

## Research Article

**Cite this article:** Tullume-Vergara PO, Ludwig A, Yurchenko V, Coelho AC, Coser EM, Krieger MA, Teixeira MMG, Shaw JJ, Alves JMP (2025) Comparative analysis of the mobilome yields new insights into its diversity, dynamics and evolution in parasites of the Trypanosomatidae family. *Parasitology* **152**, 602–617. <https://doi.org/10.1017/S0031182025100231>



First published online: 13 June 2025

## Keywords:

CRE; INGI; mobilome; SLACS; TATE; transposable elements; trypanosomatids; VIPER

**Corresponding author:** Joao M.P. Alves;  
Email: [jotajj@usp.br](mailto:jotajj@usp.br)

# Comparative analysis of the mobilome yields new insights into its diversity, dynamics and evolution in parasites of the Trypanosomatidae family

Percy Omar Tullume-Vergara<sup>1</sup>, Adriana Ludwig<sup>2</sup>, Vyacheslav Yurchenko<sup>3</sup> , Adriano Cappellazzo Coelho<sup>4</sup>, Elizabeth Magiolo Coser<sup>4</sup>, Marco Aurélio Krieger<sup>5</sup>, Marta M.G. Teixeira<sup>1</sup>, Jeffrey Jon Shaw<sup>1</sup> and Joao M.P. Alves<sup>1</sup> 

<sup>1</sup>Department of Parasitology, Institute for Biomedical Sciences, University of Sao Paulo, Sao Paulo, Brazil; <sup>2</sup>Department of Evolution, Ecology and Behaviour, University of Liverpool, Liverpool, UK; <sup>3</sup>Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic; <sup>4</sup>Departamento de Biologia Animal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil and <sup>5</sup>Vice Presidency of Production and Innovation in Health (VPPIS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

## Abstract

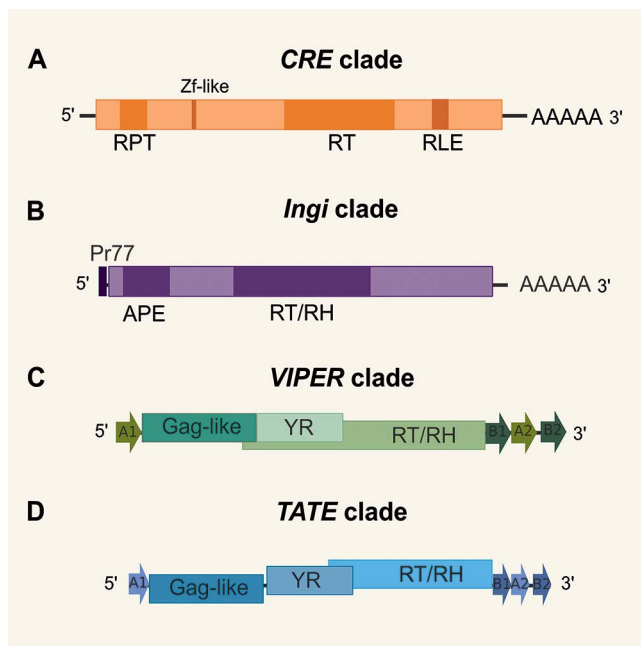
Transposable elements (TEs) have the ability to move and amplify inside the host genome, making them a pivotal source of genome plasticity. Presently, only 4 TE clades (all classified as Class I retrotransposons) have been identified in trypanosomatids. We predicted repeat content and manually curated TEs across the genomes of 57 trypanosomatids, shedding light on their proportions, diversity and dynamics. Our analysis yielded 214 TE consensus sequence models across the dataset, with abundance ranging from 0.1% to 7.2%. We found evidence of recent transposon activity in most species, with notable bursts in the *Vickermania*, *Lafontella*, *Porcisia* and *Angomonas* spp., along with *Leishmania* (*Mundinia*) *chancei*, *L. (M.) orientalis* and *L. (M.) procaviensis*. We confirmed that the 4 TE clades have colonized virtually all lineages of trypanosomatids, potentially playing a role in shaping their genome architecture. The effort of this work culminated in the establishment of the Trypanosomatid TE Database 1.0, a resource designed to standardize the TE annotation process that can serve as a foundation for future studies on trypanosomatid TEs.

## Introduction

Eukaryotic genomes harbour a significant fraction of repetitive elements (REs), which play a determinant role in driving genetic innovation in the genome (Kazazian, 2004; Bourque et al., 2018). Significant repeat categories that comprise the ‘repeatome’ include transposable elements (TEs) and certain protein-coding gene families, which are dispersed throughout the genome. Another major category consists of tandem repeats, such as satellite DNA, ribosomal RNA and simple repeats, which are arranged in consecutive copies along the genomic DNA (Woo et al., 2007; Biscotti et al., 2015). In certain organisms, TEs are the predominant type of RE within the eukaryotic genome. They were first discovered by Barbara McClintock, who referred to them as ‘controlling elements’ (McClintock, 1984). It is well known that TEs are constituted by a large variety of families that can be categorized into 2 major classes based on their transposition mechanisms: Class I TEs (retrotransposons), which relocate into the genome through a ‘copy-and-paste’ mechanism involving an RNA intermediate, and Class II TEs (DNA transposons), which move *via* a DNA intermediate mostly using ‘cut-and-paste’ mechanism of mobilization (Finnegan, 1989; Bourque et al., 2018; Gilbert et al., 2021). The retrotransposons are divided into 5 orders distinguished by the major organizational structures of their coding and non-coding domains: LTR (long terminal repeat) retrotransposons, LINE (long interspersed nuclear elements), DIRS (*Dictyostelium* intermediate repeat sequence) elements, PLE (*Penelope*-like elements) and SINE (short interspersed nuclear elements) (Wicker et al., 2007).

The family Trypanosomatidae contains a number of parasitic protistan lineages that can be divided into 2 major non-taxonomic groups: monoxenous (only 1 host, parasitizing mostly invertebrates) and dixenous (alternating between invertebrates, vertebrates and, sometimes, plants) organisms (Kaufer et al., 2017; Kostygov et al., 2021). Earlier studies have evaluated the repetitive content of the main medically important trypanosomatid species (*Trypanosoma brucei*, *T. cruzi* and *Leishmania major*) and reported retrotransposons (2–5%) as the only colonizers, while not documenting

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



**Figure 1.** Schematic structure of autonomous elements from each known clade of TE in trypanosomatids. ORFs are shown as long rectangles, and the size of elements varies within clades (not drawn to scale). (A) *CRE* clade, containing: reverse transcriptase (RT), restriction enzyme-like endonuclease motif (RLE), a poly-A tail, 1 or 2 internal repeat regions (RPT); most copies present 1 or 2 zinc finger-like structures. (B) Autonomous *INGI* elements, containing: apurinic/aprimidinic endonuclease (APE), RT and RNase h (RH), a highly conserved 77-nt sequence (pr77), a variable-length poly(A) tail. (C) *VIPER* (vestigial interposed retroelement) and (D) *TATE* (telomerase-associated transposable element) clades have similar structures, containing: a putative gag-like gene, 2 overlapped ORFs encoding for tyrosine recombinase (YR) and RT/RH, split direct repeats (SDRs), represented as arrows A1 at the 5' end and B1, A2 and B2 at the 3' end.

the presence of DNA transposons (Bringaud et al., 2007; Macías et al., 2018). More recently, using graph-based clustering of short reads, a proportion of TEs in the *T. brucei* and *T. cruzi* genomes was estimated to be even higher, at ~6% and 13%, respectively (Pita et al., 2019).

Concerning TE diversity in trypanosomatids, older studies have reported only 4 major clades, namely *INGI*, *CRE*, *VIPER* and *TATE* (Aksoy, 1991; Vazquez et al., 2000; Bringaud et al., 2002; Lorenzi et al., 2006; Peacock et al., 2007) (Figure 1). The *INGI* and *CRE* clades belong to the LINE order (also known as non-LTR retrotransposons) (Kojima, 2020). The *CRE* clade (Figure 1A) comprises a group of elements originally identified by different names (SLACS, CZAR, CRE1 and CRE2) which consistently insert at the same relative position in the spliced leader (SL) RNA genes (Aksoy, 1991; Teng et al., 1995). While these elements were previously thought to encode 2 open reading frames (ORFs) in *T. cruzi*, we verified a single ORF in most species. These elements encode a reverse transcriptase (RT) and possess a restriction enzyme-like endonuclease motif (RLE). In general, the 3' end is characterized by a poly-A tail. Most copies present 1 or 2 zinc finger (ZF)-like motifs (Fujiwara, 2015). Additionally, 1 or 2 internal repeat regions (RPT) are found in *CRE* elements of some species. These elements generate target-site duplications (TSDs, not represented in Figure 1) that vary in size.

Potentially complete *INGI* elements were found in *T. brucei* (Tbingi), *T. congolense* (Tcoingi), *T. cruzi* (L1Tc) and *T. vivax* (Tvingi) (Bringaud et al., 2009), while only remnants of the

complete *INGI* were detected in *L. major* (DIREs) (Bringaud et al., 2008). Autonomous *INGI* elements (Figure 1B) encode an apurinic/aprimidinic endonuclease (APE), RT and RNase H (RH). In the 5' end, these elements possess a highly conserved 77-nt sequence (Pr77) that works as a DNA promoter (Heras et al., 2007) and has a ribozyme activity (Sánchez-Luque et al., 2011). They have a variable-length poly(A) tail and generate TSDs (not represented in the figure). Short non-autonomous versions of *INGI* known as TbrIME, NARTc, LmsIDER and LbsIDER (not represented in Figure 1) are found in some species (Bringaud et al., 2002; Smith et al., 2009).

While there were no reports of typical LTR elements in trypanosomatid genomes, *VIPER* (vestigial interposed retroelement) and *TATE* (telomerase-associated transposable element) elements of the DIRS order (Wicker et al., 2007) were included in the LTR group in the Repbase classification (Finnegan, 1989; Kojima, 2020). *VIPER* retrotransposon was initially described as a degenerated TE family in *T. cruzi* (Vazquez et al., 2000) and more recently found in several trypanosomatid genomes, being potentially active in some species, including *T. cruzi* itself (Ribeiro et al., 2019). Besides, the *TATE* superfamily was initially discovered in *Leishmania* spp. of the subgenus *Viannia* (Peacock et al., 2007; Llanes et al., 2015) and later degenerate and potentially active copies of *TATE* were found in other trypanosomatid genomes (Ribeiro et al., 2019). Autonomous *VIPER* and *TATE* elements have a similar structure (Figure 1C and D). They encode a first ORF considered a putative Gag-like gene and 2 additional overlapped ORFs encoding for tyrosine recombinase (YR) and RT/RH. Both clades have split direct repeats (SDRs) represented as arrows A1 at the 5' end and B1, A2 and B2 at the 3' end. *VIPER* and *TATE* elements do not generate TSDs upon insertion. A short version of *VIPER*, called SIRE (short interspersed repetitive element) was identified in *T. cruzi* (Vazquez et al., 2000) and corresponds to the region encompassing A2 and B2 repeats (Ribeiro et al., 2019).

Currently, when genome assemblies are published, their repetitive content is often treated superficially, without any curation step, or overlooked altogether, which results in incomplete and, sometimes, inaccurate characterization of repeat elements, specifically TEs (Goubert et al., 2022). There is little doubt that an in-depth analysis of the evolutionary history of TEs across a broad range of non-model trypanosomatid species is essential to elucidate the diversity and dynamics of these sequences within this group. The current study delves into the mobilome of 57 genomes of the members of the family Trypanosomatidae and explores the abundance, superfamily composition and evolutionary dynamics of TEs. In addition, we report the establishment of a custom expanded TE library for genome annotation of new trypanosomatid sequences that will provide a valuable resource for future studies on TEs.

## Materials and methods

### Acquisition, quality check and preprocessing of the genomic dataset

In this study, we have compiled 2 datasets comprising assembled and unassembled genomes utilized for the analysis of TEs across 57 species from the family Trypanosomatidae (Albanaz et al., 2023; Kostygov et al., 2024). The first dataset was constituted by genome assemblies, retrieved from the National Center for Biotechnology Information (NCBI) GenBank (Sayers et al., 2021) [accessed 07/15/2024] and from the TriTryp database release 58.0 (Shanmugasundram et al., 2023). We assessed

gene completeness by employing the 'Benchmarking Universal Single Copy Orthologs' (BUSCO) tool v. 5.4.3 (Seppey *et al.*, 2019) using mode: -genome, against the Euglenozoa database (Euglenozoa\_db10), comprising 130 BUSCO orthologous genes (Kuznetsov *et al.*, 2023). Furthermore, to estimate the contiguity of the genomes, we calculated the contig N<sub>50</sub> values (Supplementary Figure S1). To build RE models with RepeatModeler v. 1.0.8 (Flynn *et al.*, 2020), we included only high-quality assemblies (37 genomes in total), selecting only 1 genome per species based on the longest contig N<sub>50</sub> value (Supplementary Table S1).

A second dataset comprising short-read genomic libraries from 20 trypanosomatid species was downloaded from the NCBI Sequence Read Archive (Katz *et al.*, 2022). The short reads were trimmed with fastp v. 0.19.4 (Chen *et al.*, 2018) using the following settings: -length-limit 50 -f 5 -t 5 -F 5 -T 5 in order to eliminate low-quality sequences and read quality was checked using fastQC v. 0.11.8 (Andrews, 2010) before and after the trimming step. We estimated the haploid genome size for the unassembled short-read genomes by applying *k*-mer frequency counting using Jellyfish v. 2.2.10 (Marçais and Kingsford, 2011) and this estimate was further used to infer the repetitive content fraction. The genome size distribution profiles were visualized using the Genoscope2 web tool (Ranallo-Benavidez *et al.*, 2020) (Supplementary Figure S2). A list of all genomes used in this study, including the source, name of isolate, contig N<sub>50</sub> value, genome coverage, assembler and publication references is shown in Supplementary Table S1.

### Identifying RE content from genome assemblies, and curation and classification of TEs

Genome assemblies were used to build a *de novo* repeat library for each of the 37 species using RepeatModeler v. 1.0.8 (Flynn *et al.*, 2020). To enhance the probability of finding new TEs, we additionally employed tools with a structure-based component: LTR\_retriever v. 2.9.0 (Ou and Jiang, 2018), which uses accurate REs input from the LTR\_Finder v. 1.07 (Xu and Wang, 2007), and LTR\_harvest v. 1.5.10 (Ellinghaus *et al.*, 2008), to predict complete LTR retrotransposons.

For each resulting RE library, filtering steps were applied to remove spurious candidate models not associated with TEs (i.e., multicopy genes or tandem repeats). Tandem repeats were identified with Tandem Repeats Finder v. 4.09 (Benson, 1999) with a threshold of 30%. Additionally, we used tRNAscan-SE v. 2.0.3 and cmscan (Chan *et al.*, 2021) on each raw library to identify and quantify tRNA, rRNAs and snoRNAs, and the Blastx tool against Uniref-90 (UniProt Reference Clusters) (Suzek *et al.*, 2015) and non-redundant NCBI databases [assessed 12/20/2023] to identify putative TEs and protein families (GP63, amastin, sialidases, GP46, tubulin, heat shock protein, etc.) within the raw libraries.

Sequences classified as unknown that were not eliminated in the previous filtering steps or confirmed as having some TE-related domain(s) were further screened using the web version of Censor (Kohany *et al.*, 2006) to analyse potential similarities to known TEs, including non-autonomous ones. TE-aid (Goubert *et al.*, 2022) was used to visualize the structure (potential ORFs and terminal repeats) and genomic coverage of the consensus. Online Conserved Domain search (Wang *et al.*, 2023) was employed to search for protein domains in sequences with a higher potential to be TEs, such as those with a defined structure or of a larger size. Only sequences with TE domains or TE-similarity were maintained in the consensus sequence curation steps.

Next, we performed a manual curation of all the potential TEs (consensus sequences) within each library generated by RepeatModeler following the main steps outlined previously (Goubert *et al.*, 2022). Briefly (i) to remove redundant consensus sequences, we ran CD-HIT-EST v. 4.8.1 (Weizhong and Godzik, 2006) with the following settings: -c 0.80 -n 5 -M 0 -aS 0.80 -g 1 -G 0 over the filtered REs library expecting to meet the majority 80-80-80 rule for the classification of the TE family (Wicker *et al.*, 2007); (ii) next, the TE library was split into individual consensus sequences using split-fasta v. 3.6 (<https://pypi.org/project/split-fasta/>); (iii) to find all members of each family, the bash script 'make\_fasta\_from\_blast.sh' (Goubert *et al.*, 2022) was used, performing a Blastn v. 2.5.0 + (Camacho *et al.*, 2009) search against the formatted genomes and retrieving extended sequences ( $\pm 0.5$ –1.5 kbp) to include as much of the ends of the TEs as possible; (iv) all recovered sequences were aligned using MAFFT v. 7.453 (Katoh and Standley, 2013); (v) the resulting MAFFT alignments were manually inspected in Aliview.jar v. 2021 (Larsson, 2014) to delimit the elements and remove flanking sequences and indels; (vi) finally, the cons tool from the EMBOSS package v. 6.6.0.0 (Rice *et al.*, 2000) was used to obtain the consensus sequence of autonomous and non-autonomous TEs. In the aforementioned steps, we also used Bedtools v. 2.27.1 (Aaron *et al.*, 2010; Quinlan, 2014) or 'make\_fasta\_from\_blast.sh' script, to extract TE copy from genome assembly and extend the length of their flanking regions. During our curation analysis, TE-trimmer, a tool to aid the manual curation of TE libraries (Qian *et al.*, 2024), was published. We tested it and found it helpful in visualizing the PFAM protein domains (Mistry *et al.*, 2021); however, we observed that the consensus sequences were frequently incorrect, particularly shrinking the DIRS and extending the CRE elements.

To help delimit some elements, we used Blastn with the parameters 'align 2 or more sequences' and 'somewhat similar sequences (blastn)', adjusting the *e*-value threshold to 10 and using the same sequence as the query and subject. This strategy was used to find the target site duplications (TSDs) generated by CRE elements since their consensus is often extended due to their insertion in a repetitive region. Moreover, this was also used to identify the SDRs of VIPER and TATE.

For the purpose of classification, the consensus sequences were categorized according to their similarity to known TEs in the Repbase Database (Bao *et al.*, 2015) utilizing the web browser version of Censor (Kohany *et al.*, 2006). The sequences were named sequentially using a format that includes the species name abbreviation, a unique identifier, and the TE family classification with the order and clade/superfamily information (INGI, CRE, TATE or VIPER) (e.g., Tcru-1#LINE/CRE). Some known TE families were not repetitive enough in some of the genomes to be identified by RepeatModeler. To overcome this, a Tblastn search was performed against each genome. The TATE, VIPER, CRE and INGI canonical proteins from Repbase Database were used as queries using an *E*-value threshold of 1e-9. The sequences were retrieved using getfasta from the Bedtools package, and the steps described above were followed to retrieve the copies and make a consensus sequence, whose classification was confirmed using the Censor webtool.

Using the approach specified above, 1 accurate species-specific TE library was obtained for each genome assembly. These custom TE libraries were merged for downstream analysis by dnaPipeTE (Goubert, 2023). All of these procedures resulted in the creation of the custom Trypanosomatid TE DataBase v. 1.0 available on GitHub ([https://github.com/percytullume/TEs\\_trypanosomatids](https://github.com/percytullume/TEs_trypanosomatids)).



### Improving the annotation provided by RepeatMasker to infer copy number

To get the annotation of the TEs, each of the 37 curated species-specific TE libraries was mapped against its corresponding genome assembly using RepeatMasker v. 4.1.2 (Tarailo-Graovac and Chen, 2009) with the following options: `-s -excln -a -gccal -norna -lib -nolow`. To improve the accuracy of copy number estimation, we utilized the 'One-code-to-find-them-all' script (Bailly-Bechet et al., 2014), which provides precise TE copy coordinates and accurate quantification of TE families (Table 1). There was no inference of TE copy numbers for unassembled genomes since the dnaPipeTE tool uses a low coverage genome ( $\sim 0.15 \times$ ); therefore, the retrieved elements are frequently fragmented, generating unreliable estimates of copy number.

### Kimura distance-based distribution analysis

From the RepeatMasker output, the .tbl file was used to estimate the TE coverage (proportions) for all trypanosomatid genomes. In addition, the .align file was used to estimate the divergence of copies and their family consensus sequence using the Kimura distance model (Kimura 2-parameter [K2P]) (Kimura, 1980) with the calDivergenceFromAlign.pl script from the RepeatMasker package. We employed the Kimura distances with correction for CpG pairs for all trypanosomatid genera except *Leishmania*, where the parameter `-noCpGMod` was used, since DNA methylation has been documented for *Trypanosoma* sp. but not *Leishmania* spp. (Militello et al., 2008; Cuypers et al., 2020). We also used the 'createRepeatLandscape.pl' script from RepeatMasker to generate landscape bar plots illustrating the temporal activity of TEs within the genomes. Furthermore, TE families were grouped into the orders LINE (*INGI* and *CRE*) and DIRS (*TATE* and *VIPER*).

### Repeat context analysis using the dnaPipeTE pipeline on the unassembled dataset

The second dataset included raw sequence read libraries from 20 trypanosomatid species. The abundance and proportion of each RE were estimated with the dnaPipeTE v. 1.4c (Goubert, 2023). The pipeline utilizes high-quality short-read sequencing libraries (either forward or reverse reads). Initially, to avoid overestimating REs, sequence reads aligning to the respective mitochondrial genomes were excluded for each species using the BBDuk package v. 37.62 from BBTools (Bushnell et al., 2017). dnaPipeTE uses Trinity v. 2.5.1 (Grabherr et al., 2011) to assemble RE contigs from low genome coverage ( $<1 \times$  subsamples), enabling the identification of these sequences in species lacking high-quality genome assemblies. We performed tests of low coverage from  $0.05 \times$  to  $0.21 \times$  in intervals of  $0.02 \times$  (9 runs) on all datasets, as suggested by Goubert (2023) (summarized in Supplementary Table S2), to find the highest contig  $N_{50}$  (optimal assembly) in the assembly step of the pipeline (best coverage species-specific). To infer genome size for dnaPipeTE, we employed Genoscope2 (Supplementary Table S4; Supplementary Figure S2). To improve the classification accuracy and annotation of TEs, we ran dnaPipeTE twice: (i) in the first run, we used the trypanosomatid RepeatModeler library (obtained in this work as described above), with the following parameter: `-RM_lib` (custom library). This custom RepeatModeler library gathered 37 curated TE libraries to identify potential candidate TEs; (ii) in the second run, we employed the correctly classified species-specific TE library identified in the first dnaPipeTE run to

**Table 1.** Diversity of Class I TEs in trypanosomatid genomes

Species	<i>INGI</i>	<i>CRE</i>	<i>TATE</i>	<i>VIPER</i>
<i>Angomonas deanei</i>	33	11	104	1
<i>Blechnomonas nonstop</i>	11	5	182	0
<i>Crithidia bombi</i>	20	1	9	21
<i>Crithidia expoeiki</i>	111	6	14	37
<i>Crithidia fasciculata</i>	49	10	299	157
<i>Herpetomonas samuelpes-soai</i>	37	5	132	98
<i>Kentomonas sorsogonicus</i>	18	6	352	166
<i>Lafontella mariadeanei</i>	63	533	248	1185
<i>Leishmania aethiopica</i>	1671	0	8	0
<i>Leishmania amazonensis</i>	1077	0	3	0
<i>Leishmania braziliensis</i>	1308	11	69	37
<i>Leishmania chancei</i>	435	0	782	1
<i>Leishmania donovani</i>	1200	0	4	21
<i>Leishmania enriettii</i>	1118	0	445	2
<i>Leishmania guyanensis</i>	1422	7	102	118
<i>Leishmania infantum</i>	1293	0	4	43
<i>Leishmania lainsoni</i>	977	27	19	12
<i>Leishmania major</i>	1240	0	2	0
<i>Leishmania martiniquensis</i>	116	0	176	1
<i>Leishmania mexicana</i>	1437	0	7	0
<i>Leishmania orientalis</i>	996	0	751	2
<i>Leishmania procaviensis</i>	1002	0	465	3
<i>Leishmania shawi</i>	1345	6	48	106
<i>Leishmania tarentolae</i>	543	0	0	0
<i>Leishmania tropica</i>	1350	0	3	51
<i>Leptomonas pyrrhocoris</i>	2	14	64	250
<i>Lotmaria passim</i>	98	11	0	0
<i>Porcisia hertigi</i>	50	4	255	0
<i>Trypanosoma brucei</i>	1314	65	0	76
<i>Trypanosoma (b.) evansi</i>	295	6	0	56
<i>Trypanosoma (b.) equiperdum</i>	312	8	0	22
<i>Trypanosoma congolense</i>	451	7	0	188
<i>Trypanosoma cruzi</i>	439	775	0	1764*
<i>Trypanosoma melophagium</i>	148	3	22	167
<i>Trypanosoma vivax</i>	1189	78	0	805
<i>Vickermania ingenoplastis</i>	5	864	1237	863
<i>Zelonia costaricensis</i>	43	14	210	24

The number of copies in 4 superfamilies was assessed by RepeatMasker and the 'one-code-to-find-them-all' script. \*The copy number of *VIPER* in *T. cruzi* also includes the short version, SIRE.

infer accurately the TE% coverage in the final dnaPipeTE library for each species. Additionally, these TE sequences were confirmed with the Censor webtool and Blastx. Gene families and satellites

were removed as described above, and the classification of TEs was confirmed using the Censor webtool. The resulting TE libraries from unassembled genomes were added to the Trypanosomatid TE DataBase v. 1.0.

### Estimation of the relationship between genome size and RE abundance

The abundances of RE and TEs vs genome assembly size were used for the correlation tests after the data were transformed to a logarithmic scale using tidyverse in R v. 4.2.1 (Wickham, 2016). To estimate correlation and the Spearman rank sum, with  $\alpha = 0.005$  (*lm* method), we used the Ape package in phytools 2.0 (Revell, 2024) and ggplot2 in R (Paradis and Schliep, 2019). We applied the Spearman rank correlation test as our data does not follow the normal distribution, as assessed by the Kolmogorov–Smirnov (KS) test. Additionally, using linear regression equations, we further inferred the relationship between genome size and the aforementioned traits with the Hiplot web tool (Li et al., 2022). Lastly, we also applied a phylogenetically independent contrasts (PICs) method (Felsenstein, 1985) to test a probable phylogenetic effect over the correlation among TEs/RE and genome size by using the pic function of Ape.

### Phylogenetic analysis of the family Trypanosomatidae

We employed a set of genes recovered through BUSCO analysis to infer the species tree. We extracted 40 single-copy genes from all the 57 trypanosomatid species from the GenBank [accessed 07/15/2024] (Supplementary Table S1). To validate the accuracy of the orthologs, we ran OrthoFinder v. 2.5.4 with the default settings (Emms and Kelly, 2019). The resulting proteins from the orthologous groups were aligned using MAFFT v. 7.453 with the auto option (Kato and Standley, 2013), followed by the trimming step in TrimAl v. 1.4 (Capella-Gutiérrez et al., 2009) to remove gaps using -option 'automated1'. Next, the alignments were concatenated using FASconCAT v. 1.04 (Kück and Longo, 2014) to build a supermatrix of sequences from the 57 species, resulting in an alignment of 24 659 amino acids in length and 3.7% missing data. The best substitution model was automatically selected by the ModelFinder with the MFP option (Kalyanamoorthy et al., 2017). A maximum likelihood (ML) tree was inferred using IQTree2 v. 2.0.4 with 100 bootstrap samples, and later on 1000 UFBS (ultra-fast bootstrap) with default options (Minh et al., 2020). Additionally, we ran RAXML-NG v. 1.1.0 with 100 bootstrap samples to compare topologies (Kozlov et al., 2019). By employing 2 likelihood-based phylogenetic inference tools we aimed to uncover potential disparities in tree topologies. Finally, the phylogenetic trees were rendered in the iTOL v. 6 (Letunic and Bork, 2024) web server and further refined with Inkscape v. 0.92.5 to add the status (presence/absence) of each TE clade based on the analyses described above.

## Results

### Trypanosomatidae species representation, genome quality and TE library construction

Recently, the taxonomy of the family Trypanosomatidae has undergone a revision resulting in a system with 7 subfamilies and 24 genera (Maslov et al., 2019; Kostygov et al., 2024). In light of this, our dataset included 57 species representing 19 genera (Supplementary

Figure S3; Supplementary Table S1). Highlighting 2 medically important genera, there were 25 *Leishmania* spp. belonging to all 4 subgenera (*Leishmania*, *Mundinia*, *Sauroleishmania* and *Viannia*), and 9 *Trypanosoma* spp. We also used 4 species whose genome sequences were obtained in our laboratory, namely *Lafontella mariadeanei*, *Herpetomonas samuelpessoai*, *Sergeia* sp. (isolate 2467) and *Blechnomonas* sp. (isolate 303E). Most species belonging to *Leishmania* genus had high BUSCO values (Supplementary Figure S4), except for *Leishmania lainsoni*, which had 10 fragmented genes. The BUSCO scores were similar to those reported for *L. major* (Friedlin), which serves as a benchmark for completeness (Camacho et al., 2021). Missing genes were documented in *Blastocrithidia nonstop* (7), *Trypanosoma b. equiperdum* (7), *Trypanosoma congolense* (5), *H. samuelpessoai* (5), *Kentomonas sorsogonicus* (4) and *L. mariadeanei* (1) (Supplementary Figure S4).

To provide a comprehensive overview of TEs in the family Trypanosomatidae, we ran RepeatModeler and dnaPipeTE pipelines across 37 genome assemblies and 20 unassembled short-read libraries, respectively. As a result, we obtained 12 301 consensus sequence models with RepeatModeler, while dnaPipeTE retrieved 10 484 contig models. The consensus models depicted the overview of REs. Because we focused solely on TEs, the raw repeat libraries underwent several filtering steps, to remove potential false positives (Figure 2A).

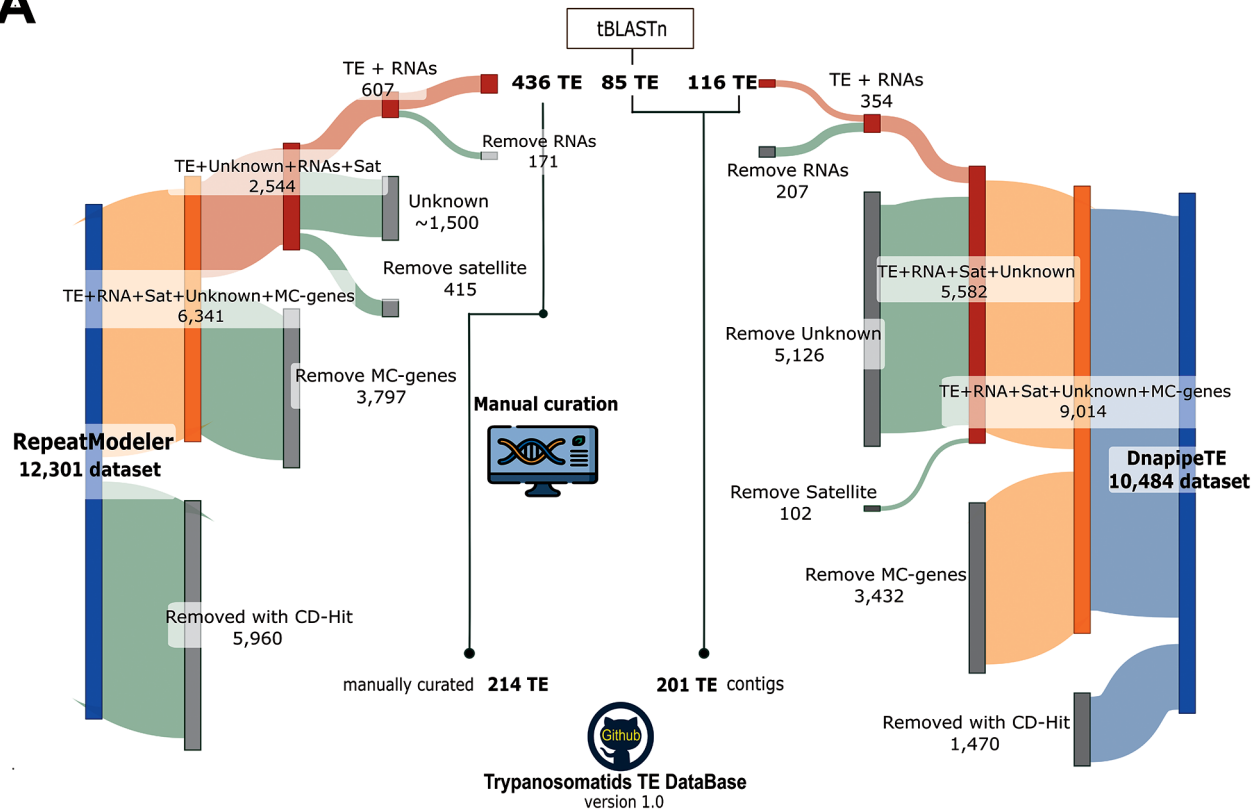
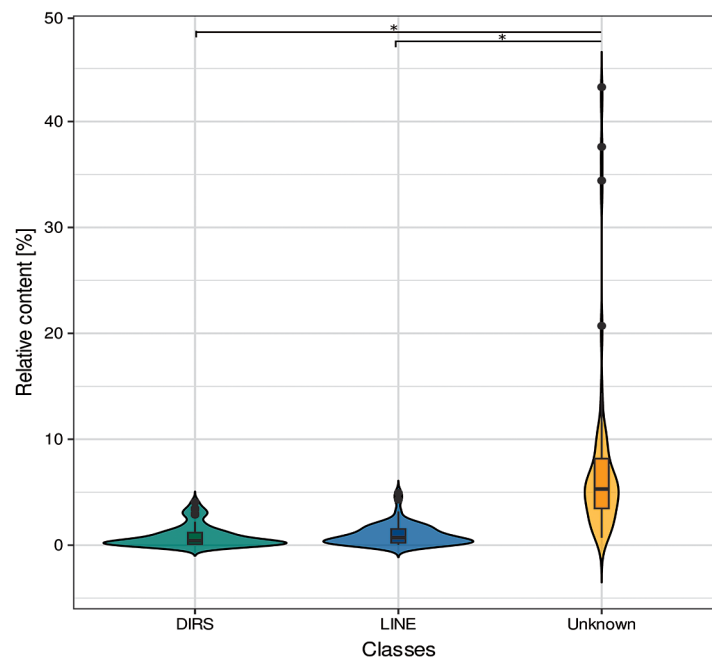
From the RepeatModeler approach, we recovered a set of 436 sequences with predicted TE-related proteins using Blastx. Additionally, some TEs were recovered using Tblastn, adding 85 model sequences. We also tried to classify the sequences annotated by RepeatModeler as 'unknown' (~1500 sequences). Several of them were discarded because they represented additional multicopy genes or tandem repeats, while a few were confirmed as TEs. The remaining sequences lacked any traits of TEs or similarity to known TEs. We chose to retain only sequences with confirmed classification in the final TE library, totalling 214 TE families (Figure 2A). As expected, the sequences recovered from dnaPipeTE were TE contigs (very fragmented), and, after the filtering steps, only 116 TE contigs were confirmed.

From all REs found, we confirmed only 4 previously known trypanosomatid TE clades (*ING1*, *CRE*, *VIPER* and *TATE*). A few potential TEs classified as DNA transposon (*Helitron*) and LTRs (*Gypsy* and *Copia* elements) were not confirmed after the curation steps, as they were confirmed to be multicopy genes (false positives). Notably, the majority of the REs found were classified as unknown (Figure 2B).

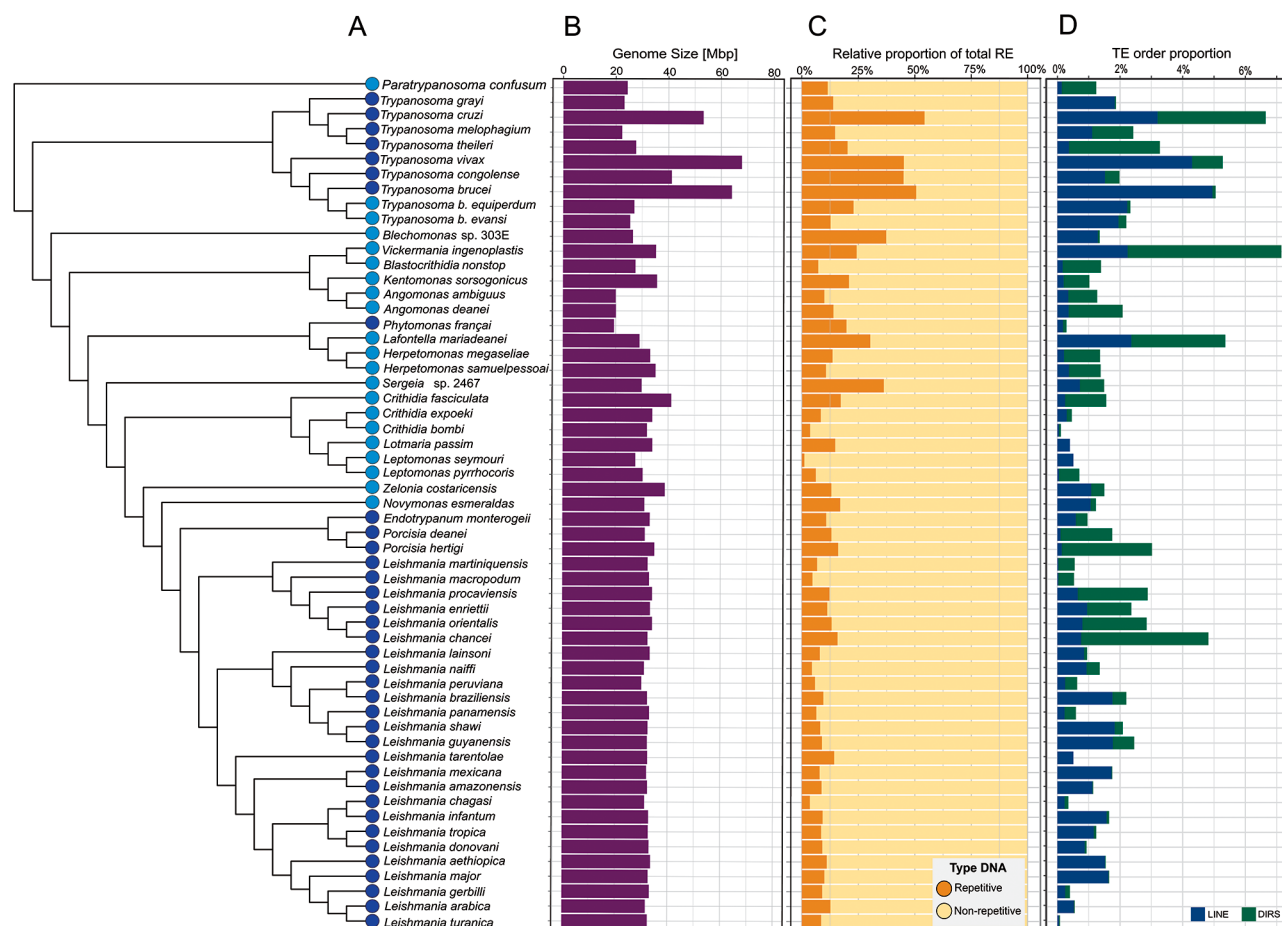
### RE and TE content of trypanosomatid genomes

The proportion of RE (multicopy genes, unknown, RNAs) and TE content was mapped onto the phylogenetic tree of trypanosomatids (Figure 3A; Supplementary Figure S5). The genome size of trypanosomatids varies widely, ranging from ~20 Mbp in *Angomonas deanei* and *Phytomonas françai* to 67 Mbp in *Trypanosoma vivax* (Figure 3B; Supplementary Table S4). The genus *Leishmania* displays a lesser variation in length, ranging from ~30 to ~35 Mbp. Overall, the haploid genome size in trypanosomatids was spread around the mean of 33.0 Mbp with a standard deviation (s.d.) of 8.3 Mbp.

The repetitive content also varies widely among the trypanosomatid taxa (Figure 3C, Supplementary Table S4) from 3.7% (*Crithidia bombi*) to 56.1% (*T. cruzi*) with a mean of 15.2% (s.d. of 11.9%). As expected because of their pivotal role, rRNA and snoRNAs gene families were detected among the RE across all

**A****B**

**Figure 2.** Overview of the steps to curate the final TE database and comparison among RE types. (A) Sankey plot displaying the raw RE libraries built by RepeatModeler (RM) and dnaPipeTE pipelines. For each phase, the grey portion indicates the number of consensus models removed by the filtering process, while the coloured segment depicts the number that continued to the next step. (1) Initial clustering reduced the number of family copies (5960 for RM and 1470 for dnaPipeTE) to streamline curation and minimize redundant TE models. (2) Multicopy genes (3797 for RM and 3432 for dnaPipeTE) were identified and removed using a homology-based approach. (3) Potential satellite sequences (415 for RM and 102 for dnaPipeTE) and ‘unknown’ sequences (1500 for RM and 5126 for dnaPipeTE) were excluded. (4) RNA-related families (171 for RM and 207 for dnaPipeTE) were detected and separated from the TE dataset. Following this pipeline, 436 TE sequences were manually curated, resulting in 214 canonical TE models. Additionally, a small number of TEs were incorporated based on Tblastn results. (B) Violin plot representing the genome occupancy of 2 TE orders as DIRS, LINE, along with unknown from 57 assessed trypanosomatid genomes. Pairwise Wilcoxon rank test with Bonferroni correction was used for comparison among classes, where \* ( $P$ -value < 0.01) indicates significance (Supplementary Table S3).



**Figure 3.** Contribution of REs and TEs to trypanosomatid genomes. (A) A cladogram displays the relationship of the 57 trypanosomatid species of 7 subfamilies used in this study. Monoxenous and dioxenous parasites are marked by light and dark blue circles, respectively. (B) Genome sizes are shown in Mbp. (C) Proportion of repetitive and non-repetitive content in each species. (D) Total proportion of each TE order, DIRS (green) and LINE (blue).

genomes. Moreover, the multicopy gene families represent the most abundant elements in the interspersed repeats fraction of the trypanosomatid genomes, ranging from 44.1% in *T. cruzi* to 0.7% in *C. bombi* (Supplementary Figure S5; Supplementary Table S4).

In terms of TE proportions, trypanosomatids presented a mean of 1.8% (s.d. of 1.6%) with most species having less than 3% of their genomes present as TEs. Higher proportions of TEs were found in *Vickermania ingenoplastis* (7.2%), *T. cruzi* (6.7%), *Lafontella mariadeanei* (5.4%), *T. vivax* (5.4%), *T. brucei* (5.0%), *Leishmania chancei* (4.8%), *Trypanosoma theileri* (3.3%) and *Porcisia hertigi* (3.0%). Conversely, a lower proportion was documented in *C. bombi* with ~0.1% (Figure 3D; Supplementary Table S4). Notably, we detected greater proportions of TEs in some species of the subgenus *Mundinia*, including *L. chancei*, *Leishmania procavensis* (2.9%), *Leishmania orientalis* (2.9%) and *Leishmania enriettii* (2.3%), but with the exception of *Leishmania macropodum* (0.1%).

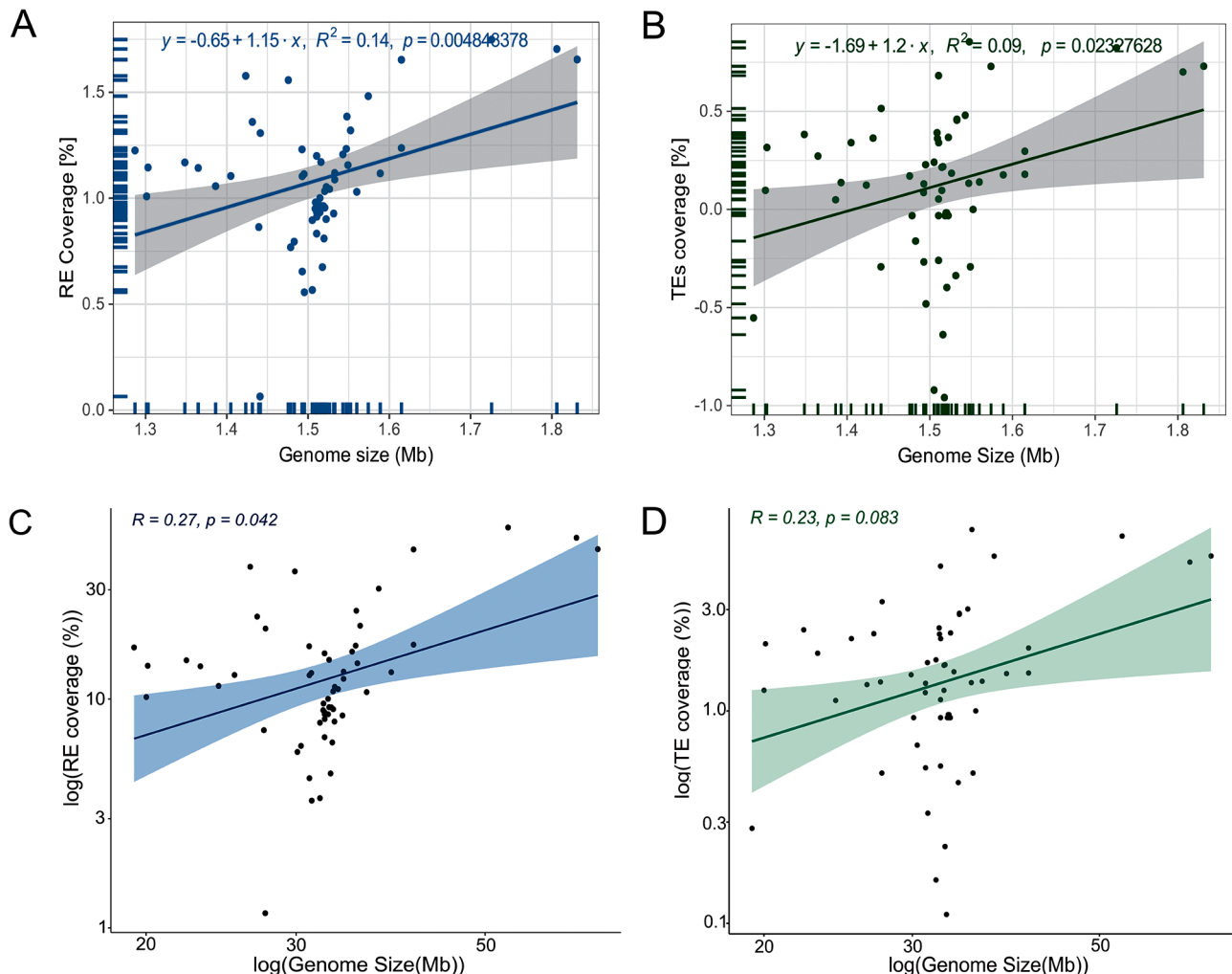
Our results show that LINEs are more widely distributed than DIRS, and their proportions vary among the genomes (Figure 3D; Supplementary Table S4) comprising up to 4.9% in *T. b. brucei*, 4.3% in *T. vivax*, and 3.2% in *T. cruzi* genomes. In contrast, their proportions in *L. macropodum* and *L. martiniquensis* accounted for ~0.02%. The DIRS elements were present at a higher proportion in *L. chancei* (4.1%) and *T. cruzi* (3.5%) and were either absent or present at a low proportion (0.01%) in some *Leishmania* spp.

This study is the first to report the TE proportions in several trypanosomatid species, with some showing relatively high values not previously reported for trypanosomatid species: *T. b. equiperdum* (2.3%), *T. vivax* (5.4%), *Trypanosoma melophagium* (2.4%), *L. mariadeanei* (5.4%), *B. nonstop* (1.4%), *L. guyanensis* (2.5%), *L. shawi* (2.2%), *L. lainsoni* (0.9%) and *V. ingenoplastis* (7.2%).

#### Correlation test of REs and TEs vs trypanosomatid genome size

The differences in trypanosomatid genome size prompted us to inquire how much REs and TEs contribute to this trait. The linear regression model revealed that RE and TE coverages have a weak correlation with the genome sizes ( $R^2 = 0.14$ ,  $P = 0.00484$ ;  $R^2 = 0.09$ ,  $P = 0.02327$ , respectively) (Figure 4A, B). We confirmed that our data does not follow the normal distribution based on KS statistics (KS = 0.247,  $P = 0.0014$ ). Additionally, the Spearman rank correlation test revealed a significant, albeit modest, positive correlation between the abundance of all REs and genome size (Spearman's rank sum test  $\rho = 0.27$ ,  $P = 0.042$ ) (Figure 4C). In contrast, the correlation between genome size and TEs was not statistically significant (Spearman,  $\rho = 0.23$ ,  $P = 0.083$ ) (Figure 4D). We further tested PICs to correct for the non-independency of traits among species. These tests revealed significant relationships of genome size with RE ( $R^2 = 0.5316$ ,  $P = 7.639 \times 10^{-11}$ )





**Figure 4.** Scatterplots of correlation between genome size vs REs and TEs [%] across 57 trypanosomatid genomes. (A) Linear regression plot between genome size and the percentage of REs. (B) Linear regression plot between assembly genome size and the percentage of TEs. (C) Correlation plot between genome size and the percentage of REs. (D) correlation plot between genome size and the percentage of TEs. Lines: linear regression, shaded area: confidence interval.

and TEs ( $R^2 = 0.4897$ ,  $P = 8.367 \times 10^{-10}$ ) (Supplementary Figure S6).

#### Distribution of TEs across the family Trypanosomatidae

In the last decade, molecular phylogenetic analyses have significantly enhanced our understanding of the extended relationships within the family Trypanosomatidae, providing valuable insights into the evolutionary processes in this group (Yurchenko et al., 2016; Kostygov and Yurchenko, 2017; Espinosa et al., 2018; Kaufer et al., 2019; Kostygov et al., 2020). In this sense, our study contributed to extending this analysis by including newly sequenced genomes of *Lafontella mariadeanei*, *Herpetomonas samuelpessoai*, *Blechnomonas* spp. *Sergeia* spp. along with *L. shawi* and *L. guyanensis*.

The ML method was applied to the supermatrix to recover the best and most robust trypanosomatid phylogenetic tree, as depicted in Figure 5. As expected, all the subfamilies formed well-supported clades (100%), but there were some exceptions, such as observed between the Leishmaniinae and *Sergeia* spp. clades, with a bootstrap support of 78%. Moreover, the clade that included *T. cruzi* and *T. grayi* had a low bootstrap support (51%).

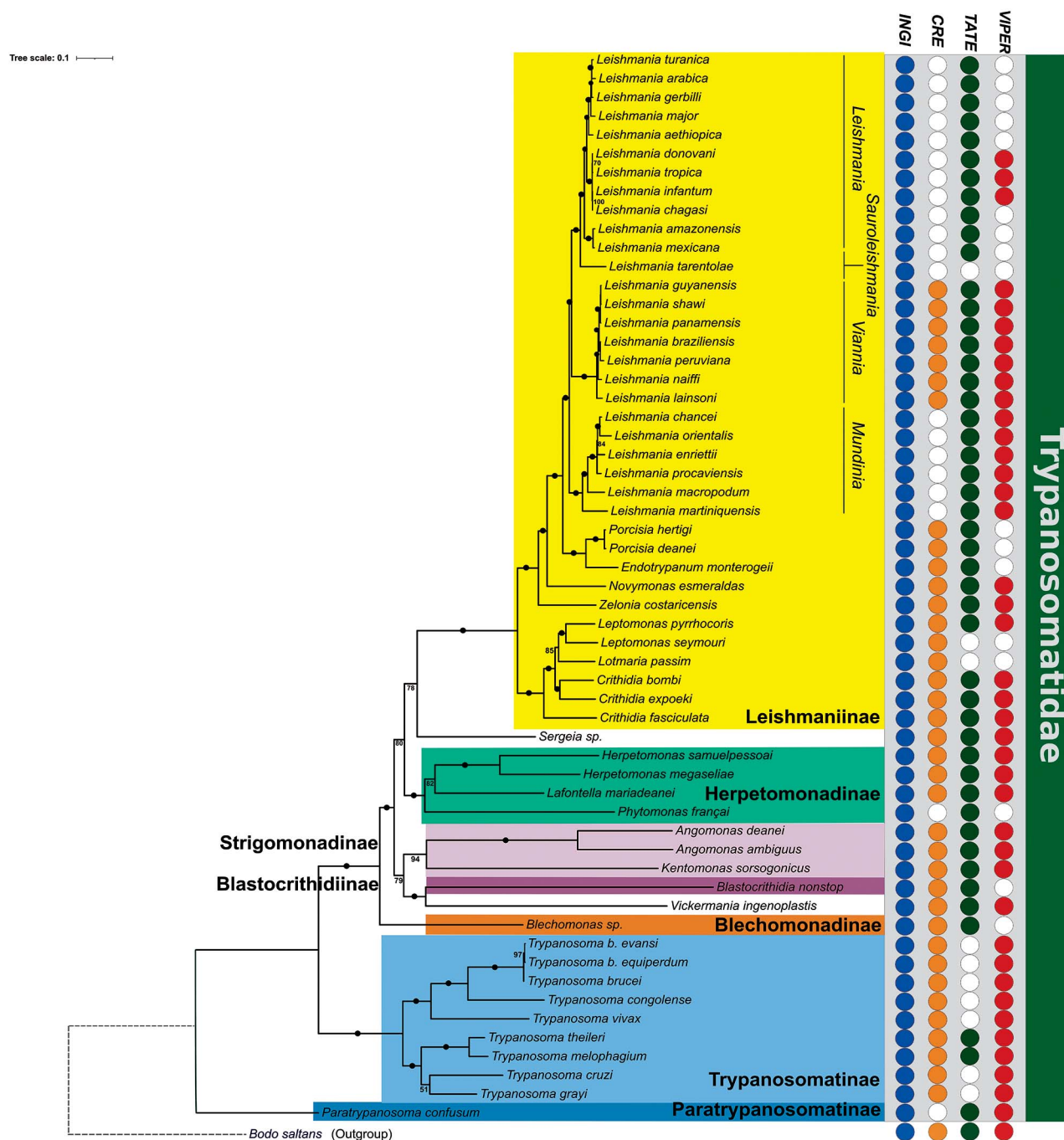
Additionally, the RAxML tool showed a different topology for this clade (Supplementary Figure S7).

Understanding these phylogenetic relationships is essential to visualize the evolutionary pattern of the TEs (Figure 5). Notably, *Bodo saltans*, a free-living kinetoplastid species, is also known to harbour the 4 TE clades (*CRE*, *INGI*, *TATE* and *VIPER*), indicating that these elements were likely present in the last common ancestor of all trypanosomatids (Jackson et al., 2008, 2016; Ribeiro et al., 2019).

The *INGI* superfamily is the most widespread, found in all 57 lineages analysed with a number of copies varying from only a few in *Leptomonas pyrrocoris* to 1671 in *L. aethiopica*. However, these abundant copies in *Leishmania* appear to be mostly non-autonomous (Table 1). In contrast, *CRE* clade showed a patchy distribution, with high numbers in *V. ingenoplastis*, *T. cruzi* and *L. mariadeanei*. Four independent events of loss could explain the distribution pattern of these elements: in *P. confusum*, in *P. françai*, in the ancestor of the subgenus *Mundinia*, and in the common ancestor of the subgenera *Leishmania* and *Sauroleishmania*.

*VIPER* and *TATE* also displayed a patchy distribution pattern. Here, we reported for the first time *VIPER* in multiple monoxenous genera and remnants in several *Leishmania* spp. High copy





**Figure 5.** Phylogenetic relationships across 57 species of trypanosomatids. The bold letters show the position of the 7 subfamilies that currently constitute the family Trypanosomatidae. Bootstrap supports are shown at nodes, and maximum bootstrap support (100%) is shown with black circles. The distribution of the 4 retrotransposon clades is shown on the right, with coloured circles indicating the presence and white circles indicating the absence of a given element. *Bodo saltans*, a free-living phagotroph, was added as an outgroup. The scale bar indicates the number of substitutions per site.

numbers in *T. cruzi*, *L. mariadeanei*, and *V. ingenoplastis* may reflect retained activity in these lineages. Moreover, we found TATE elements in all *Leishmania* spp. investigated (except *L. tarentolae*) with high loads in the *Mundinia* subgenus and *V. inglenoplastis*. In *Trypanosoma* spp., TATE has been previously detected only in *T. theileri* (Ribeiro *et al.*, 2019). Here, we also identified this element in the genome of a closely related species, *T. melophagium*.

Our analysis revealed a highly variable number of TE copies across 37 assembled trypanosomatid genomes. However, it is

important to recognize that these counts do not necessarily reflect complete or functional copies, as they may include remnant fragments. The copy numbers can also be influenced by the quality of the genome assemblies and TE libraries. For instance, we noticed that the TE consensus sequences from *Zelonia costaricensis* are fragmented due to the fragmentation of the genome itself (Tullume-Vergara *et al.*, 2023), potentially leading to overestimation of the copy numbers. On the other hand, the CRE elements could be underrepresented in certain species if the SL-RNA region,

where these elements typically insert, is not well-covered in the genome assembly. Therefore, while this provides an overall view of TE load across species, these findings should be interpreted with caution.

### TE transposition activity during trypanosomatid evolution

The K2P-based copy divergence analysis was performed to assess the diversity and dynamics of the trypanosomatid mobilome in detail. The TE landscapes illustrate the distribution of genome coverage of copies for LINE and DIRS sequences relative to their divergence from the consensus model sequence. The shape of a distribution landscape can be categorized as follows: (i) recent events of TE activity (transposition bursts) are characterized by low divergence scores (<5% divergence from the consensus) and are depicted with L-shaped peaks (Barrón et al., 2014); (ii) bimodal (2 peaks) and (iii) 'bell-shaped' curves depict an equilibrium between transposition and excision events over evolutionary time (Le Rouzic and Capy, 2005).

The TE landscape distribution for 30 trypanosomatid genomes was compared among genera and species to describe some main events (Figure 6; Supplementary Figure S8). The CRE superfamily has acquired the majority of copies relatively recently (compared to the other retrotransposons) in almost all genomes possessing these elements. Very recent activity of this superfamily (including some notable bursts) can be seen in *H. samuelpessoai*, *L. mariadeanei*, *L. braziliensis*, *L. guyanensis*, *L. lainsoni*, *Lotmaria passim*, *T. brucei*, *T. congolense*, *T. cruzi*, *T. melophagium*, *T. vivax* and *V. ingenoplastis*. The other TE families exhibit different evolutionary trajectories depending on the species.

Within the genus *Trypanosoma*, the landscape distributions generally show a multimodal shape with INGI and VIPER being well represented. As expected due to their phylogenetic proximity, *T. brucei*, *T. b. equiperdum* and *T. b. evansi*, show a similar pattern with a very low number of ancient insertions, and a recent activity peak of INGI (K2P of 3) and CRE (K2P 0), although the CRE peak is more prominent in *T. brucei*, *T. congolense*, *T. cruzi* and *T. vivax* have a recent activity peak of CRE, VIPER and INGI and additional, more ancient peaks (K2P 9 and 16 in *T. cruzi*; K2P 12 in *T. vivax*; K2P 5 and 20 in *T. congolense*). Interestingly, in contrast to other *Trypanosoma* spp., *T. melophagium* does not present a very recent peak of INGI elements. On the other hand, in this species, the TATE elements appear to be the most recently active TEs, followed by the VIPERs and CREs.

The divergence landscape for *V. ingenoplastis* and *K. sorsogonicus* displayed a multimodal distribution with a significant proportion of sequences presenting divergence below 5%, which suggests a recent activity for the CRE and VIPER clades. *Blastocrithidia nonstop* showed a multimodal shape, with the first peak occurring earlier (from 10 to 5 of K2P), being dominated by TATE elements. This pattern of *B. nonstop* is likely to be associated with the accumulation of TEs in its genome.

Strikingly, we observed an L-shape distribution for *A. deanei*, *C. fasciculata*, *H. samuelpessoai*, *L. mariadenaei*, *L. passim*, *Porcisia hertigi* and *Z. costaricensis* with increasing trend spanning from ~10% to 0% of K2P divergence. This pattern indicates a more recent burst of activity with a very low quantity of older copies in some of these species. The CRE, TATE and VIPER elements were well noticeable in *H. samuelpessoai* and *L. mariadenaei*. The TATE elements are the most abundant in *A. deanei*, *P. hertigi* and *Z. costaricensis* genomes, while TATE and CRE stand out in *L. passim*.

Within the genus *Leishmania*, similar patterns can be observed for the species belonging to the same subgenera. In the subgenus *Mundinia*, which includes *L. chancei*, *L. enriettii*, *L. orientalis* and *L. procaviensis*, multi-peaked distributions of TATE and INGI were observed. The TATE elements are highly abundant in this group although this expansion was primarily due to some more ancient events of transposition (highest peaks of K2P 5–10). Furthermore, in subgenus *Viannia*, bimodal peaks were observed in *L. braziliensis*, *L. guyanensis*, *L. lainsoni* and *L. shawi* (although less pronounced in the latter). Recent activity of CRE and TATE elements is indicated for almost all these species, while INGI elements presented more ancient activity. Lastly, the TE landscape in the *Leishmania* subgenus is dominated by the INGI clade exhibiting an ancient peak, similar to what is observed in the subgenera *Mundinia* and *Viannia*. This pattern is expected, given that only non-autonomous INGI-related elements (SIDERs and DIREs) are present in the genomes of *Leishmania* spp. Interestingly, *L. donovani*, *L. infantum* and *L. tropica* still contain remnants of the VIPER elements, reported for the first time in this study.

Unexpectedly, a low percentage of copies with very low divergence (K2P 0) is observed despite the absence of active TEs. Noteworthy, this could suggest that some TEs are being duplicated through mechanisms other than transposition, such as through segmental genomic duplications. This possibility could help explain the persistence of non-autonomous TEs even when active elements are absent and merits further investigation. In addition, some of these observations could result from false duplications in the assemblies. Even applying, purging steps, such artefacts are known to occur (Ko et al., 2022).

For several species, we observe a very low amount of ancient elements. Possible explanations include (1) loss of active TEs followed by a recent invasion of active families from other species, although no evidence of horizontal TE transfer has been documented for these species; and (2) continuous production of new copies by active TEs with rapid turnover, which may lead to the elimination of older, inactive copies.

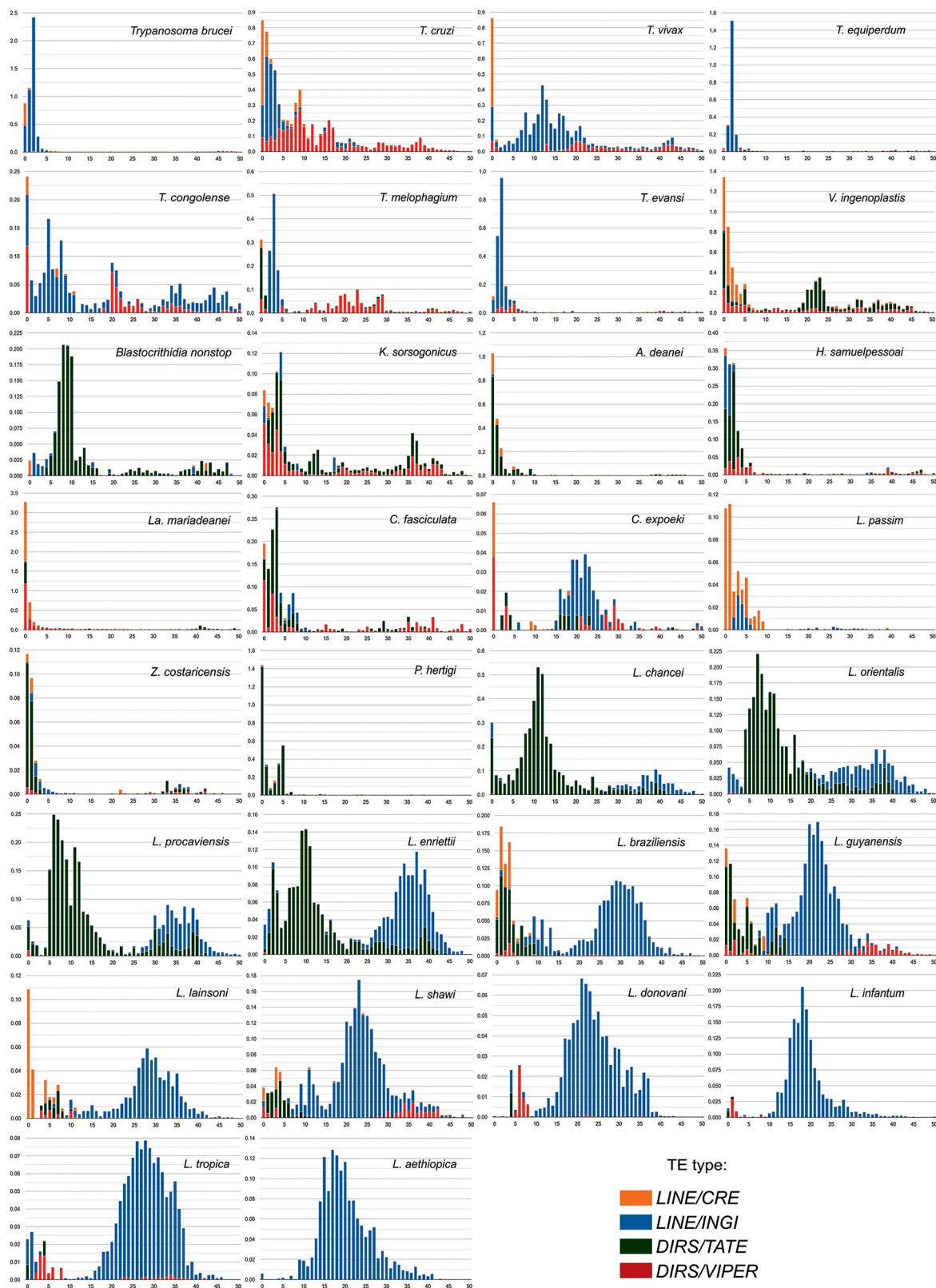
## Discussion

### Trypanosomatid TE database 1.0: A curated high-quality resource for future research

In recent years, the volume of trypanosomatid genomic data has increased dramatically. While the study of TEs is crucial for understanding the evolutionary dynamics and functionality of genomes, they are often analysed superficially. Our work significantly extended the understanding of the trypanosomatid mobilome through a comparative analysis encompassing 57 genomes that included new non-model trypanosomatids. It revealed TE abundance, diversity, activity and evolution.

To ensure the reliability of outcomes, we executed 2 pipelines based on repetitiveness, RepeatModeler and dnaPipeTE. The latter prevents underestimation of the TE proportion, a common problem for genome assemblies based on short reads as they can be fragmented and present collapsed contigs (Alkan et al., 2011; Peona et al., 2018; Shahid and Slotkin, 2020).

The TE prediction approach employed in this work also addressed common issues associated with similarity-based methods. Specifically, the TE databases often harbour a low representation of curated models for non-model microorganisms and, consequently, rely solely on similarity-based methods, which can limit the detection of TEs in these underrepresented or novel



**Figure 6.** TE age distribution in trypanosomatid genomes based on K2P divergence analysis. The y-axis displays the percentage of the genome (abundance) for different clades of TEs, and the x-axis shows the Kimura substitution level ( $k$ -value from 0 to 50) of copies with their respective consensus sequences. Likewise, a low degree of divergence indicates recent activity ( $<5\%$ ), whereas higher divergence scores suggest that the copies derive from older transposition events.



species (Storer et al., 2021). On the other hand, some TEs might be missed by repetitiveness-based methods due to their insufficient repetitiveness and, therefore, we complemented our analyses with BLAST-based searches.

Confirming sequences as genuine TEs and classifying them correctly is crucial but complex, as programs like RepeatModeler can detect a wide range of repetitive DNAs and sometimes misclassify sequences (Almutairi et al., 2021a). Distinguishing between the total repetitive content of the genome (RE) and the TE content is critical in any discussion concerning transposons. Moreover, since trypanosomatid genomes exhibit low diversity (Bringaud et al., 2008), new TEs identified by such programs require careful scrutiny. In this study, manual curation ensured accurate TE classification. Additionally, while we attempted to classify unknown sequences, most remained unclassified. Although some of them might be genuine TEs, the absence of typical protein domains and TE structures suggests they are unlikely to be conventional elements, making their characterization particularly challenging.

In this work, after a concerted effort to overcome the challenges in characterizing trypanosomatid TEs, we ultimately confirmed 214 consensus TE families across 37 species. These curated sequences are now part of the Trypanosomatid TE Database 1.0, presenting a valuable resource for future studies.

### RE and TE content in trypanosomatid genomes

Based on the first genomic sequences of trypanosomatids, significant variation in the proportion of the repetitive genomic content has been noted, with the vast majority of it allocated to the multi-genic families (Ivens et al., 2005; Pita et al., 2019). Considering the total proportion of REs, our data agree with previous works for *C. fasciculata* (Albanaz et al., 2023), *T. cruzi* (Pita et al., 2019), and *L. major* and *L. martiniquensis* (Almutairi et al., 2021a; Albanaz et al., 2023). A higher proportion of REs than previously reported was found in *H. samuelpessoai* and *T. brucei* compared to previous reports (Berriman et al., 2005; Pita et al., 2019; Albanaz et al., 2023). According to our results, the highest repetitive contents are reported in the genus *Trypanosoma*.

Earlier studies suggested that TEs comprise up to 5% of the genomic content in trypanosomatids (Bringaud et al., 2007) while later analyses elevated this number to 12% (Pita et al., 2019). A recent work estimated that 48% of the *T. cruzi* genome are TEs (Hoyos Sanchez et al., 2024). However, this estimate was obtained with raw RepeatModeler libraries that likely included other repetitive sequences, such as the large gene families or unknown sequences. In our study, we report a lower proportion for *T. cruzi*, which can be explained by the different TE prediction tools used and the manual curation of consensus sequences. The highest proportion of TEs in this work was ~7%, for *V. ingenoplastis*. Although this proportion is much smaller than what is found in mammals and plants, it is comparable to or slightly lower than what was found in some unicellular protists, such as *Entamoeba* sp. (5–8% (Pritham, 2009)), apicomplexans (up to 5.4% (Rodríguez and Makalowski, 2022)) and the amoebozoan *Dictyostelium discoideum* (~10%; Glöckner et al., 2001). Multiple interconnected factors are known to contribute to TE abundance, including transposition activity, historical accumulation of TEs, silencing mechanisms, competition between TEs, occasional positive selection for beneficial insertions (Betancourt et al., 2024) and the strength of purifying selection acting against the TEs, which itself is affected by population size (Lynch and Conery, 2003; Betancourt et al., 2024).

### Phylogeny of trypanosomatid species

Our reconstructed phylogenetic tree was broadly consistent with previous reports based on various nuclear or kDNA markers (Kaufer et al., 2019; Maslov et al., 2019; Kostygov et al., 2021). The topology of the clade encompassing *T. cruzi* and *T. grayi* (with a low support in our work) is in line with the phylogenomics analysis of Kelly et al. (2014) that relied on a supermatrix of 959 single-copy nuclear genes. Moreover, in this work, we report the first multilocus-based analysis of the subfamily Herpetomonadinae. Similarly to the previous inferences based on 18S ribosomal RNA/gGAPDH sequences (Yurchenko et al., 2016), it placed *H. samuelpessoai* and *L. mariadeaneai* as sister taxa.

### Diversity and evolution of TEs in Trypanosomatidae

To date, DNA transposons have not been reliably identified or characterized in trypanosomatid genomes; however, Merlin DNA transposons were recently discovered in the genomes of trypanosomatid-related *B. saltans* and *Perkinsella* sp. (Lopes et al., 2021). In this work, we found no evidence of any DNA transposons across the 57 analysed nuclear genomes. Thus, we could not confirm the presence of helitrons detected in the *L. martiniquensis* genome (Almutairi et al., 2021a) or other DNA transposons reported in *T. cruzi* (Hakim et al., 2024; Hoyos Sanchez et al., 2024), suggesting that they were false positives. Our findings corroborate the idea that trypanosomatid genomes are devoid of class II TEs. Considering that DNA transposons are present in the last common ancestor of kinetoplastids (Lopes et al., 2021), we propose that these elements were eliminated very early from the genomes of trypanosomatids. Our data support the hypothesis that the mobilome of a trypanosomatid ancestor was of low diversity and limited to the *INGI*, *CRE*, *TATE* and *VIPER*. This aligns with a recent report showing low TE diversity in the *Paradiplonema papillatum* genome, a species from a sister group to kinetoplastids (Valach et al., 2023).

Considering the distribution of the 4 TE clades, we conclude that these elements were generally effective in colonizing most trypanosomatid species. Nevertheless, some TEs were either ablated or degenerated during evolution. Events of complete TE loss occurred mostly independently either in single species (for example, *VIPER* and *TATE* in *Lotmaria passim* and *Leptomonas seymouri*) or in the ancestors of the species group (for example, *CRE* in *Mundinia*). The persistence of all 4 elements in several species across the family Trypanosomatidae suggests that either these elements were more active in certain species or their remnants were retained. This can be well exemplified by the *INGI* elements, which, despite the absence of active copies, were retained in the genomes as non-autonomous counterparts, likely due to their role in the control of gene expression (Bringaud et al., 2007; Heras et al., 2007).

Considering the mode of TE transmission, there have been no documented cases of horizontal transfer (HT) in trypanosomatids, despite the fact that this mechanism was reported for several trypanosomatid genes, such as catalase or proline racemase (Oppendoes and Michels, 2007; Caballero et al., 2015; Chmelová et al., 2021). In our dataset, we found no clear evidence of HT, such as the presence of unexpected elements that might represent acquisitions from non-trypanosomatid sources. Such acquisitions could be anticipated, given the ecological associations of these parasites with diverse host and vector species. HT of TEs has been observed in other systems with close ecological interactions; for example, in *Rhodnius prolixus*, a major vector of *T. cruzi*, multiple TE families



have been identified as nearly identical to those found in its mammalian hosts (Gilbert et al., 2010; Schaack et al., 2010). While the possibility of TE lateral transfer among trypanosomatid species remains an area for further investigation, the apparent absence of this process raises important questions about the mechanisms that constrain TE transmission to a vertical model.

Despite manual curation of the presented database, we recognize that trypanosomatid TEs and their posited activities must be further characterized using wet-lab methods. Furthermore, future research could benefit from detailed phylogenetic analyses of the 4 trypanosomatid TE superfamilies to better understand their evolutionary trajectories and roles within the trypanosomatid genomes.

In this work, we provided the first in-depth report on the abundance and distribution of TEs in a large array of trypanosomatid species. We generated valuable custom TE libraries that can be employed by the trypanosomatid research community to improve the annotation of the mobilome for new genome assembly. Our comparative study provided new perspectives to understanding the events of gains and loss in the TE repertoire, elucidating the dynamics in trypanosomatid genome architecture.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/S0031182025100231>.

**Acknowledgements.** The authors would like to thank Valentina Peona for all comments and suggestions to improve our work.

**Author contributions.** PT and JA: Conceptualization, project administration, methodology, manuscript drafting. PT and AL: wrote the first draft of the manuscript and visualization. PT and AL: Data analysis. AL: Results validation. PT, AC, and EC: Analyses. AL, VY, MT, JS, and JA: Reviewing and editing. MT, MK, JS, and AL: Resources and supervision. All authors read and approved the final manuscript.

**Financial support.** POTV and this study were supported by the Brazilian National Council of Scientific and Technological Development – CNPq (140430/2021-0) and Brazilian Federal Agency for Support and Evaluation of Graduate Education – CAPES from the government of Brazil. VY was supported by the European Union Operational Program 'Just Transition' (LERCO CZ.10.03.01/00/22\_003/0000003).

**Competing interests.** The authors declare there are no conflicts of interest.

**Ethical standards.** Not applicable.

## References

- Aaron R, Quinlan I and Hall M (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842. doi:[10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033)
- Aksoy S (1991) Site-specific retrotransposons of the trypanosomatid protozoa. *Parasitology Today* **7**(10), 281–285. doi:[10.1016/0169-4758\(91\)90097-8](https://doi.org/10.1016/0169-4758(91)90097-8)
- Albanaz ATS, Carrington M, Frolov AO, Ganyukova AI, Gerasimov ES, Kostygov AY and Butenko A (2023) Shining the spotlight on the neglected: New high-quality genome assemblies as a gateway to understanding the evolution of Trypanosomatidae. *BMC Genomics* **24**(1), 471. doi:[10.1186/s12864-023-09591-z](https://doi.org/10.1186/s12864-023-09591-z)
- Alkan C, Sajjadian S and Eichler EE (2011) Limitations of next-generation genome sequence assembly. *Nature Methods* **8**(1), 61–65. doi:[10.1038/nmeth.1527](https://doi.org/10.1038/nmeth.1527)
- Almutairi H, Urbaniak MD, Bates MD, Jariyapan N, Al-Salem WS, Dillon RJ, Bates PA and Gatherer D (2021a) Chromosome-scale assembly of the complete genome sequence of *Leishmania* (Mundinia) *martiniquensis*, Isolate LSCM1, Strain LV760. *Microbiology Resource Announcements* **10**(24), e0005821. doi:[10.1128/MRA.00058-21](https://doi.org/10.1128/MRA.00058-21)
- Andrews S (2010) FastQC: A Quality Control tool for High Throughput Sequence Data. Babraham Bioinformatics, Cambridge, UK. Distributed by the author: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed 15 March 2024).
- Bailey-Bechet M, Haudry A and Lerat E (2014) “One code to find them all”: A perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**, 13. doi:[10.1186/1759-8753-5-13](https://doi.org/10.1186/1759-8753-5-13)
- Bao W, Kojima KK and Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11. doi:[10.1186/s13100-015-0041-9](https://doi.org/10.1186/s13100-015-0041-9)
- Barrón MG, Fiston-Lavier AS, Petrov DA and González J (2014) Population genomics of transposable elements in *Drosophila*. *Annual Review Genetics* **48**, 561–581. doi:[10.1146/annurev-genet-120213-092359](https://doi.org/10.1146/annurev-genet-120213-092359)
- Benson G (1999) Tandem repeats finder: A program to analyse DNA sequences. *Nucleic Acids Research* **27**(2), 573–580. doi:[10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573)
- Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renaud H, Bartholomeu DC, Lennard NJ and El-Sayed NM (2005) The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**(5733), 416–422. doi:[10.1126/science.1112642](https://doi.org/10.1126/science.1112642)
- Betancourt AJ, Wei KH, Huang Y and Lee YCG (2024) Causes and consequences of varying transposable element activity: An evolutionary perspective. *Annual Review of Genomics & Human Genetics* **25**(1), 1–25. doi:[10.1146/annurev-genom-120822-105708](https://doi.org/10.1146/annurev-genom-120822-105708)
- Biscotti MA, Olmo E and Heslop-Harrison JS (2015) Repetitive DNA in eukaryotic genomes. *Chromosome Research* **23**, 415–420. doi:[10.1007/s10577-015-9499-z](https://doi.org/10.1007/s10577-015-9499-z)
- Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL and Feschotte C (2018) Ten things you should know about transposable elements. *Genome Biology* **19**(1), 199. doi:[10.1186/s13059-018-1577-z](https://doi.org/10.1186/s13059-018-1577-z)
- Bringaud F, Berriman M and Hertz-Fowler C (2009) Trypanosomatid genomes contain several subfamilies of ingi-related retrotransposons. *Eukaryotic Cell* **8**(10), 1532–1542. doi:[10.1128/EC.00183-09](https://doi.org/10.1128/EC.00183-09)
- Bringaud F, García-Pérez JL, Heras SR, Ghedin E, El-Sayed NM, Andersson B, Baltz T and Lopez MC (2002) Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. *Molecular and Biochemistry Parasitology* **124**(1–2), 73–78. doi:[10.1016/s0166-6851\(02\)00167-6](https://doi.org/10.1016/s0166-6851(02)00167-6)
- Bringaud F, Ghedin E, El-Sayed NM and Papadopoulos B (2008) Role of transposable elements in trypanosomatids. *Microbes and Infection* **10**(6), 575–581. doi:[10.1016/j.micinf.2008.02.009](https://doi.org/10.1016/j.micinf.2008.02.009)
- Bringaud F, Müller M, Cerqueira GC, Smith M, Rochette A, El-Sayed NM, Papadopoulos B and Ghedin E (2007) Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. *PLoS Pathogens* **3**(9), 1291–1307. doi:[10.1371/journal.ppat.0030136](https://doi.org/10.1371/journal.ppat.0030136)
- Bushnell B, Rood J and Singer E (2017) BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One* **12**(10), e0185056. doi:[10.1371/journal.pone.0185056](https://doi.org/10.1371/journal.pone.0185056)
- Caballero ZC, Costa-Martins AG, Ferreira RC, Alves JMP, Serrano MG, Camargo EP, Buck GA, Minoprio P and G Teixeira MM (2015) Phylogenetic and syntenic data support a single horizontal transference to a *Trypanosoma* ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences. *Parasite and Vectors* **12**(8), 222. doi:[10.1186/s13071-015-0829-y](https://doi.org/10.1186/s13071-015-0829-y)
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421. doi:[10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421)
- Camacho E, González-de la Fuente S, Solana JC, Rastrojo A, Carrasco-Ramiro F, Requen AJM and Aguado B (2021) Gene Annotation and transcriptome delineation on a de novo genome assembly for the reference *Leishmania major* Friedlin Strain. *Genes (Basel)* **12**(9), 1359. doi:[10.3390/genes12091359](https://doi.org/10.3390/genes12091359)
- Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T (2009) TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**(15), 1972–1973. doi:[10.1093/bioinformatics/btp348](https://doi.org/10.1093/bioinformatics/btp348)

- Chan PP, Lin BY, Mak AJ and Lowe TM (2021) tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**(16), 9077–9096. doi:10.1093/nar/gkab688
- Chen S, Zhou Y, Chen Y and Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890. doi:10.1093/bioinformatics/bty560
- Chmelová L, Bianchi C, Albanaz ATS, Režnarová J, Wheeler R, Kostygov AY, Kraeva N and Yurchenko V (2021) Comparative analysis of three trypanosomatid catalases of different origin. *Antioxidants (Basel)* **11**(1), 46. doi:10.3390/antiox11010046
- Cuyper B, Dumetz F, Meysman P, Laukens K, De Muylder G, Dujardin JC and Domagalska MA (2020) The absence of C-5 DNA methylation in *Leishmania* donovani allows DNA enrichment from complex samples. *Microorganisms* **8**(8), 1252. doi:10.3390/microorganisms8081252
- Ellinghaus D, Kurtz S and Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18. doi:10.1186/1471-2105-9-18
- Emms DM and Kelly S (2019) OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**(1), 238. doi:10.1186/s13059-019-1832-y
- Espinosa OA, Serrano MG, Camargo EP, Teixeira MMG and Shaw JJ (2018) An appraisal of the taxonomy and nomenclature of trypanosomatids presently classified as *Leishmania* and *Endotrypanum*. *Parasitology* **145**(4), 430–442. doi:10.1017/S0031182016002092
- Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist* **125**, 1–15. doi:10.1086/284325
- Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5**(4), 103–107. doi:10.1016/0168-9525(89)90039-5
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C and Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceeding of the National Academy of Sciences of the United States America* **117**(17), 9451–9457. doi:10.1073/pnas.1921046117
- Fujiwara H (2015) Site-specific non-LTR retrotransposons. *Microbiology Spectrum* **3**(2), MDNA3–0001–2014. doi:10.1128/microbiolspec.MDNA3-0001-2014
- Gilbert C, Peccoud J and Cordaux R (2021) Transposable elements and the evolution of insects. *Annual Review Entomology* **66**, 355–372. doi:10.1146/annurev-ento-070720-074650
- Gilbert C, Schaack S, Pace JK, Brindley PJ and Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* **464**(7293), 1347–1350. doi:10.1038/nature08939
- Glöckner G, Szafranski K, Winckler T, Dingermaier T, Quail MA, Cox E, Eichinger L, Noegel AA and Rosenthal A (2001) The complex repeats of *Dictyostelium discoideum*. *Genome Research* **11**(4), 585–594. doi:10.1101/gr.162201
- Goubert C (2023) Assembly-Free detection and quantification of transposable elements with dnaPipeTE. *Methods Molecular Biology* **2607**, 25–43. doi:10.1007/978-1-0716-2883-6\_2
- Goubert C, Craig RJ, Bilal AF, Peona V, Vogan AA and Protasio AV (2022) A beginner's guide to manual curation of transposable elements. *Mobile DNA* **13**(1), 7. doi:10.1186/s13100-021-00259-7
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N and Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**(7), 644–652. doi:10.1038/nbt.1883
- Hakim JMC, Gutierrez Guarnizo SA, Málaga Machaca E, Gilman RH and Mugnier MR (2024) Whole-genome assembly of a hybrid *Trypanosoma cruzi* strain assembled with Nanopore sequencing alone. *G3 Bethesda* **14**(6), jkae076. doi:10.1093/g3journal/jkae076
- Heras SR, López MC, Olivares M and Thomas MC (2007) The L1Tc non-LTR retrotransposon of *Trypanosoma cruzi* contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. *Nucleic Acids Research* **35**(7), 2199–2214. doi:10.1093/nar/gkl1137
- Hoyos Sanchez MC, Ospina Zapata HS, Suarez BD, Ospina C, Barbosa HJ, Carranza Martinez JC, Vallejo GA, Urrea Montes D and Duitama J (2024) A phased genome assembly of a Colombian *Trypanosoma cruzi* TcI strain and the evolution of gene families. *Scientific Reports* **14**(1), 2054. doi:10.1038/s41598-024-52449-x
- Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E and Myler PJ (2005) The genome of the kinetoplastid parasite. *Leishmania major* *Science* **309**(5733), 436–442. doi:10.1126/science.1112680
- Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A and Berriman M (2016) Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Current Biology* **26**(2), 161–172. doi:10.1016/j.cub.2015.11.055
- Jackson AP, Quail MA and Berriman M (2008) Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa). *BMC Genomics* **9**, 594. doi:10.1186/1471-2164-9-594
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jermini LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**(6), 587–589. doi:10.1038/nmeth.4285
- Katoh K and Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology Evolution* **30**(4), 772–780. doi:10.1093/molbev/mst010
- Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR and O'Sullivan C (2022) The sequence read archive: A decade more of explosive growth. *Nucleic Acids Research* **50**(D1), D387–D390. doi:10.1093/nar/gkab1053
- Kaufer A, Ellis J, Stark D and Barratt J (2017) The evolution of trypanosomatid taxonomy. *Parasites and Vectors* **10**(1), 287. doi:10.1186/s13071-017-2204-7
- Kaufer A, Stark D and Ellis J (2019) Evolutionary insight into the Trypanosomatidae using alignment-free phylogenomics of the kinetoplast. *Pathogens* **8**(3), 157. doi:10.3390/pathogens8030157
- Kazanian HH (2004) Mobile elements: Drivers of genome evolution. *Science* **303**(5664), 1626–1632. doi:10.1126/science.1089670
- Kelly S, Ivens A, Manna PT, Gibson W and Field MC (2014) A draft genome for the African crocodilian trypanosome *Trypanosoma grayi*. *Scientific Data* **1**, 140024. doi:10.1038/sdata.2014.24
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120. doi:10.1007/BF01731581
- Ko BJ, Lee C, Kim J, Rhie A, Yoo DA, Howe K, Wood J, Cho S, Brown S, Formenti G, Jarvis ED and Kim H (2022) Widespread false gene gains caused duplication errors in genome assemblies. *Genome Biology* **23**(1), 205. doi:10.1186/s13059-022-02764-1
- Kohany O, Gentles AJ, Hankus L and Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474. doi:10.1186/1471-2105-7-474
- Kojima KK (2020) Structural and sequence diversity of eukaryotic transposable elements. *Genes and Genetic Systems* **94**(6), 233–252. doi:10.1266/ggs.18-00024
- Kostygov AY, Albanaz ATS, Butenko A, Gerasimov ES, Lukeš J and Yurchenko V (2024) Phylogenetic framework to explore trait evolution in Trypanosomatidae. *Trends Parasitology* **40**(2), 96–99. doi:10.1016/j.pt.2023.11.009
- Kostygov AY, Frolov AO, Malysheva MN, Ganyukova AI, Chistyakova IV, Tashyreva D, Tesařová M, Spodareva VV, Režnarová J, Macedo DH, Butenko A, d'Ávila-Levy CM, Lukeš J and Yurchenko V (2020) *Vickermania* gen. nov., trypanosomatids that use two joined flagella to resist midgut peristaltic flow within the fly host. *BMC Biology* **18**(1), 187. doi:10.1186/s12915-020-00916-y
- Kostygov AY, Karnkowska A, Votýpka J, Tashyreva D, Maciszewski K, Yurchenko V and Lukeš J (2021) Euglenozoa: Taxonomy, diversity and ecology, symbioses and viruses. *Open Biology* **11**(3), 200407. doi:10.1098/rsob.200407
- Kostygov AY and Yurchenko V (2017) Revised classification of the subfamily Leishmaniinae (Trypanosomatidae). *Folia Parasitologica (Praha)* **64**, 020. doi:10.14411/fp.2017.020

- Kozlov AM, Darriba D, Flouri T, Morel B and Stamatakis A (2019) RAXML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21), 4453–4455. doi:10.1093/bioinformatics/btz305
- Kück P and Longo GC (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontier in Zoology* 11(1), 81. doi:10.1186/s12983-014-0081-x
- Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV and Zdobnov EM (2023) OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* 51(D1), D445–D451. doi:10.1093/nar/gkac998
- Larsson A (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30(22), 3276–3278. doi:10.1093/bioinformatics/btu531
- Le Rouzic A and Capy P (2005) The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. *Genetics* 169(2), 1033–1043. doi:10.1534/genetics.104.031211
- Letunic I and Bork P (2024) Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* 52(W1), W78–W82. doi:10.1093/nar/gkac268
- Li J, Miao B, Wang S, Dong W, Xu H, Si C, Wang W, Duan S, Lou J, Bao Z, Zeng H, Yang Z, Cheng W, Zhao F, Zeng J, Liu XS, Wu R, Shen Y, Chen Z, Chen S, Wang M and Consortium H (2022) Hiplot: A comprehensive and easy-to-use web service for boosting publication-ready biomedical data visualization. *Briefings in Bioinformatics* 23(4), bbac261. doi:10.1093/bib/bbac261
- Llanes A, Restrepo CM, Del Vecchio G, Anguizola FJ and Leonart R (2015) The genome of *Leishmania panamensis*: Insights into genomics of the L. (Viannia) subgenus. *Scientific reports* 5, 8550. doi:10.1038/srep08550
- Lopes ALK, Kriegová E, Lukeš J, Krieger MA and Ludwig A (2021) Distribution of Merlin in eukaryotes and first report of DNA transposons in kinetoplastid protists. *PLoS One* 16(5), e0251133. doi:10.1371/journal.pone.0251133
- Lorenzi HA, Robledo G and Levin MJ (2006) The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-encoding retrotransposons. *Molecular Biochemistry Parasitology* 145(2), 184–194. doi:10.1016/j.molbiopara.2005.10.002
- Lynch M and Conery JS (2003) The origins of genome complexity. *Science* 302(5649), 1401–1404. doi:10.1126/science.1089370
- Macías F, Afonso-Lehmann R, López MC, Gómez I and Thomas MC (2018) Biology of Trypanosoma cruzi retrotransposons: From an enzymatic to a structural point of view. *Current Genomics* 9(2), 110–118. doi:10.2174/1389202918666170815150738
- Marçais G and Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6), 764–770. doi:10.1093/bioinformatics/btr011
- Maslov DA, Oppendoes FR, Kostygov AY, Hashimi H, Lukeš J and Yurchenko V (2019) Recent advances in trypanosomatid research: Genome organization, expression, metabolism, taxonomy and evolution. *Parasitology* 146(1), 1–27. doi:10.1017/S0031182018000951
- McClintock B (1984) The significance of responses of the genome to challenge. *Science* 226(4676), 792–801. doi:10.1126/science.15739260
- Militello KT, Wang P, Jayakar SK, Pietrasik RL, Dupont CD, Dodd K, King AM and Valenti PR (2008) African trypanosomes contain 5-methylcytosine in nuclear DNA. *Eukaryotic Cell* 7(11), 2012–2016. doi:10.1128/EC.00198-08
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37(5), 1530–1534. doi:10.1093/molbev/msaa015
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD and Bateman A (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research* 49(D1), D412–D419. doi:10.1093/nar/gkaa913
- Oppendoes FR and Michels PA (2007) Horizontal gene transfer in trypanosomatids. *Trends in Parasitology* 23(10), 470–476. doi:10.1016/j.pt.2007.08.002
- Ou S and Jiang N (2018) LTR\_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant physiology* 176(2), 1410–1422. doi:10.1104/pp.17.01310
- Paradis E and Schliep K (2019) Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi:10.1093/bioinformatics/bty633
- Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA and Berriman M (2007) Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature Genetics* 39(7), 839–847. doi:10.1038/ng2053
- Peona V, Weissensteiner MH and Suh A (2018) How complete are “complete” genome assemblies?—An avian perspective. *Molecular Ecology Resources* 18(6), 1188–1195. doi:10.1111/1755-0998.12933
- Pita S, Díaz-Viraqué F, Iraola G and Robello C (2019) The tritryps comparative repeatome: Insights on repetitive element evolution in trypanosomatid pathogens. *Genome Biology Evolution* 11(2), 546–551. doi:10.1093/gbe/evz017
- Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. *Journal of Heredity* 100(5), 648–655. doi:10.1093/jhered/esp065
- Qian J, Xue H, Ou S, Storer J, Fürtauer L, Wildermuth MC, Kusch S and Panstruga R (2024) TETrimmer: A novel tool to automate the manual curation of transposable elements. *bioRxiv* doi:10.1101/2024.06.27.600963
- Quinlan AR (2014) BEDTools: The Swiss-army tool for genome feature analysis. *Current protocols in Bioinformatics* 47(1), 1–34. doi:10.1002/0471250953.bi1112s47
- Ranallo-Benavidez TR, Jaron KS and Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communication* 11(1), 1432. doi:10.1038/s41467-020-14998-3
- Revell LJ (2024) Phytools 2.0: An updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12, e16505. doi:10.7717/peerj.16505
- Ribeiro YC, Robe LJ, Veluza DS, Dos Santos CMB, Lopes ALK, Krieger MA and Ludwig A (2019) Study of VIPER and TATE in kinetoplastids and the evolution of tyrosine recombinase retrotransposons. *Mobile DNA* 10, 34. doi:10.1186/s13100-019-0175-2
- Rice P, Longden I and Bleasby A (2000) EMBOS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6), 276–277. doi:10.1016/S0168-9525(00)00204-2
- Rodríguez M and Makalowski W (2022) Mobilome of apicomplexa parasites. *Genes (Basel)* 13(5), 887. doi:10.3390/genes13050887
- Sánchez-Luque FJ, López MC, Macías F, Alonso C and Thomas MC (2011) Identification of an hepatitis delta virus-like ribozyme at the mRNA 5′-end of the L1Tc retrotransposon from *Trypanosoma cruzi*. *Nucleic Acids Research* 39(18), 8065–8077. doi:10.1093/nar/gkr478
- Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST and Karsch-Mizrachi I (2021) GenBank. *Nucleic Acids Research* 49(D1), D92–D96. doi:10.1093/nar/gkaa1023
- Schaack S, Gilbert C and Feschotte C (2010) Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology and Evolution* 25(9), 534–546. doi:10.1016/j.tree.2010.06.001
- Seppey M, Manni M and Zdobnov EM (2019) BUSCO: Assessing genome assembly and annotation completeness. *Methods Molecular Biology* 1962, 227–245. doi:10.1007/978-1-4939-9173-0\_14
- Shahid S and Slotkin RK (2020) The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology* 54, 49–56. doi:10.1016/j.pbi.2019.12.012
- Shanmugasundram A, Starns D, Böhme U, Amos B, Wilkinson PA, Harb OS, Warrenfeltz S, Kissinger JC, McDowell MA, Roos DS, Crouch K and Jones AR (2023) TriTrypDB: An integrated functional genomics resource for kinetoplastida. *PLoS Neglected Tropical Diseases* 17(1), e0011058. doi:10.1371/journal.pntd.0011058
- Smith M, Bringaud F and Papadopoulos B (2009) Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. *BMC Genomics* 10, 240. doi:10.1186/1471-2164-10-240



- Storer J, Hubley R, Rosen J, Wheeler TJ and Smit AF (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* 12(1), 2. doi:10.1186/s13100-020-00230-y
- Suzek BE, Wang Y, Huang H, McGarvey PB and Wu CH (2015) UniProt Consortium. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31(6), 926–932. doi:10.1093/bioinformatics/btu739
- Tarailo-Graovac M and Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* Chapter 4, 4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s25.
- Teng SC, Wang SX and Gabriel A (1995) A new non-LTR retrotransposon provides evidence for multiple distinct site-specific elements in *Crithidia fasciculata* minixen arrays. *Nucleic Acids Research* 23(15), 2929–2936. doi:10.1093/nar/23.15.2929
- Tullume-Vergara PO, Caicedo KYO, Tantalean JFC, Serrano MG, Buck GA, Teixeira MMG, Shaw JJ and Alves JMP (2023) Genomes of *Endotrypanum monerogeei* from Panama and *Zelonina costaricensis* from Brazil: Expansion of multigene families in Leishmaniinae parasites that are close relatives of *Leishmania* spp. *Pathogens* 12(12), 1409. doi:10.3390/pathogens12121409
- Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, Nenarokova A, Prokopchuk G, Batstone T, Lapébie P, Lemogo L, Sarrasin M, Stretenowich P, Tripathi P, Yazaki E, Nara T, Henrissat B, Lang BF, Gray MW, Williams TA, Lukeš J and Burger G (2023) Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes. *BMC Biology* 21(1), 99. doi:10.1186/s12915-023-01563-9
- Vazquez M, Ben-Dov C, Lorenzi H, Moore T, Schijman A and Levin MJ (2000) The short interspersed repetitive element of *Trypanosoma cruzi*, SIRE, is part of VIPER, an unusual retroelement related to long terminal repeat retrotransposons. *Proceedings of the National Academic of Sciences of the United State American* 97(5), 2128–2133. doi:10.1073/pnas.050578397
- Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ and Marchler-Bauer A (2023) The conserved domain database in 2023. *Nucleic Acids Research* 51(D1), D384–D388. doi:10.1093/nar/gkac1096
- Weizhong L and Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13), 1658–1659. doi:10.1093/bioinformatics/btl158
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B and Schulman AH (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8(12), 973–982. doi:10.1038/nrg2165
- Wickham H (2016) Ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. Retrieved from <https://ggplot2.tidyverse.org> (accessed 1 July 2024).
- Woo TH, Hong TH, Kim SS, Chung WH, Kang HJ, Kim CB and Seo JM (2007) Repeatome: A database for repeat element comparative analysis in human and chimpanzee. *Genome Informatics* 5, 179–187.
- Xu Z and Wang H (2007) LTR\_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* 35(Web Server issue), W265–8. doi:10.1093/nar/gkm286
- Yurchenko V, Kostygov A, Havlová J, Grybchuk-Ieremenko A, Ševčíková T, Lukeš J, Ševčík J and Votýpka J (2016) Diversity of Trypanosomatids in cockroaches and the description of *Herpetomonas tarakana* sp. n. *The Journal of Eukaryotic Microbiology* 63(2), 198–209. doi:10.1111/jeu.12268