, Dictyoptera, Cryptocercidae). Naturwissenschaften 95:859–867.doi: 10.1007/s00114-008-0395-7

Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM (2023) Ortho DB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Research 51: D445-D451

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013) Software for computing and annotating genomic ranges. PLoS Comput Biol 9: e1003118. doi:10.1371/journal.pcbi.1003118

Lewis SH, Ross L, Bain SA, Pahita E, Smith SA, Cordaux R, et al. (2020) Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. PLoS Genet 16: e1008864.

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100. https://doi.org/10.1093/bioinformatics/bty191

Lin S, Werle J, Korb J (2021) Transcriptomic analyses of the termite, *Cryptotermes secundus*, reveal a gene network underlying a long lifespan and high fecundity Communication Biology 4:384. DOI:10.1038/s42003-021-01892-x.

Martelossi J, Forni G, Iannello M, Savojardo C, Martelli PL, Casadio R, Mantovani B, Luchetti A, Rota-Stabelli (2023) Wood feeding and social living: draft genome of the subterranean termite *Reticulitermes lucifugus* (Blattodea; Termitoidae). Insect Mol Biol 32:118-131. https://doi.org/10.1111/imb.12818 32, Issue2

Monroy Kuhn JM, Meusemann K, Korb J (2019) Long live the queen, the king and the commoner? Transcript expression differences between old and young in the termite *Cryptotermes secundus*. PLoS One 14:e0210371

Nobré T, Rouland-Lefèvre C, Aanen DK (2011) Comparative biology of fungus cultivation in termites and ants. In Biology of termites: a modern synthesis (eds DE Bignell, Y Roisin, N Lo), pp.193–210. Dordrecht, The Netherlands: Springer.

Okwaro LA, Korb J (2023) Epigenetic regulation and division of labor in social insects. Curr Opin Insect Sci 58:1010051

Piégu B, Bire S, Arensburger P, Bigo Y (2015) A survey of transposable element classification systems – a call for a fundamental update to meet the challenge of their diversity and complexity. Molecular Phylogenetics and Evolution 86:90–109.

Post F, Bornberg-Bauer E, Vasseur-Cognet M, Harrison MC (2023) More effective transposon regulation in fertile, long-lived termite queens than in sterile workers. Mol Ecol 32:369-380.

Poulsen et al. (2014) Holobiomic division of labor in fungus-farming termites Proc Natl Acad Sci USA 111:14500-14505.

Pringle RM, Doak DF, Brody AK, Jocqué R, Palmer TM (2010) Spatial pattern enhances ecosystem functioning in an African savannah. PLoS Biology 8:e1000377

Rau V, Flatt T, Korb J (2023) The remoulding of dietary effects on the fecundity / longevity trade-off in a social insect. BMC Genomics 24:244

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shigenobu S, et al. (2022) Genomic and transcriptomic analyses of the subterranean termite *Reticulitermes speratus*: gene duplication facilitates social evolution. Proc Natl Acad Sci USA119: e2110361119

Smit AFA, Hubley R, Green P (2013-2015) *RepeatMasker Open-4.0*. <http://www.repeatmasker.org>.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12:2

Sturm A, Ivics Z, Vellai T (2015) The mechanism of ageing: primary role of transposable elements in genome disintegration. Cell Mol Life Sci 72:1839-1847.

Su N-Y (2019) Development of baits for population management of subterranean termites. Annu Rev Entomol 64:115-130.

Terrapon N et al. (2014) Molecular traces of alternative social organization in a termite genome. Nature Comm 5:3636

Wood JG, Helfand SL (2013) Chromatin structure and transposable elements in organismal aging. Front Genet 4:e274.

Wood JG et al. (2016) Chromatin-modifying genetic interventions suppress age-associated transposable element activation and extend life span in *Drosophila*. Proc Natl Acad Sci USA 113:11277-11282.

Yagound B, Remnant EJ, Buchmann G, Oldroyd BP (2020) Intergenerational transfer of DNA methylation marks in the honey bee. Proc Natl Acad Sci USA 117: 32519-32527. https://doi.org/10.1073/pnas.201709411

Zhang (2003) Evolution by gene duplication: an update. Trend Ecol Evol18:292-298

Zhou W, Liang G, Molloy PL, Jones PA (2020) DNA methylation enables transposable element-driven genome expansion Proc Natl Acad Sci USA 117:19359-19366. https://doi.org/10.1073/pnas.192171911