Report

Current Biology

Robust termite phylogenies built using transposable element composition and insertion events

Highlights

- Trees inferred from transposable element composition surpass mitogenome-based trees
- Trees based on TE insertion rivaled those from thousands of ortholog alignments
- Transposable elements represent an untapped source of phylogenetic markers

Authors

Cong Liu, Simon Hellemans, Yi-Ming Weng, ..., Mark C. Harrison, Dino P. McMahon, Thomas Bourguignon

Correspondence

congliu37@outlook.com (C.L.), thomas.bourguignon@oist.jp (T.B.)

In brief

Phylogenetic trees are generally reconstructed from conserved sequence alignments. Liu et al. use transposable elements to build termite phylogenetic trees as accurate as those inferred from over one thousand single-copy orthologous gene alignments, showing they represent robust markers for phylogenetic reconstructions.

Liu et al., 2025, Current Biology *35*, 1–7
November 17, 2025
https://doi.org/10.1016/j.cub.2025.10.019

CellPress

Current Biology

CellPress

Report

Robust termite phylogenies built using transposable element composition and insertion events

Cong Liu,[1,*] Simon Hellemans,[1] Yi-Ming Weng,[1] Alina A. Mikhailova,[2,4] Cédric Aumont,[3,4] Aleš Buček,[1,5] Jan Šobotník,[5,6] Mark C. Harrison,[7] Dino P. McMahon,[3,4] and Thomas Bourguignon[1,5,8,*]

[1]Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son 904-0495, Okinawa, Japan
[2]Institute for Evolution and Biodiversity, University of Münster, Hüfferstraße 1, 48149 Münster, Germany
[3]Institute of Biology, Freie Universität Berlin, 14195 Berlin, Germany
[4]Department for Materials and the Environment, BAM Federal Institute for Materials Research and Testing, 12205 Berlin, Germany
[5]Institute of Entomology, Biology Centre, Czech Academy of Sciences, České Budějovice 37001, Czech Republic
[6]Faculty of Tropical AgriSciences, Czech University of Life Sciences, Kamýcká 129, 16521 Prague, Czech Republic
[7]Centre for Health & Life Sciences, Coventry University, CV1 5FB Coventry, UK
[8]Lead contact
*Correspondence: congliu37@outlook.com (C.L.), thomas.bourguignon@oist.jp (T.B.)
https://doi.org/10.1016/j.cub.2025.10.019

SUMMARY

Phylogenetic trees are typically reconstructed using conserved sequence alignments.[1,2] Other genomic elements, such as transposable elements (TEs), make up a large fraction of eukaryotic genomes[3] but are ignored for phylogenetic reconstruction, despite potentially containing phylogenetic information,[4,5] which could be used to resolve nodes that remain contentious. Here, we reconstructed accurate phylogenetic trees of 45 termites and two cockroaches using two types of characters derived from the TE landscape: (1) genome-wide presence and absence of 37,966 TE families and (2) presence/absence data of 37,966 TE families in the flanking regions of orthologous ultraconserved elements (UCEs), which was a proxy for TE insertions. The topologies of our TE-based phylogenetic trees were largely congruent with phylogenetic trees inferred from alignments of UCEs and single-copy orthologous genes, only differing for a few nodes variably reconstructed in other phylogenetic analyses. Notably, trees based on genome-wide TE family composition were more accurate than trees inferred from mitochondrial genome alignments, and trees based on TE family composition in regions flanking UCEs achieved comparable accuracy with trees inferred from single-copy orthologous gene alignments. Our results demonstrate that the TE landscape is phylogenetically informative, representing an additional set of markers for robust phylogenetic reconstructions, with potential use to resolve ambiguous nodes in the tree of life.

RESULTS AND DISCUSSION

Our understanding of the tree of life is largely based on phylogenetic trees reconstructed from alignments of conserved genetic sequences, often derived from protein-coding genes.[1,2] Modern phylogenetic trees tend to be robust, as they are reconstructed from alignments of thousands of sequences, allowing for the resolution of the relationships among many, but not all, lineages. Yet some specific nodes remain difficult to resolve due to underlying evolutionary processes, such as rapid speciation events leading to incomplete lineage sorting[6] and the loss of phylogenetic signal for deep divergences.[7] New characters other than substitutions inferred from sequence alignments may be used to elucidate ambiguous nodes in the tree of life. For example, the early branching in the animal tree was inferred using conserved gene synteny.[8] The identification of new molecular markers is one avenue to further improve future phylogenetic reconstructions.

Genes typically make up a small fraction of eukaryotic genomes, unlike transposable elements (TEs), which often represent more than half of eukaryotic genomes.[3] TEs are acquired vertically through parental inheritance or through horizontal transfers, often across distantly related organisms.[9] Although the rate of horizontal TE transfers between species has previously been qualified as "massive" for insects,[10] TE landscapes do contain phylogenetic information.[9] For example, TE insertions in orthologous genomic regions have been used as phylogenetic characters in a few lineages of vertebrates.[4,5] Because the probability of insertion of closely related TEs in the same location is negligible, TE insertions are homoplasy-free characters, theoretically ideal for phylogenetic reconstruction.[11,12] Furthermore, most TE insertions are nearly neutral,[13] and TE activity can introduce numerous genetic variations in a short period of time,[14,15] which may allow the resolution of divergences that arose during rapid speciation events.

While the use of TEs for phylogenetic reconstruction experienced initial success, it has largely been abandoned, possibly because the characterization of TE insertion events was originally arduous, each requiring efforts similar to the production

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

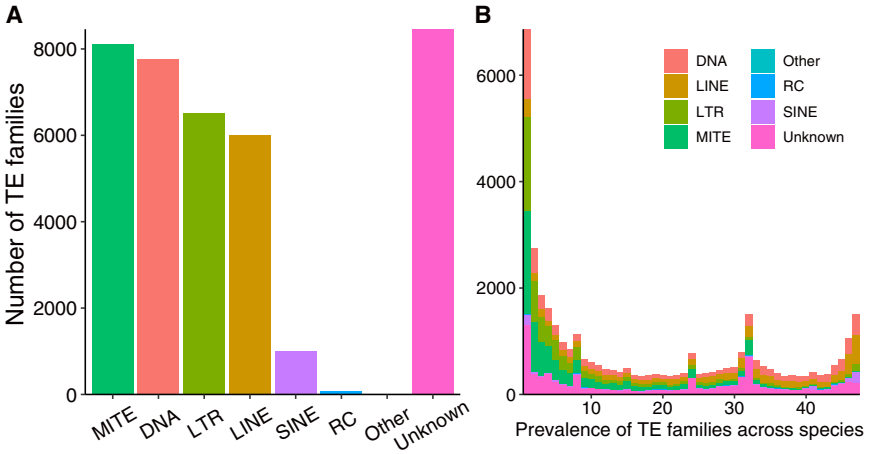

Figure 1. TE family diversity across the genomes of 45 termites and two cockroach outgroups

(A) Number of TE families assigned to each of the main TE classes in the pan-genome of 47 species. (B) Prevalence of TE families among the genomes of the 47 species studied here, with each bar representing the number of TE families found in a given number of species (the sum of all the bars is 37,966, the number of TE families identified in this study).

of a DNA sequence containing hundreds of characters. However, genome sequencing and TE annotation have become less challenging tasks due to improved sequencing methods and genome assembly and annotation tools. Eukaryotic genomes often contain millions of TEs, each potentially representing independent characters for phylogenetic analyses. TE insertions inferred from genomic data have occasionally been used to build phylogenetic trees of closely related species,[16] but it remains unclear whether the TE landscape can be used to reconstruct accurate phylogenetic trees at a larger timescale.

Here, we used whole-genome assemblies and improved TE annotation methods to reconstruct accurate phylogenetic trees of termites based on TEs. We used two types of characters derived from the TE landscape: (1) genome-wide presence/absence data of TE families and (2) presence/absence data of TE families in the flanking regions of orthologous ultraconserved elements (UCEs), a proxy for TE insertions. Termites represent an ideal use-case lineage. First, termite phylogenies have previously been reconstructed using mitochondrial genomes,[17,18] transcriptomes,[19] and UCEs,[20,21] leaving only a few nodes unresolved and thereby providing a solid basis for comparison. Second, whole-genome assemblies of 45 termite species representing 11 of the 13 families of termites and 12 of the 18 subfamilies of Termitidae were recently generated by Liu et al.,[22] allowing a genome-level characterization of TEs for an insect lineage that originated ~150 million years ago.[17,19]

Phylogenetic trees inferred from TE family composition are more accurate than trees inferred from mitochondrial genome sequence alignments

We first characterized the TE landscapes of the 45 termites and 2 cockroach outgroups (*Blatta orientalis* and *Cryptocercus meridianus*). We built one TE library for each genome individually using the *de novo* TE annotation pipeline EDTA[23] and aggregated all libraries into a pan-genome TE library using the software panEDTA.[24] Our pan-genome TE library contained a total of 37,966 sequences, each representing a TE family following the 80-80-80 criteria[25]—two sequences belonged to the same TE family if over 80% of one sequence aligned to the other sequence with identity and length above 80% and 80 bp, respectively. We used our pan-genome TE library to annotate all genome assemblies and retained both intact and fragmented TEs. The main TE classes included miniature inverted-repeat TEs (MITEs), DNA transposons, long terminal repeats (LTRs), and long and short interspersed nuclear elements (LINEs and SINEs) (Figure 1A). Many TE families were endemic to specific termite clades. For example, many TE families were specific to Termitidae and Neoisoptera, forming 2 peaks in the distribution of TE family prevalence among termites at 24 and 32 species, the number of species of Termitidae and Neoisoptera used in this study (Figure 1B). These results show that the TE family composition of termite genomes contains a phylogenetic signal supporting monophyly of major termite clades.

Our first approach to inferring TE-based phylogenetic trees relied upon the composition of TE families across the genomes of 47 blattodean genomes. We inferred two maximum likelihood trees from binary matrices that have the presence/absence of TE families coded as 1/0. One tree was inferred using a substitution model for binary data (TEcB), and the other tree was based on a substitution model for morphological data (TEcM). Both trees were largely congruent with the UCE-based phylogenetic tree of Liu et al.,[22] with a few exceptions (Figures 2A, S1H, and S1I). In TEcM, (1) Hodotermopsidae was found sister to Stolotermitidae, instead of sister to Archotermopsidae in the UCE-based phylogeny; (2) *Paraneotermes* was found sister to other Kalotermitidae, instead of *Kalotermes* in the UCE-based phylogeny and previous mitogenome-based phylogenetic trees[26]; (3) *Dolichorhinotermes* was sister to *Prorhinotermes* + Heterotermitidae + Termitidae, instead of forming a monophyletic group with *Glossotermes* in the UCE-based phylogeny; (4) Sphaerotermitinae was sister to all other Termitidae subfamilies, while it represented the sister group of Macrotermitinae in UCE- and transcriptome-based phylogenies[19–21] and the sister group of non-Macrotermitinae non-Foraminitermitinae Termitidae in mitogenome-based phylogenies[17,18]; and (5) the relationships among subfamilies of Termitidae forming a clade sister to Apicotermitinae, which were largely unresolved in previous sequence-based phylogenetic trees,[18,19,21] presented a unique branching pattern, while the relationships among genera of Nasutitermitinae and Syntermitinae + Microcerotermitinae were congruent with the UCE-based tree. The topology of TEcB was similar to that of TEcM, except for the relationships among genera of Kalotermitidae, which differed from the UCE-based phylogeny and previous mitogenome-based phylogenetic trees.[26] Notably, incongruencies between TE-family-content-based and sequence-based trees primarily affected nodes variably reconstructed in previous phylogenetic analyses, while nodes that

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

CellPress

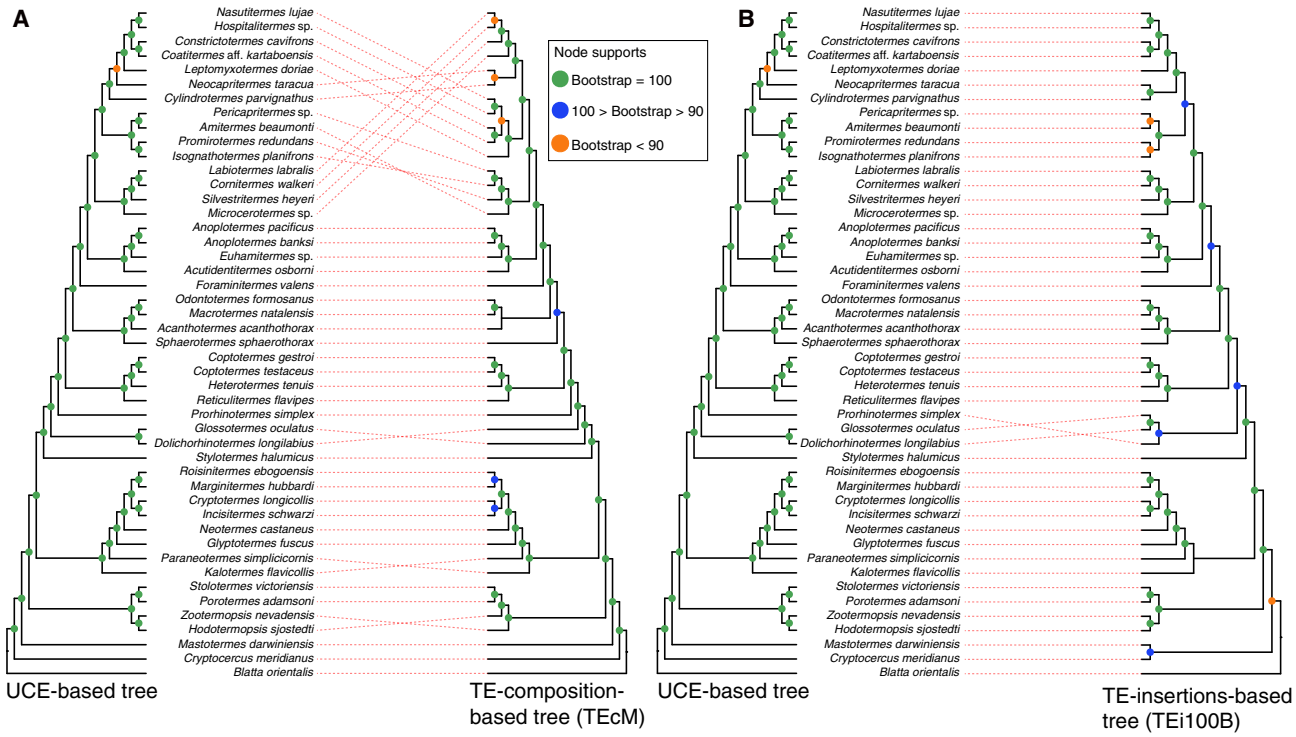

Figure 2. Comparisons between sequence-based and TE-based phylogenetic trees

Tanglegrams between a phylogenetic tree (left) based on the alignment of 27,610 ultraconserved elements (UCEs)[22] with (A) a tree inferred from presence/absence data of 37,966 TE families reconstructed using a maximum likelihood model for morphological data and (B) a tree reconstructed using a maximum likelihood model for binary data and inferred from TE insertion information, using presence/absence data of 37,966 TE families in the 100-bp region flanking 13,227 UCEs as a proxy.

See also Figure S1 and Table S1.

received strong support were generally retrieved with identical topology.

To better characterize the accuracy of the two trees inferred from TE family content, we reconstructed six phylogenetic trees for comparison, including three trees inferred from 1,410 nuclear single-copy orthologous genes (scHOGs) and three trees inferred from the 13 mitochondrial protein-coding genes. The six trees were reconstructed using protein sequence alignments and DNA sequence alignments analyzed with and without third codon positions (Figures S1B–S1G; Table S1). We calculated the normalized Robinson-Foulds distance[27] (nRF) between each of these eight trees and a reference tree based on 27,610 UCEs reconstructed by Liu et al.,[22] which arguably is the closest approximation of the actual species tree for this 47-species dataset (Figure 3). The resulting UCE-nRF metric reflects the disimilarity of each tree with the reference UCE tree, with values of 0 indicating identical tree topologies. TEcM and TEcB had a UCE-nRF of 0.205 and 0.273, respectively. These values were higher than the UCE-nRF values obtained for trees inferred from 1,410 scHOGs (0.023 for the protein-based tree, 0.023 for the tree based on nucleotide sequence with third codon positions included, and 0.045 for the nucleotide-sequence-based tree without third positions). However, the UCE-nRF values obtained for the two trees inferred from TE family content were lower than the UCE-nRF values obtained with three trees inferred from the 13 mitochondrial protein-coding genes (0.295 for the protein-sequence-based tree and 0.341 and 0.319 for the DNA-sequence-based tree with and without the third position, respectively). This may be explained by the nuclear origin of both UCEs and the TE family content, while mitochondrial genes are linked and represent a single marker often experiencing introgression, which leads to discordance between mitogenome trees and species trees.[28,29] Overall, these results support the genomic content in TE families as valid characters for phylogenetic tree reconstruction, yielding trees with higher accuracy than trees inferred from mitochondrial genomes (Figure 3).

Phylogenetic trees inferred from TE family composition in the flanking regions of orthologous UCEs, a proxy for TE insertions, have comparable accuracy to trees inferred from thousands of nuclear sequence alignments

Previous studies used TE insertion events instead of the TE family composition to extract phylogenetic information from the TE landscape.[4,5] We designed a new approach to identify a large number of orthologous TE insertions across the 47 blattodean genome set. We used presence/absence data of TE families in UCE flanking regions as a set of characters for phylogenetic tree reconstruction. Our approach is based on the premise that TEs belonging to the same family and located near orthologous genomic elements, such as UCEs, are themselves orthologous. We inferred 16 maximum likelihood trees using

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

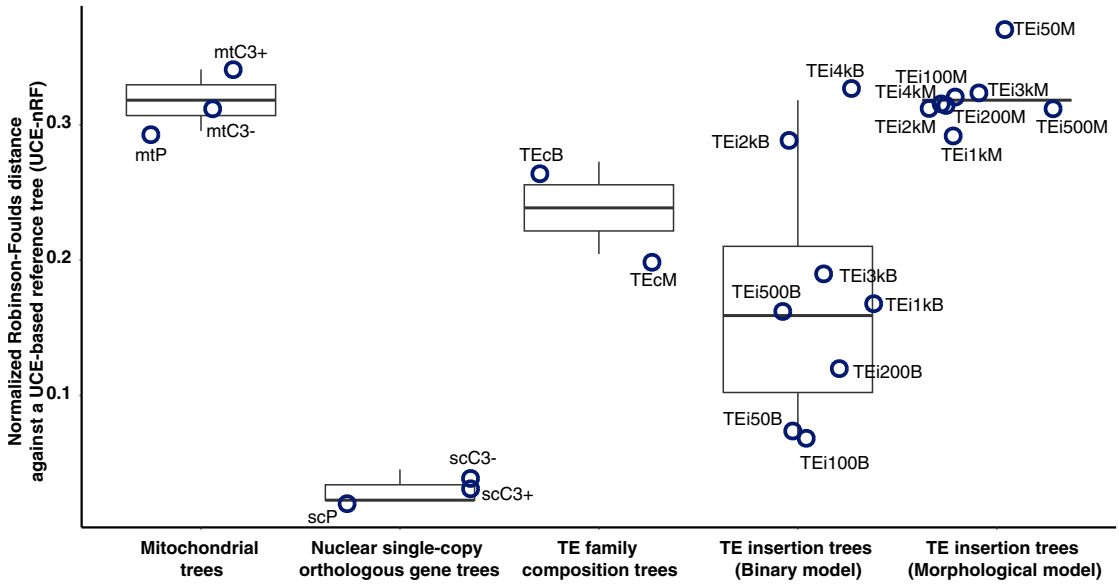

Figure 3. Comparisons between a reference phylogenetic tree based on the alignment of 27,610 UCEs and 24 trees reconstructed in this study

The values on the *y* axis are the normalized Robinson-Foulds distances between the reference phylogenetic tree[22] and 24 trees reconstructed in this study (UCE-nRF). The 24 trees included three trees based on 13 mitochondrial protein-coding genes and inferred from protein alignments (mtP) and nucleotide alignments with (mtC3+) and without (mtC3−) third codon positions; three trees based on 1,410 nuclear scHOGs and inferred from protein alignments (scP) and nucleotide alignments with (scC3+) and without (scC3−) third codon positions; two trees inferred from presence/absence data of 37,966 TE families using models for binary (TEcB) and morphological data (TEcM); eight trees inferred from presence/absence data of 37,966 TE families in a UCE flanking region of 50 bp (TEi50B), 100 bp (TEi100B), 200 bp (TEi200B), 500 bp (TEi500B), 1,000 bp (TEi1kB), 2,000 bp (TEi2kB), 3,000 bp (TEi3kB), and 4,000 bp (TEi4kB) using a model for binary data; and eight trees inferred from presence/absence data of 37,966 TE families in a UCE flanking region of 50 bp (TEi50M), 100 bp (TEi100M), 200 bp (TEi200M), 500 bp (TEi500M), 1,000 bp (TEi1kM), 2,000 bp (TEi2kM), 3,000 bp (TEi3kM), and 4,000 bp (TEi4kM) using a model for morphological data.
See also Figure S2.

presence/absence data of 37,966 TE families in UCE flanking regions. We used two models for binary and morphological data and eight UCE flanking region lengths: 50, 100, 200, 500, 1,000, 2,000, 3,000, and 4,000 bp (Figures S1J–S1Y; Table S1). As a first step to evaluate the 16 trees, we calculated the UCE-nRF metric for each tree. The UCE-nRF values obtained for trees inferred with the model for morphological data varied between 0.295 and 0.364, which are within the range of values obtained for the trees inferred from the 13 mitochondrial protein-coding genes (Figure 3). The UCE-nRF values of the trees inferred with the model for binary data were variable, ranging from 0.068 to 0.318, with values decreasing as the UCE flanking region length decreases, possibly reflecting the effect of recombination, which invalidates the assumption of orthology among TEs located farther away from orthologous UCE loci. Notably, the two trees reconstructed with a UCE flanking region length of 50 bp (TEi50B) and 100 bp (TEi100B) and the binary model were topologically identical and had a UCE-nRF value of 0.068, which is close to the values obtained with trees based on 1,410 nuclear scHOGs (Figure 3). Therefore, TE insertion events in the neighborhood of UCEs can be used to produce phylogenetic trees comparable to state-of-the-art phylogenies reconstructed from thousands of nuclear loci.

The topology of TEi50B and TEi100B was similar to the UCE-based phylogenetic tree of Liu et al.,[22] with three exceptions (Figure 2B): (1) *Cryptocercus* and *Mastotermes* formed a monophyletic group sister to other termites, while *Cryptocercus* was consistently sister to *Mastotermes* + other termites in previous phylogenetic reconstructions[17,19,21,30–33]; (2) *Prorhinotermes* was sister to *Dolichorhinotermes* + *Glossotermes* instead of sister to Heterotermitidae + Termitidae in the UCE-based phylogeny; and (3) *Cylindrotermes* was sister to *Neocapritermes* instead of sister to *Neocapritermes* + Nasutitermitinae in the UCE-based phylogeny. The latter two incongruencies affected species that lie in unresolved parts of the termite tree; however, the former incongruency requires an explanation. *Cryptocercus* and *Mastotermes* were reconstructed as monophyletic with bootstrap supports lower than 100%, unlike most of the branches in TEi50B and TEi100B, pointing toward potentially incorrectly inferred topology (Figures 2B and S1J). A sister relationship between *Cryptocercus* and *Mastotermes* is not parsimonious, as it would require two independent origins of eusociality or one origin and one loss. Our approach was therefore unable to resolve nodes older than ~120 million years, possibly because of saturation of the phylogenetic signal at this time scale. At a shorter time scale, our TE-insertion-based phylogenetic reconstruction approach is comparable to the currently most robust and data-intensive reconstructions based on UCE- and transcriptome-based phylogenies (Figure 3).

Integrating TEs with sequence-based phylogenies

Our results show that TEs can be used as phylogenetic characters alone, which could bring new insights into the topology of nodes ambiguously reconstructed by previous analyses. They

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

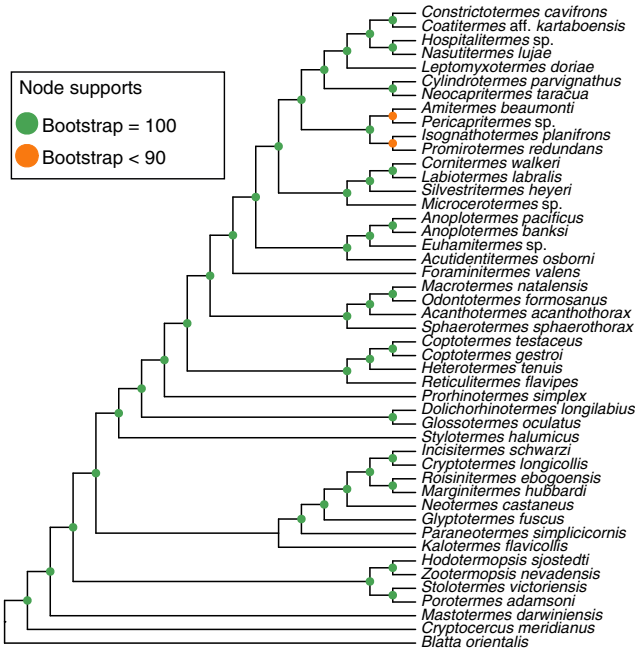

Figure 4. Reconstruction of a total-evidence phylogenetic tree

Maximum likelihood phylogenetic tree of termites based on a total-evidence approach combining TE family content, aligned scHOG DNA sequences with the third codon position included, 13 mitochondrial protein alignments, and TE family content in the 100-bp regions flanking UCEs.

See also Table S1.

could also be used in combination with sequence alignments, rather than as an alternative, as a total-evidence approach similar to phylogenetic studies that combine morphological and molecular data, as has been done for termites.[32,34,35] We reconstructed a maximum likelihood phylogenetic tree that followed this philosophy and was based on four types of data partitioned into four subsets: TE family content, aligned scHOG DNA sequences with third codon position included, 13 mitochondrial protein alignments, and TE family content in the 100-bp regions flanking UCEs. The topology of our combined tree was identical to the UCE-based phylogenetic tree of Liu et al.,[22] except for the position of *Neocapritermes*, which was sister to *Cylindrotermes* in the combined tree, instead of sister to Nasutitermitinae in the UCE-based tree (Figure 4). The phylogenetic position of *Neocapritermes* remains partially unresolved, as it branches in a polytomic part of the tree with short internodes characteristic of rapid diversification events.[19,21] The sister position of the South American *Neocapritermes* and *Cylindrotermes* inferred from the total-evidence approach represents a new phylogenetic hypothesis that invites future testing using morphological and other molecular data. In summary, our results show that TEs represent an additional source of phylogenetic characters, which can be used to supplement sequence alignments and generate alternative hypotheses for lineages that have been resisting phylogenetic resolution.

Conclusions

Alignments of conserved genetic sequences are the primary source of characters for phylogenetic reconstruction.[1,2] Other molecular characters have been used occasionally, but they have been difficult to characterize and generally ignored. Our results show that TE landscapes provide an additional source of molecular characters, which can be used to reconstruct robust phylogenies as an alternative to, or in combination with, sequence alignments. Our approach takes advantage of the improvements in sequencing methods, which have led to the generation of many genome assemblies of sufficient quality for a thorough characterization of TE landscapes. While our TE-based phylogenies are comparable in accuracy to state-of-the-art phylogenies reconstructed from thousands of nuclear sequence alignments, they may be further improved in several ways by future studies, such as by the development of evolutionary models better adapted to TE evolution. Overall, our study shows that eukaryotic genomes contain useful phylogenetic information in their entirety, setting the stage for the development of integrative phylogenetic analyses that combine new types of genomic characters and alignments of conserved sequences like protein-coding genes and UCEs.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Thomas Bourguignon (Thomas.bourguignon@oist.jp).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The genome assemblies of the 47 species used in this study have been published in a previous study.[22] They are available on GeneBank under bioproject PRJNA1198669. The mitochondrial genomes of the 31 species sequenced in this study are available on GeneBank under accession numbers PV938871-PV938899, PX255553, and PX260181. The accession numbers of the 16 species sequenced in previous studies are available on GeneBank under accession numbers provided in Table S2.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the OIST's Scientific Computing & Data Analysis Section (SCDA) for providing access to the OIST computing cluster. This work was supported by subsidiary funding from OIST to T.B. and by funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to D.P.M. (MC 436/5-1 and MC 436/7-1) and M.C.H. (HA 8997/1-1). A.B. was supported by the Czech Science Foundation (GAČR) grant Junior STAR no. 23-08010M.

AUTHOR CONTRIBUTIONS

C.L. and T.B. conceptualized the experiments. C.L. and S.H. analyzed the data. C.L. and T.B. wrote the manuscript. S.H., Y.M.-W., A.A.M., C.A., A.B., J.Š., M.C.H., and D.P.M. read and edited the manuscript and accepted the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Annotation of TEs
  - Supermatrices for phylogenetic tree inference
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Phylogenetic tree inference

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2025.10.019.

Received: July 4, 2025
Revised: September 12, 2025
Accepted: October 7, 2025

REFERENCES

1. Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. *6*, 361–375. https://doi.org/10.1038/nrg1603.
2. Kapli, P., Yang, Z., and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. Nat. Rev. Genet. *21*, 428–444. https://doi.org/10.1038/s41576-020-0233-0.
3. Elliott, T.A., and Gregory, T.R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140331. https://doi.org/10.1098/rstb.2014.0331.
4. Murata, S., Takasaki, N., Saitoh, M., and Okada, N. (1993). Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. Proc. Natl. Acad. Sci. USA *90*, 6995–6999. https://doi.org/10.1073/pnas.90.15.6995.
5. Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., and Okada, N. (1997). Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature *388*, 666–670. https://doi.org/10.1038/41759.
6. Whitfield, J.B., and Lockhart, P.J. (2007). Deciphering ancient rapid radiations. Trends Ecol. Evol. *22*, 258–265. https://doi.org/10.1016/j.tree.2007.01.012.
7. Rokas, A., and Carroll, S.B. (2006). Bushes in the tree of life. PLoS Biol. *4*, e352. https://doi.org/10.1371/journal.pbio.0040352.
8. Schultz, D.T., Haddock, S.H.D., Bredeson, J.V., Green, R.E., Simakov, O., and Rokhsar, D.S. (2023). Ancient gene linkages support ctenophores as sister to other animals. Nature *618*, 110–117. https://doi.org/10.1038/s41586-023-05936-6.
9. Gilbert, C., Peccoud, J., and Cordaux, R. (2021). Transposable elements and the evolution of insects. Annu. Rev. Entomol. *66*, 355–372. https://doi.org/10.1146/annurev-ento-070720-074650.
10. Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. USA *114*, 4721–4726. https://doi.org/10.1073/pnas.1621178114.
11. Shedlock, A.M., and Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. BioEssays *22*, 148–160. https://doi.org/10.1002/(SICI)1521-1878(200002)22:2<148::AID-BIES6>3.0.CO;2-Z.
12. Nikaido, M., Nishihara, H., and Okada, N. (2022). SINEs as credible signs to prove common ancestry in the tree of life: a brief review of pioneering case studies in retroposon systematics. Genes *13*, 989. https://doi.org/10.3390/genes13060989.
13. Arkhipova, I.R. (2018). Neutral theory, transposable elements, and eukaryotic genome evolution. Mol. Biol. Evol. *35*, 1332–1337. https://doi.org/10.1093/molbev/msy083.
14. Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature *513*, 375–381. https://doi.org/10.1038/nature13726.
15. Belyayev, A. (2014). Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol. *27*, 2573–2584. https://doi.org/10.1111/jeb.12513.
16. Lammers, F., Gallus, S., Janke, A., and Nilsson, M.A. (2017). Phylogenetic conflict in bears identified by automated discovery of transposable element insertions in low-coverage genomes. Genome Biol. Evol. *9*, 2862–2878. https://doi.org/10.1093/gbe/evx170.
17. Bourguignon, T., Lo, N., Cameron, S.L., Šobotník, J., Hayashi, Y., Shigenobu, S., Watanabe, D., Roisin, Y., Miura, T., and Evans, T.A. (2015). The evolutionary history of termites as inferred from 66 mitochondrial genomes. Mol. Biol. Evol. *32*, 406–421. https://doi.org/10.1093/molbev/msu308.
18. Bourguignon, T., Lo, N., Šobotník, J., Ho, S.Y.W., Iqbal, N., Coissac, E., Lee, M., Jendryka, M.M., Sillam-Dussès, D., Krížková, B., et al. (2017). Mitochondrial phylogenomics resolves the global spread of higher termites, ecosystem engineers of the tropics. Mol. Biol. Evol. *34*, 589–597. https://doi.org/10.1093/molbev/msw253.
19. Buček, A., Šobotník, J., He, S., Shi, M., McMahon, D.P., Holmes, E.C., Roisin, Y., Lo, N., and Bourguignon, T. (2019). Evolution of termite symbiosis informed by transcriptome-based phylogenies. Curr. Biol. *29*, 3728–3734.e4. https://doi.org/10.1016/j.cub.2019.08.076.
20. Hellemans, S., Wang, M., Hasegawa, N., Šobotník, J., Scheffrahn, R.H., and Bourguignon, T. (2022). Using ultraconserved elements to reconstruct the termite tree of life. Mol. Phylogenet. Evol. *173*, 107520. https://doi.org/10.1016/j.ympev.2022.107520.
21. Hellemans, S., Rocha, M.M., Wang, M., Romero Arias, J., Aanen, D.K., Bagnères, A.G., Buček, A., Carrijo, T.F., Chouvenc, T., Cuezzo, C., et al. (2024). Genomic data provide insights into the classification of extant termites. Nat. Commun. *15*, 6724. https://doi.org/10.1038/s41467-024-51028-y.
22. Liu, C., Aumont, C., Mikhailova, A.A., Audisio, T., Hellemans, S., Weng, Y.M., He, S., Clitheroe, C., Wang, Z., Haifig, I., et al. (2025). Unravelling termite evolution with 47 high-resolution genome assemblies. Preprint at bioRxiv. https://doi.org/10.1101/2025.01.20.633303.
23. Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. *20*, 275. https://doi.org/10.1186/s13059-019-1905-y.
24. Ou, S., Scheben, A., Collins, T., Qiu, Y., Seetharam, A.S., Menard, C.C., Manchanda, N., Gent, J.I., Schatz, M.C., Anderson, S.N., et al. (2024). Differences in activity and stability drive transposable element variation in tropical and temperate maize. Genome Res. *34*, 1140–1153. https://doi.org/10.1101/gr.278131.123.
25. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982. https://doi.org/10.1038/nrg2165.
26. Buček, A., Wang, M., Šobotník, J., Hellemans, S., Sillam-Dussès, D., Mizumoto, N., Stiblík, P., Clitheroe, C., Lu, T., González Plaza, J.J.G., et al. (2022). Molecular phylogeny reveals the past transoceanic voyages of drywood termites (Isoptera, Kalotermitidae). Mol. Biol. Evol. *39*, msac093. https://doi.org/10.1093/molbev/msac093.
27. Robinson, D.F., and Foulds, L.R. (1981). Comparison of phylogenetic trees. Math. Biosci. *53*, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2.

28. Linnen, C.R., and Farrell, B.D. (2007). Mitonuclear discordance is caused by rampant mitochondrial introgression in Neodiprion (Hymenoptera: Diprionidae) sawflies. Evolution *61*, 1417–1438. https://doi.org/10.1111/j.1558-5646.2007.00114.x.

29. Degnan, J.H., and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. *24*, 332–340. https://doi.org/10.1016/j.tree.2009.01.009.

30. Lo, N., Tokuda, G., Watanabe, H., Rose, H., Slaytor, M., Maekawa, K., Bandi, C., and Noda, H. (2000). Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. Curr. Biol. *10*, 801–804. https://doi.org/10.1016/s0960-9822(00)00561-3.

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

32. Legendre, F., Whiting, M.F., Bordereau, C., Cancello, E.M., Evans, T.A., and Grandcolas, P. (2008). The phylogeny of termites (Dictyoptera: Isoptera) based on mitochondrial and nuclear markers: implications for the evolution of the worker and pseudergate castes, and foraging behaviors. Mol. Phylogenet. Evol. *48*, 615–627. https://doi.org/10.1016/j.ympev.2008.04.017.

33. Legendre, F., Nel, A., Svenson, G.J., Robillard, T., Pellens, R., and Grandcolas, P. (2015). Phylogeny of Dictyoptera: dating the origin of cockroaches, praying mantises and termites with molecular data and controlled fossil evidence. PLoS One *10*, e0130127. https://doi.org/10.1371/journal.pone.0130127.

34. Inward, D.J.G., Vogler, A.P., and Eggleton, P. (2007b). A comprehensive phylogenetic analysis of termites (Isoptera) illuminates key aspects of their evolutionary biology. Mol. Phylogenet. Evol. *44*, 953–967. https://doi.org/10.1016/j.ympev.2007.05.014.

35. Mizumoto, N., and Bourguignon, T. (2021). The evolution of body size in termites. Proc. Biol. Sci. *288*, 20211458. https://doi.org/10.1098/rspb.2021.1458.

36. Nishimura, D. (2000). RepeatMasker. Biotech Software & Internet Report *1*, 36–39. https://doi.org/10.1089/152791600319259.

37. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

38. Katoh, K., Misawa, K., Kuma, K.I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.

39. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609–W612. https://doi.org/10.1093/nar/gkl315.

40. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

41. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834. https://doi.org/10.1101/gr.213959.116.

42. Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., and Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Mol. Ecol. Resour. *20*, 892–905. https://doi.org/10.1111/1755-0998.13160.

43. Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. Trends Genet. *16*, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

44. Kück, P., and Longo, G.C. (2014). FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. *11*, 81. https://doi.org/10.1186/s12983-014-0081-x.

45. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

46. Schliep, K.P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics *27*, 592–593. https://doi.org/10.1093/bioinformatics/btq706.

47. Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.

48. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
EDTA	Ou et al.[23]	https://github.com/oushujun/EDTA
panEDTA	Ou et al.[24]	https://github.com/oushujun/EDTA
RepeatMasker	Nishimura[36]	https://www.repeatmasker.org/
OrthoFinder	Emms and Kelly[37]	https://github.com/davidemms/OrthoFinder
MAFFT	Katoh et al.[38]	https://mafft.cbrc.jp/alignment/server/index.html
PAL2NAL	Suyama et al.[39]	https://www.bork.embl.de/pal2nal/
Fastp	Chen et al.[40]	https://github.com/OpenGene/fastp
metaSPAdes	Nurk et al.[41]	https://github.com/ablab/spades
MitoFinder	Allio et al.[42]	https://github.com/RemiAllio/MitoFinder
EMBOSS	Rice et al.[43]	https://emboss.sourceforge.net/
FASconCAT-G	Kück and Longo[44]	https://github.com/PatrickKueck/FASconCAT-G
IQ-TREE	Minh et al.[45]	https://iqtree.github.io/
Phangorn	Schliep[46]	https://klausvigo.github.io/phangorn/
R	N/A	https://www.r-project.org/

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used the genome assemblies of 45 termites and two cockroaches (Table S2).[22]

METHOD DETAILS

Annotation of TEs

TEs were annotated in two steps. First, we annotated each genome individually using the sensitive mode of EDTA v2.2.0 with default parameters.[23] The TE annotations of all genomes were combined to create a pan-genome TE library with panEDTA run with default settings.[24] Second, we used our pan-genome TE library to reannotate each genome individually with RepeatMasker v4.1.2-p1 run with default parameters.[36] We ran RepeatMasker separately rather than the version implemented in panEDTA for computational efficiency. The genome annotations obtained with RepeatMasker were filtered and curated with EDTA with default parameters. Both intact and fragmented TEs were retained. Annotated repeats that were not TEs, including simple repeats, low complexity regions, satellite DNA, ribosomal RNAs, small nuclear RNAs, and transfer RNAs, were removed. These analyses classified TEs populating each genome into families.

Supermatrices for phylogenetic tree inference

We generated 16 supermatrices from the genomes of 45 termites and two cockroach species[22] for phylogenetic tree reconstruction. One supermatrix was built from the content in TE families; three supermatrices were built from alignments of single-copy orthologous genes (scHOGs); three supermatrices were built from alignments of mitochondrial protein-coding gene sequences; eight supermatrices were built from the TE family content in the regions flanking UCE loci; one supermatrix consisted of the other supermatrices concatenated, including the TE family content, aligned scHOG DNA sequences with third codon position included, mitochondrial protein alignments, and the TE family content in the 100-bp regions flanking UCEs.

The first supermatrix was built using our pan-genome TE library, which contained 37,966 TE families of 45 termites and two cockroach species. We converted the genomic composition in TE families of each species into a binary matrix. The presence of a family was coded as 1 and the absence as 0.

The second, third, and fourth supermatrices were built using alignments of scHOGs. We ran OrthoFinder v.2.5.4[37] with the species tree topology and the genome annotations[22] and identified 1,410 scHOGs shared by all 47 species. We aligned their protein sequences using MAFFT v.7.508[38] with the *–auto* option. The nuclear protein alignments were concatenated as the second supermatrix. We also converted the 1,410 protein alignments of scHOGs into DNA alignments using PAL2NAL v.14[39] and concatenated them into a supermatrix. The third supermatrix consisted of the concatenated DNA sequences of 1,410 scHOGs with the third codon positions included. The fourth supermatrix consisted of the concatenated DNA sequences of 1,410 scHOGs without the third codon positions.

Please cite this article in press as: Liu et al., Robust termite phylogenies built using transposable element composition and insertion events, Current Biology (2025), https://doi.org/10.1016/j.cub.2025.10.019

The fifth, sixth, and seventh supermatrices were built using the 13 protein-coding genes of mitochondrial genomes. We obtained the mitochondrial genomes of the 47 species used herein. We used 16 previously published mitogenome sequences (Table S2). The remaining 31 mitogenomes were assembled from the short reads.[22] Briefly, raw reads were quality-trimmed using fastp v.0.20.1[40] and assembled with metaSPAdes v3.13.[41] Mitogenomes were identified and annotated with MitoFinder v.1.4.[42] The mitochondrial genomes and their annotation are available on GenBank (Table S2). The nucleotide sequences of the 13 mitochondrial protein-coding genes were translated into amino acids with the *transeq* function of EMBOSS v.6.6.0.[43] Protein sequences were aligned with MAFFT and converted into codon alignments using PAL2NAL. The protein and nucleotide sequence alignments were concatenated with FASconCAT-G_v1.04.pl.[44] The fifth supermatrix consisted of the concatenated protein sequences of the 13 mitochondrial protein-coding genes. The sixth and seventh supermatrices consisted of the concatenated DNA sequences of the 13 mitochondrial protein-coding genes with and without third codon positions, respectively.

Eight supermatrices were built from the TE family content in the flanking regions of the UCEs.[22] Each supermatrix differed in UCE flanking region lengths. We used eight flanking lengths: 50, 100, 200, 500, 1000, 2000, 3000, and 4000 bp. We only retained UCEs associated with entire flanking regions in all genome assemblies. Therefore, UCEs located near the end of contigs were not considered. We extracted presence/absence data of each TE family in the flanking regions of each UCE. More precisely, we generated a binary matrix composed of 47 rows and 37,966 columns for each UCE, with one row and one column for each of the 47 genomes and 37,966 TE families considered in this study. Presence and absence were coded as 1 and 0, respectively. Columns only composed of 1 or 0 were removed, and all remaining columns were concatenated into a supermatrix.

The last supermatrix included the four types of data concatenated in a single supermatrix. It included the TE family content, aligned scHOG DNA sequences with third codon positions included, 13 mitochondrial protein alignments, and the TE family content in the 100-bp regions flanking UCEs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Phylogenetic tree inference

We inferred 26 phylogenetic trees using the 16 supermatrices with IQ-TREE v.2.3.6[45] with the option "*-B 1000*" for bootstrap.[47] The evolutionary models were selected by ModelFinder[48] implemented in IQ-TREE v.2.3.6. For each of the nine TE-based supermatrices, we built two trees using the options "*–seqtype BIN*" and "*–seqtype* MORPH" separately. For the protein alignments of nuclear and mitochondrial genes, we selected the best amino acid substitution models with the options "*–msub nuclear*" and "*–msub mitochondrial*," respectively. For the four DNA supermatrices, we used the option "*–msub GTR*." The last supermatrix was partitioned into four subsets: one for the TE family content, one for aligned scHOG DNA sequences with third codon positions included, one for 13 mitochondrial protein alignments, and one for the TE family content in the 100-bp regions flanking UCEs. We assigned to each partition the corresponding models selected for phylogenetic analyses run with a single type of data. The options used for each IQ-TREE run, and the phylogenetic tree in Newick format, are summarized in Table S1. We quantified the topological differences between the phylogenetic tree of Liu et al.[22] based on the alignment of 27,610 UCEs and all 26 trees except the last tree, composed of four partitions, using the normalized Robinson-Foulds distance computed with the *RF.dist* function implemented in the R package *phangorn*.[46] We also quantified the topological differences between all 26 trees except the last tree, composed of four partitions, using the normalized Robinson-Foulds distance (Figure S2).