BMC Biology

RESEARCH

Open Access

Comprehensive analysis of the Kinetoplastea intron landscape reveals a novel intron-containing gene and the first exclusively *trans*-splicing eukaryote

Alexei Yu. Kostygov[1,2†], Karolína Skýpalová[1†], Natalia Kraeva[1], Elora Kalita[1], Cameron McLeod[3], Vyacheslav Yurchenko[1], Mark C. Field[3,4], Julius Lukeš[4,5] and Anzhelika Butenko[1,4,5*]

Abstract

Background In trypanosomatids, a group of unicellular eukaryotes that includes numerous important human parasites, *cis*-splicing has been previously reported for only two genes: a poly(A) polymerase and an RNA helicase. Conversely, *trans*-splicing, which involves the attachment of a spliced leader sequence, is observed for nearly every protein-coding transcript. So far, our understanding of splicing in this protistan group has stemmed from the analysis of only a few medically relevant species. In this study, we used an extensive dataset encompassing all described trypanosomatid genera to investigate the distribution of intron-containing genes and the evolution of splice sites.

Results We identified a new conserved intron-containing gene encoding an RNA-binding protein that is universally present in Kinetoplastea. We show that *Perkinsela* sp., a kinetoplastid endosymbiont of Amoebozoa, represents the first eukaryote completely devoid of *cis*-splicing, yet still preserving *trans*-splicing. We also provided evidence for reverse transcriptase-mediated intron loss in Kinetoplastea, extensive conservation of 5′ splice sites, and the presence of non-coding RNAs within a subset of retained trypanosomatid introns.

Conclusions All three intron-containing genes identified in Kinetoplastea encode RNA-interacting proteins, with a potential to fine-tune the expression of multiple genes, thus challenging the perception of *cis*-splicing in these protists as a mere evolutionary relic. We suggest that there is a selective pressure to retain *cis*-splicing in trypanosomatids and that this is likely associated with overall control of mRNA processing. Our study provides new insights into the evolution of introns and, consequently, the regulation of gene expression in eukaryotes.

Keywords Kinetoplastea, Trypanosomatidae, Introns, Splicing, Poly(A) polymerase, RNA helicase, RNA-binding protein

[†]Alexei Yu. Kostygov and Karolína Skýpalová contributed equally to this work.

*Correspondence:
Anzhelika Butenko
anzhelika.butenko@paru.cas.cz
Full list of author information is available at the end of the article

Background

Introns are noncoding gene elements that are excised from nascent RNA during maturation by various mechanisms depending on the intron type [1]. Spliceosomal introns are excised from precursor mRNA (pre-mRNA) by the spliceosome, a large ribonucleoprotein complex comprising five small nuclear RNAs (snRNAs) and, typically, over a hundred proteins [2]. In the predominant *cis*-splicing reaction, fragments of the same RNA molecule are ligated together after intron removal, while *trans*-splicing, a less common modality, results in the fusion of RNA fragments transcribed from non-contiguous loci [3]. Introns excised by the spliceosome contain conserved splicing signals, usually GT/AG dinucleotides at their 5′ and 3′ ends, a branchpoint/lariat sequence, and a polypyrimidine tract [2]. From here on, we will refer to introns removed during *cis*-splicing as *cis*-spliceosomal introns [4].

Cis-spliceosomal introns can serve multiple functions, such as enhancing protein diversity through alternative splicing, hosting diverse non-coding RNAs, regulating gene expression, RNA stability and targeting, and influencing mRNA transport and chromatin assembly [5]. Eukaryotic lineages differ drastically in their intron content and length, with the genomes of vertebrates, plants, and some fungi being intron-rich, whereas those of some intracellular endosymbionts are highly reduced and devoid of introns [6]. For instance, human genes contain an average of eight introns. In contrast, the nucleomorph (a highly reduced remnant of algal endosymbiont nucleus) of the cryptophyte *Hemiselmis andersenii* and nuclei of some microsporidia (obligate intracellular parasites related to fungi) apparently lack *cis*-spliceosomal introns [7–9]. However, complete loss of introns is extremely rare and has never been documented in eukaryotes endowed with *trans*-splicing [10].

The phylum Euglenozoa, a group of unicellular eukaryotes (protists) including kinetoplastids, diplonemids, euglenids, and symbiontids [11], serves as an excellent example of a lineage that unites organisms with drastic differences in intron content. Introns are widespread in diplonemid and euglenid genomes, whose genes contain conventional (with typical GT/AG borders, excised by the spliceosome) and non-conventional (with atypical dinucleotides at the borders, stable secondary structure and an unknown excision mechanism) introns, or a combination thereof [12–15]. Conversely, in the genomes of trypanosomatids (Kinetoplastea: Trypanosomatidae), *cis*-spliceosomal introns are reportedly confined to only two genes: poly(A) polymerase (*PAP1*) and a putative RNA helicase (*DBP2B*) [16]. Genome-wide analysis, employing transcriptomic data from different life stages of *Trypanosoma brucei*, strongly supported the notion that no additional intervening sequences are present in the thoroughly investigated genome of this parasite [17]. PAP1 is involved in the polyadenylation of noncoding RNAs, particularly small nucleolar RNAs (snoRNAs) and long noncoding RNAs (lncRNAs) [18]. Silencing of *PAP1* in *T. brucei* reduces the levels of mature snoRNAs [18], which are important translation regulators in this parasite [19]. Eleven nucleotides (nt) at the 5′ splice site of the *PAP1* intron sequence are highly conserved between *T. brucei* and *T. cruzi* and point mutations within this region abolish splicing [20]. However, it is not clear whether such extended conserved sequence at the 5′ splice site is characteristic for other Kinetoplastea. To the best of our knowledge, no data on the functional significance of the trypanosomatid RNA helicase and its intron were obtained since the first identification [21, 22]. However, the well-studied yeast homolog of this protein (DEAD-box ATPase DBP2) is involved in RNA metabolism, including transcription [23, 24], nonsense-mediated mRNA decay, and ribosomal RNA processing [25]. In addition, it has been reported to disrupt RNA G-quadruplex structures [26] and to interact with RNA viruses [27].

Recently, the genome sequences and transcriptomic data of virtually all formally described trypanosomatid genera have become available [28, 29]. Nevertheless, studies of intervening sequences have typically focused on a limited set of the kinetoplastid genomes belonging to medically and veterinary-relevant pathogens of the genera *Leishmania* and *Trypanosoma* [17, 20, 21]. Therefore, a systematic analysis of introns in Kinetoplastea is timely. Here, we investigate the distribution of intron-containing genes and the respective introns using a phylogenetically balanced dataset that includes numerous recently sequenced genomes [28], analyze conserved intron features, alternative transcript isoforms, and evaluate a potential role for reverse transcriptase (RT)-mediated intron loss in the evolution of these iconic protists.

Results

A novel intron-containing gene in Kinetoplastea encodes an RNA-binding protein

Genome-wide analysis of *cis*-splicing in Kinetoplastea has been limited to a handful of trypanosomatids that infect vertebrates, such as *T. brucei* and *Leishmania major* [17, 21]. Genomic and transcriptomic data for nearly all formally recognized trypanosomatid genera are now available [29]. This significantly extended dataset was subjected to two independent intron identification strategies: (i) based on the presence of conserved sequences at the 5′ intron border and (ii) analysis of split reads mapping [20, 30].

Using the first approach, we found an intron in the gene encoding a putative RNA-binding protein (*T. brucei* accession Tb927.8.6440) previously designated as RBP20 [31] and identified its orthologues in all analyzed genomes (Additional file 1: Table S1). The RBP20 protein orthologues contain a glutamine-rich and two glycine/arginine-rich low complexity regions of variable length and an RNA recognition motif (Fig. 1A). The length of RBP20 in Kinetoplastea ranges from 216 amino acids (aa) in the parabodonid *Trypanoplasma borreli* to around 350 aa in a monoxenous relative of *Leishmania*, the trypanosomatid *Zelonia costaricensis*. The protein length variation primarily stems from divergence within low complexity regions, with the *T. borreli* sequence lacking the glutamine-rich region being the shortest (Fig. 1A). The position of the *RBP20* intron is highly conserved across kinetoplastids (Fig. 1A). Moreover, all identified introns have canonical GT/AG borders (Fig. 1B and C), with splicing supported by the presence of split reads in transcriptomic data mapped on the genome assembly and by a sharp decrease in read coverage at intron ends (Fig. 1B).

The analysis of split read mappings identified 237 splice junction candidates for *Bodo saltans* and 373 for *Perkinsela* sp. However, manual inspection did not reveal any valid intron predictions for *Perkinsela* sp. and confirmed the three known intron-containing genes in *B. saltans*. Erroneously predicted splice junctions were associated with repetitive regions and the genome assembly artefacts.

Taken together, this evidence suggests that we have identified a third gene in Kinetoplastea harboring a *cis*-spliceosomal intron. Similar to *PAP1* and *DBP2B*, the *RBP20* gene encodes a protein likely involved in RNA metabolism, as suggested by the presence of RNA recognition motif, although its exact function remains undefined.

Distribution of *cis*-spliceosomal introns in Kinetoplastea

The information on intron-containing genes and corresponding intervening sequences in Kinetoplastea is fragmentary, with a sampling bias toward medically and veterinary important species. To achieve a comprehensive understanding with deeper implications for kinetoplastid genome evolution, we conducted a systematic analysis of intron distribution across a broad set of genomes.

We identified homologs of the three intron-containing genes in all analyzed kinetoplastid genomes, with rare exceptions that are likely related to the lower quality of genome assemblies (Fig. 2; Additional file 1: Table S1). Nevertheless, the evolutionary trajectories of each of these three genes were distinct. In the case of poly(A) polymerase, there were two basal clades, of which one was represented mostly by intron-containing genes (*PAP1*), while the other was formed exclusively by intron-less homologs (*PAP2*) (Additional file 2: Fig. S1A). Each of these two clades contained a single kinetoplastid gene and two genes from *Paradiplonema papillatum*. The phylogeny of the RNA helicase showed a gene duplication in Metakinetoplastia (Additional file 2: Fig. S1B). While *Perkinsela* sp. has a single intron-less *DBP2A* gene, Trypanosomatidae and their closest free-living relative, *B. saltans*, have both the intron-containing and intron-less homologs (Additional file 2: Fig. S1B). This duplication was independent of that in *P. papillatum*. Regrettably, due to the low quality of the assembly for *T. borreli,* it was not possible to ascertain the *DBP2B* gene(s) configuration in this species. The *RBP20* phylogeny was the simplest: all analyzed taxa possess a single gene, with the exception of *B. saltans*, which possesses an additional intron-less paralogue (Additional file 2: Fig. S1C), a duplication that likely occurred in Eubodonida.

In Kinetoplastea, the intron within *PAP1* and *RBP20* genes was independently lost at least twice, with both losses occurring outside the family Trypanosomatidae: in the prokinetoplastid *Perkinsela* sp. and in the

(See figure on next page.)

Fig. 1 A novel intron-containing gene encodes an RNA-binding protein. **A** Alignment of RBP20 proteins from several kinetoplastids. The intron position is indicated by a magenta arrow. Sequence conservation level for each alignment position is depicted with height and color of the underlying bars (short brown and high yellow bars indicate low and high conservation, respectively). The asterisk next to the *T. borreli* name indicates that its protein is encoded by an intron-less gene. The borders of RNA recognition motif and low complexity regions are defined according to the sequence from *B. saltans*. G, glycine; Q, glutamine; R, arginine. Amino acids are colored according to the Clustal scheme implemented in Jalview. **B** A snapshot of RNA-seq data mapping onto the genome assembly at the *RBP20* locus in various kinetoplastids. Exons are depicted by blue bars (drawn to scale for each species), and the upper track shows RNA-seq read coverage (with the maximum value indicated on *Y* axis). A splice junction track (SJ) features arcs connecting alignment blocks from split reads, highlighted in red and blue for the forward and the reverse strands, respectively. The height and thickness of the arcs are proportional to the coverage depth. For each gene, the genomic coordinates, strand (in brackets), and length are shown. Transcriptomic data from public databases were used to generate coverage plots (see Additional File 1: Table S1 for details). **C** RNA-seq data derived from non-polysomal RNA library mapped onto the *RBP20* gene of *Leishmania donovani*. Brown bar indicates the borders of BLAST hit to ncRNA in RNAcentral database (see Additional file 1: Table S4)

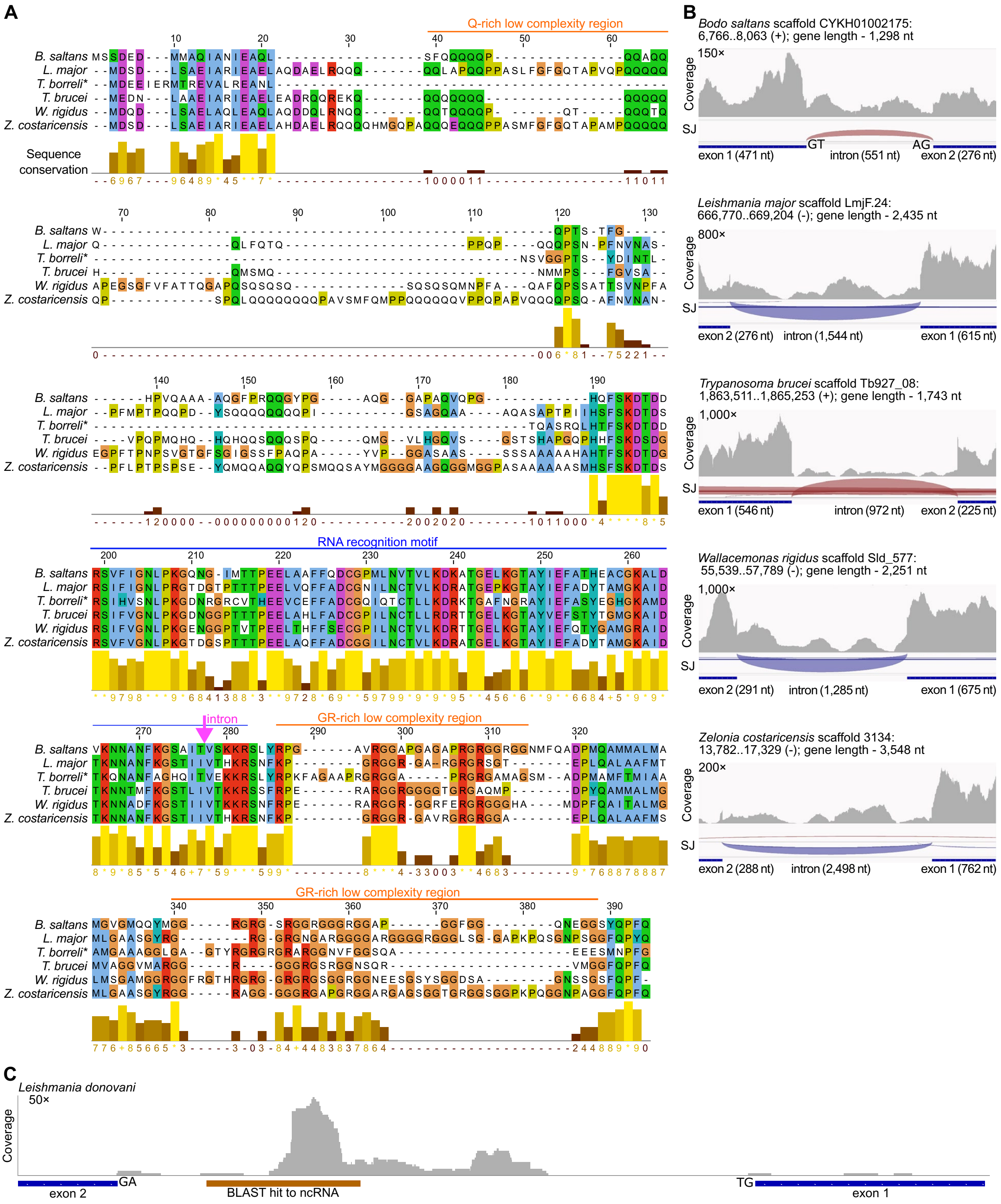

Fig. 1 (See legend on previous page.)

parabodonid *T. borreli* (Fig. 2; Additional file 1: Table S1). The intron in the *DBP2B* gene is absent from *Perkinsela* sp. and was lost at least five times independently within the family Trypanosomatidae: specifically, in the common ancestor of *Angomonas* sp., in *Porcisia hertigi*, *Sergeia podlipaevi*, *Strigomonas oncopelti*, and *Trypanosoma*

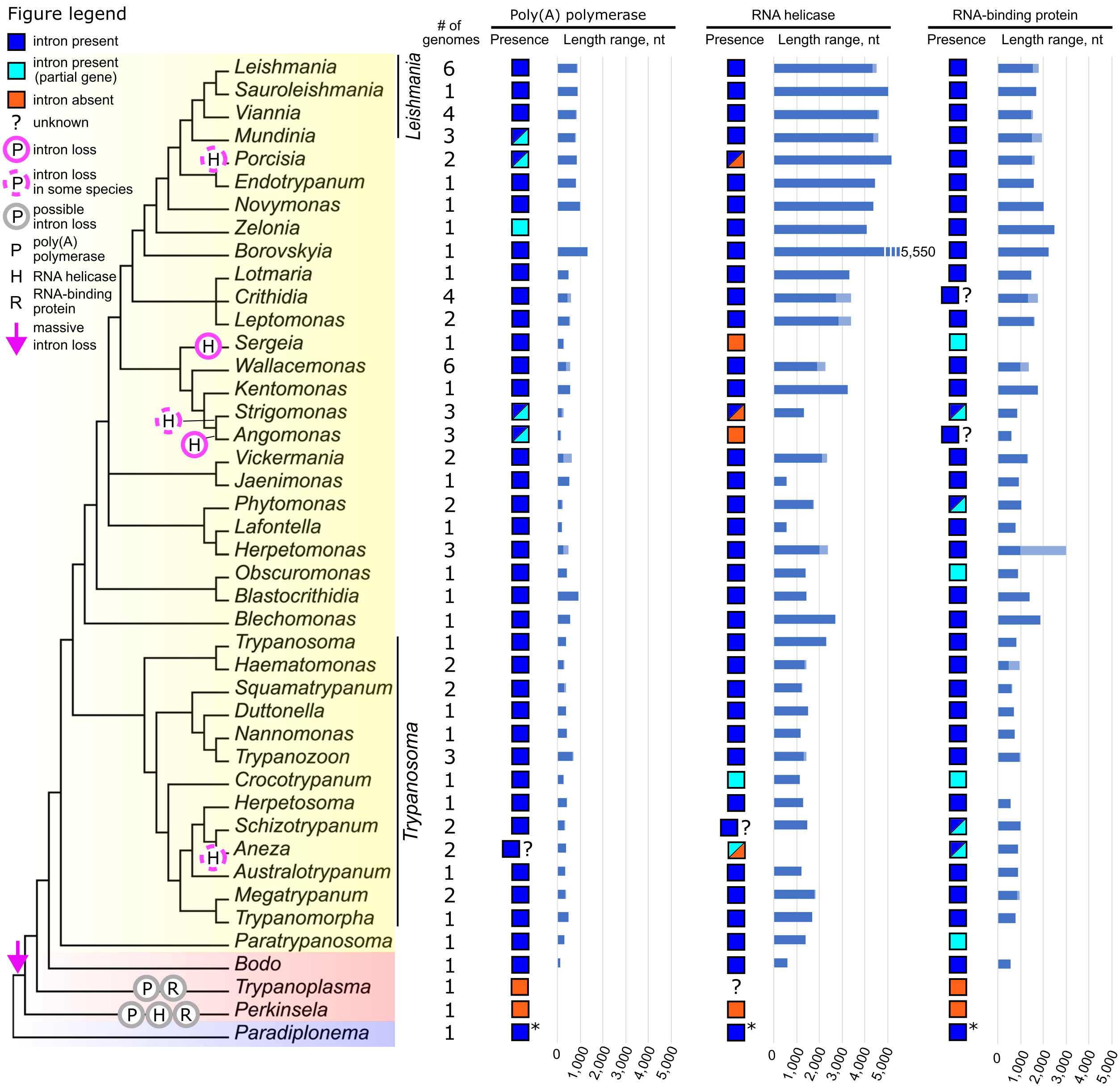

Fig. 2 Intron distribution in the genes encoding poly(A) polymerase, RNA helicase, and RNA-binding protein in Euglenozoa. Intron presence/absence is depicted using rectangles: blue, present; cyan, gene is partial with an intron fragment; orange, intron absent; question mark, no data (unavailability of genome assembly, absence of the gene, or the gene fragment too short to infer intron presence). For each clade/species intron length is displayed as bar plot (showing minimal and maximal intron length values for the group). Only the length of introns for complete genes is shown. Dashed bar is used for the intron out of scale of the graph. Intron losses on the tree are depicted by magenta circles (dashed outline in case of losses only in some species of the group: H, RNA helicase gene; P, poly(A) polymerase; R, RNA-binding protein). Magenta arrow indicates massive intron loss in the kinetoplastid common ancestor. The cladogram is based on [29] and [32]. Asterisk, although introns are present in *P. papillatum* homologs, we excluded them from the analysis as their positions differ from those in Kinetoplastea

conorhini (Fig. 2; Additional file 1: Table S1). Thus, the only kinetoplastid devoid of *cis*-spliceosomal introns in our dataset is *Perkinsela* sp., an endosymbiont of an amoebozoan with a massively reduced genome [33]. An examination of a more contiguous *Perkinsela* assembly [34] with N_{50} of ~ 1 megabase (Mb) did not reveal any evidence of introns. The fish parasite *T. borreli* lacks introns in *PAP1* and *RBP20* genes, and we were unable to identify an orthologue of *DBP2B* in this species to investigate the presence of an intron (see above). Homologs of all three genes in the *P. papillatum* genome possess multiple introns, but their positions do not correspond to the positions of those in the kinetoplastid genes, leaving the ancestral intron configuration for these genes in Glycomonada unresolved.

The location of introns varies between *PAP1*, *DBP2B*, and *RBP20* in Kinetoplastea. In *PAP1*, about three-quarters of the protein, including most of the poly(A) polymerase domain, is encoded by the second exon (Additional file 3: Fig. S2). In contrast, the other two genes exhibit a different pattern—the intron is located between the first exon encoding a conserved domain and the second exon encoding low complexity regions (glycine and glycine/proline-rich in DBP2B and glycine/arginine-rich in RBP20) (Additional file 3: Fig. S2). Introns within these three genes exhibit uniform phase (i.e., position within/between codons) across different species. However, the phase differs among genes, with *RBP20*, *DBP2B*, and *PAP1* interrupted by the intervening sequences with phases 0, 1, and 2, respectively (Additional file 1: Table S1). The introns within *PAP1* orthologues in *B. saltans*, *A. deanei*, and *P. serpens*, as well as that in the *DBP2B* gene in *T. conorhini*, do not harbor stop codons (Additional file 1: Table S1). Among the analyzed kinetoplastid genes, the *PAP1* gene's intervening sequence is the shortest, with a median length of approximately 440 nucleotides (Fig. 2). Conversely, the RNA helicase intron is the longest, with a median length of around 3350 nt, and the RBP20 intron exhibits intermediate values, being approximately 1300 nt long (Fig. 2).

Our analysis indicates the absence of *cis*-spliceosomal introns in *Perkinsela* sp., whereas the eubodonid *B. saltans* and trypanosomatids invariably harbor *PAP1* and the newly identified *RBP20* introns. The *DBP2B* intervening sequence is less persistent, having been lost multiple times within the family Trypanosomatidae.

Alternative splicing inferred from transcriptomics

We utilized short-read transcriptomic data to validate the occurrence of splicing and analyzed publicly available long-read datasets to explore the diversity of transcript isoforms. Importantly, in all species with available short-read transcriptomic data in which we could identify full-length intron-containing genes, we detected split reads indicative of functional splicing, with a few exceptions (Additional file 1: Table S1). These exceptions include *PAP1* and *RBP20* genes in *Leishmania donovani*, *Trypanosoma avium*, *T. scelopori*, and the *PAP1* gene in *Blastocrithidia nonstop*, *Trypanosoma boissoni*, and *Vickermania* spp. In most of these examples, the transcriptome coverage for the respective loci is relatively low, and all mapped reads apparently either originate from pre-mRNA or indicate intron retention. It is not possible to distinguish between these two options using the short-read data only. However, considering that RNA-seq data for the majority of these species is derived from poly(A)-selected libraries, presumably enriched for mature mRNAs, the option of intron retention is more plausible.

Additionally, we examined several publicly available long-read RNA-seq datasets and determined that among these seven datasets, only one [35, 36] contained reads of sufficient quality and length, with at least two reads covering the genes of interest. This dataset contains transcriptome data for the procyclic and bloodstream stages of *T. brucei* [36]. We only considered reads representing mature transcripts, i.e., containing the spliced leader (SL) sequence at the 5' end. Unfortunately, we were unable to identify SL-containing reads covering the *DBP2B* gene. For both *PAP1* and *RBP20* genes, we identified reads corresponding to two different transcript isoforms: SL sequence attached to two exons joined together with the intron spliced out (SL-5′ untranslated region (UTR)-exon1-exon2), and SL sequence attached to the 5′ end of the second exon (SL-exon2). In the case of the *PAP1*, one out of two and two out of six reads corresponded to the SL-exon2 isoform in the bloodstream and procyclic stages, respectively. For *RBP20*, one out of 12 and two out of 12 reads corresponded to the SL-exon2 isoform in the bloodstream and procyclic stages, respectively. Unfortunately, the sample size was insufficient for a quantitative analysis.

We used complementary DNA (cDNA) from *L. mexicana* to investigate whether the transcript isoform diversity observed in *T. brucei* is present in a broader phylogenetic context. Indeed, from polymerase chain reaction (PCR) products (refer to the Additional file 1: Table S2 for the primers used), we assembled transcripts representing SL-5′ UTR-exon1-exon2 and SL-exon2 sequences for each of the three intron-containing genes (Additional file 4: Fig. S3A-G). Additionally, among *PAP1* transcripts originating from *L. mexicana* amastigotes, we identified an isoform (SL-intronic_fragment-exon2) that originated by usage of an intra-intronic *trans*-splicing site. These data suggest an interplay between *cis*- and *trans*-splicing in Trypanosomatidae.

Conserved features of *cis*-spliceosomal introns and their potential roles as splicing signals

A previous analysis of 5′ splice sites in *T. brucei* and *T. cruzi* revealed striking sequence conservation, which was hypothesized to be important for the recognition by the spliceosomal U1 snRNA [20]. However, it was unclear whether the extended 5′ splicing sites are conserved in other kinetoplastids, and additional intronic features remained even more elusive. To clarify this, we undertook a systematic analysis of intervening sequences in diverse kinetoplastids.

Sequence conservation was observed within several nucleotides at the 5′ end of the introns in all three genes and was most pronounced in the *PAP1* gene, in which a 10-nt-long motif was predominant. In the other genes, only six to eight nucleotides showed high levels of conservation (Additional file 5: Fig. S4A). Conservation around the 3′ splice site was in general lower. Besides the invariant AG dinucleotide site itself, the *PAP1* and *RBP20* genes lack recognizable motifs at the 3′ end of the intron, while the −3 position in the *DBP2B* intron was most frequently occupied by C (Additional file 5: Fig. S4A). Sequence conservation was also observed in the exon sequences adjacent to the introns (Additional file 5: Fig. S4A).

The search for intraspecific sequence motifs at the 5′ intron end of the three intron-containing genes revealed a highly conserved pentanucleotide GTATG, with a shorter motif (GTAT) present only in *Kentomonas sorsogonicus* (Additional file 5: Fig. S4B). In several other species, this motif was longer, being either contiguous (up to nine nucleotides in *Wallacemonas rigidus*) or disrupted by 1–3 variable nucleotides (up to 17 out of 18 in *B. saltans*). In most cases, downstream of the conserved motif(s), isolated or grouped invariant pyrimidines were observed, with specific nucleotides (C or T) distinct for each species. In addition, in all genes, there was an invariable A at the position −4 (i.e., four exonic nucleotides upstream of the splice site), and in the majority of species, this was accompanied by T at the position −2 (Additional file 5: Fig. S4B).

The search for intronic gapped motifs identified additional features possibly associated with the regulation of splicing in these genes (Additional file 6: Fig. S5). Only the polypyrimidine tracts were universally present, however, without a specific position with respect to the introns' termini, similarly to other identified gapped motifs (Additional file 6: Fig. S5; Additional file 1: Table S3). This and the fact that kinetoplastids do not have a conserved recognition motif for branchpoints did not allow identification of the latter [37].

In addition to various conserved features important for correct splicing, introns in some eukaryotes contain non-coding RNA (ncRNA) species, which can indicate an evolutionary pressure for intron retention [38]. We investigated whether this is the case for kinetoplastid introns through RNAcentral database [39]. For the introns of some Leishmaniinae, trypanosomes, Herpetomonadinae, *Wallacemonas* spp., and several others, we indeed obtained hits to lncRNAs (Additional file 1: Table S4). However, the results of these searches could be impacted by low-complexity sequences within introns. Even for the hit with the highest identity (76.2%) and target coverage (86.3%), a microRNA of *Trypanosoma vivax* similar to a subsequence of *RBP20* intron in *Trypanosoma caninum*, the analysis with LocARNA-P showed only a moderate reliability (mostly below 0.5). Therefore, we chose an independent confirmation. Available transcriptome sequencing data enriched for ncRNAs are very scarce for trypanosomatids, but we found RNA-seq data enriched for the fraction of non-polysomal RNAs bound to proteins derived from *L. donovani* promastigotes [40, 41]. RNA helicase and *RBP20* introns of this species show significant hits (*e*-value below e^{-15}) to ncRNAs (Additional file 1: Table S4), and the intron of *RBP20* gene showed ~ 35 × coverage by transcriptome data at the region overlapping the respective BLAST hit, while the exon sequences were not covered by reads (Fig. 1C) strongly suggesting that a ncRNA is indeed expressed from this intronic locus.

The repertoire and characteristics of U1 snRNAs in Kinetoplastea

Conservation of several 5′-region nucleotides of the intron is consistent with their binding to the 5′ terminus of the U1 snRNA [42]. Therefore, we searched for U1 snRNAs and used these sequences to predict pairing between 5′ ends and those of the introns considering previously proposed interactions [43] and de novo predictions.

U1 snRNAs were identified in all kinetoplastids except *T. borreli*, for which only a low-quality assembly is available. In most cases, there was a single variant (Additional file 7: Fig. S6). However, in *Lafontella* sp. and *Herpetomonas muscarum*, two highly similar sequences were found, while for *Paratrypanosoma confusum*, three divergent candidates were detected (Additional file 7: Fig. S6). Guided by the previously published predictions for model trypanosomatids, we inferred secondary structures of U1 snRNAs (Fig. 3A). The majority of trypanosomatid U1 snRNAs contained typical features conserved in reference species, specifically two hairpins and three additional conserved motifs (Fig. 3A). The first conserved motif required for interaction with the intron's 5′ end (consensus: AACUCACCUGCA) consisted of 10–11 nt at the 5′ end. The second motif (consensus: CAUCAA GAAA) required for U1 small nuclear ribonucleoprotein 70 kDa (U1-70 K) binding was present within the loop of the first hairpin. The third motif (consensus: AAAUUU UGA), representing a site for Smith (Sm)-like proteins binding, was situated between the two hairpin loops.

Despite some sequence variation, these motifs were recognizable and similar to those from very distantly related eukaryotes, such as *Homo sapiens* [44, 45]. The sequences of the hairpin stems differed both in composition (but typically were G/C rich) and in size: they could include from 3 to 16 base pairs, but most frequently

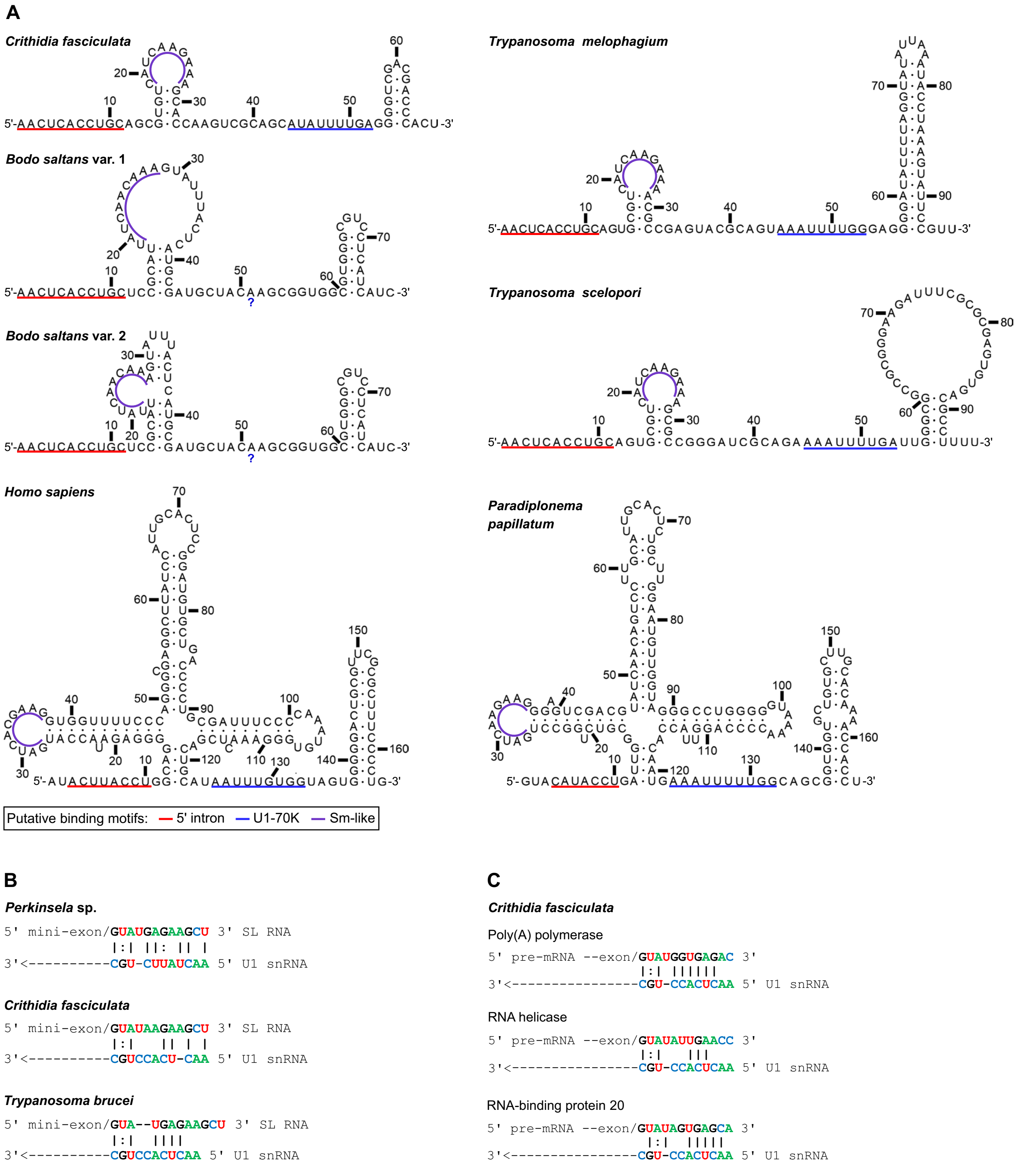
A
Crithidia fasciculata
5'-AACUCACCUGCAGCG·CCAAGUCGCAGCAUAUUUUGAGG·CACU-3'
Bodo saltans var. 1
5'-AACUCACCUGCUCC·GAUGCUACAAGCGGUGGC·CAUC-3'
Bodo saltans var. 2
5'-AACUCACCUGCUCC·GAUGCUACAAGCGGUGGC·CAUC-3'
Homo sapiens
5'-AUACUUACCUGG·CAUAAUUUGUGGUAGUG·UG-3'
Trypanosoma melophagium
5'-AACUCACCUGCAGUG·CCGAGUACGCAGUAAAUUUUGGGAGG·CGUU-3'
Trypanosoma scelopori
5'-AACUCACCUGCAGUG·CCGGGAUCGCAGAAAAUUUUGAUUG·UUUU-3'
Paradiplonema papillatum
5'-GUACAUACCUGA·UGAAAUUUUUGGCAGCG·CU-3'
Putative binding motifs: 5' intron U1-70K Sm-like
B
Perkinsela sp.
5' mini-exon/GUAUGAGAAGCU 3' SL RNA
|:| ||: || |
3'<----------CGU-CUUAUCAA 5' U1 snRNA
Crithidia fasciculata
5' mini-exon/GUAUAAGAAGCU 3' SL RNA
|:| || | |
3'<----------CGUCCACU-CAA 5' U1 snRNA
Trypanosoma brucei
5' mini-exon/GUA--UGAGAAGCU 3' SL RNA
|:| ||||
3'<----------CGUCCACUCAA 5' U1 snRNA
C
Crithidia fasciculata
Poly(A) polymerase
5' pre-mRNA --exon/GUAUGGUGAGAC 3'
|:| ||||||
3'<----------------CGU-CCACUCAA 5' U1 snRNA
RNA helicase
5' pre-mRNA --exon/GUAUAUUGAACC 3'
|:| |||
3'<----------------CGU-CCACUCAA 5' U1 snRNA
RNA-binding protein 20
5' pre-mRNA --exon/GUAUAGUGAGCA 3'
|:| |||||
3'<----------------CGU-CCACUCAA 5' U1 snRNA

Fig. 3 U1 snRNAs in Kinetoplastea. **A** Comparison of secondary structures of snRNAs in selected kinetoplastid species and two reference organisms—*Homo sapiens* (URS00006CA71A_9606, structure retrieved from RNAcentral) and *Paradiplonema papillatum* (JAPJBO010001534: 29,166..29330). For *B. saltans*, two reconstructions are shown: with and without the hairpin. **B** Potential interactions between 5′ termini of U1 snRNA and the intron of SL RNA in *Perkinsela* sp. and two trypanosomatid species. **C** Potential interactions between 5′ termini of U1 snRNA and the introns of the three intron-containing genes in *Crithidia fasciculata*
]

4 (Fig. 3A; Additional file 7: Fig. S6A). In some species, the predicted stem of the first hairpin included the first nucleotide of the U1-70 K binding site, although this nucleotide differed from the consensus (U *vs* C). Besides this, the size of the first loop was invariant among trypanosomatids, while the second loop ranged in length from three to 28 nt (Fig. 3A; Additional file 7: Fig. S6A). *B. saltans* differed from trypanosomatids in two respects: the U1 lacks an Sm-like binding motif and possesses a longer first loop, where an additional small hairpin is predicted (Fig. 3A; Additional file 7: Fig. S6A). Of note, the sequence for this hairpin was the only region with low similarity to the Sm-like motif. In contrast to kinetoplastids, their close relative *P. papillatum* has a cloverleaf-like secondary structure of U1 snRNA, which is typical for opisthokonts, with only slight differences from that of *H. sapiens* (Fig. 3A).

The alignments of U1 snRNA sequences and structures were satisfactory for all species except *P. confusum* with three low-confidence candidate sequences and a single one for *Perkinsela* sp., which was expected to lack the U1 snRNA gene, considering that it lacks *cis*-spliceosomal introns (Additional file 7: Fig. S6B). We attempted to reassemble available genomic reads for *P. confusum*, but the resulting assembly revealed the same sequences that lacked a recognizable Sm-binding site, with the deviations within conserved intron recognition and/or U1-70 K-binding motifs (Additional file 7: Fig. S6B). Predicted U1 snRNA for *Perkinsela* sp. featured eight nucleotides at the 5′ terminus matching the consensus and a recognizable U1-70 K-binding motif; however, the latter was not associated with a characteristic hairpin. Moreover, the only predictable hairpin in this area significantly overlapped with the U1-70 K binding motif. This sequence also did not contain identifiable Sm-like binding site (Additional file 7: Fig. S6B). We assessed the potential interaction of the 5’ termini of SL gene intron and U1 snRNA in *Perkinsela* sp. and concluded that it may include up to 9 base pairs, i.e., more than in reference trypanosomatid species *C. fasciculata* and *T. brucei*, in which no more than 7 bases are involved (Fig. 3B). This is within the range observed for the interactions between spliceosomal introns and U1 snRNA in trypanosomatids (Fig. 3C), suggesting an involvement of U1 snRNA into *trans*-splicing in this species. Interestingly, while the data on the potential role of U1 snRNA of Kinetoplastea in SL RNA processing is obscure, a recent in vivo study unambiguously demonstrated the interaction of U1 small nuclear ribonucleoprotein (snRNP) with SL RNA in *T. brucei* [46]. The search for U1 snRNA-associated proteins in *Perkinsela* sp. identified only U1-70 K and a divergent U1 small nuclear ribonucleoprotein A (U1A)-like sequence, while the kinetoplastid-specific U1 small nuclear ribonucleoprotein 24 kDa (U1-24 K) and the broadly conserved U1 small nuclear ribonucleoprotein C (U1-C) were absent. Notably, *Perkinsela* sp. and *T. borreli* were the only analyzed euglenozoan species lacking the U1-C protein.

It was challenging to select a single variant of pairing between 5′ splice site and U1 snRNA, as sometimes we obtained up to three competing hypothetical interactions. In most cases, the inferred interaction involved the GUA–CGU pairing as in [43] and often followed by a 1–4-nt-long bulge in one of the RNA molecules (Fig. 3B; Additional file 8). The number of base pairs varied between species and genes. In *B. saltans* characterized by the longest conserved 5′ intron sequence, the distal nucleotides of the latter showed complementarity to the 5′ end of the snRNA, thereby providing more total paired bases than predicted for *T. brucei* (Additional file 8). The strongest interaction was inferred for *DBP2B* RNA helicase gene of *Z. costaricensis*, which has only a 6-nt-long conserved motif.

Thus, the U1 snRNAs in kinetoplastids are simplified as compared to opisthokonts and are generally conserved in terms of structure and functional motifs. A very divergent form has been retained in *Perkinsela* sp. despite the absence of *cis*-spliceosomal introns, probably due to its involvement in *trans*-splicing that remains present in this highly derived organism [33]. The observed variation in interaction strength between the introns of the three genes and U1 snRNAs may determine differences in their splicing efficiency and, consequently, expression levels.

A mechanism for intron loss in Kinetoplastea

The genomes of euglenozoan relatives of Kinetoplastea, *E. gracilis* and *P. papillatum* contain numerous introns [14, 15]. This implies that the genome of the euglenozoan common ancestor was significantly more intron-rich than that of extant kinetoplastids and suggests a substantial intron loss in the kinetoplastid common ancestor (Fig. 2). Additionally, there were subsequent sporadic intron losses within Kinetoplastea, resulting in variation in intron content even between closely related species (Fig. 2). Several mechanisms of intron loss have been proposed to date, including RT-mediated loss, simple deletion, and exonization [47]. RT-mediated loss is one of the best-characterized mechanisms, leaving discernible footprints in the genome: (i) “exact” intron removal without affecting the adjacent exon sequence; (ii) preferential occurrence at the 3′ end of the gene; (iii) bias towards losing adjacent introns; and (iv) a syntenic location for intron-less genes compared to their intron-containing orthologues in closely related species [47, 48].

To assess the potential for RT-mediated intron loss in kinetoplastid evolution, we searched for footprints of this mechanism in the available genomes. We used pairs of closely related species differing in intron content and assessed whether the removal of an intervening sequence was precise or it has also impacted an exon. For this, we aligned protein sequences corresponding to the intron-containing and intron-less genes in those species' pairs, under the assumption that the intron position was conserved among them (Fig. 4A). Our analysis revealed that intron loss was exact in all cases, and the coding sequences adjacent to the putative ancestral intron were unaffected, as evidenced by their conservation at the amino acid level (Fig. 4A). This observation holds true not only for closely related species within the same genus, such as *Porcisia deanei* and *P. hertigi* or *Strigomonas culicis* and *S. oncopelti*, but also extends to distantly related *B. saltans* and *Perkinsela* sp. Furthermore, intron loss appears to be precise for all three analyzed genes—*DBP2B*, *PAP1*, and *RBP20*. However, for the latter two genes, our evidence is derived from the comparisons between only two species, *B. saltans* and *Perkinsela* sp. (Fig. 4A).

The RT-mediated mechanism assumes that the intron-less cDNA is inserted into the genome via homologous recombination, and the insertion likely occurs at the intron-containing version of the gene locus, i.e., preserving synteny. As a result, we expect that the intron-less genes will be syntenic with intron-containing orthologues in the closely related species.

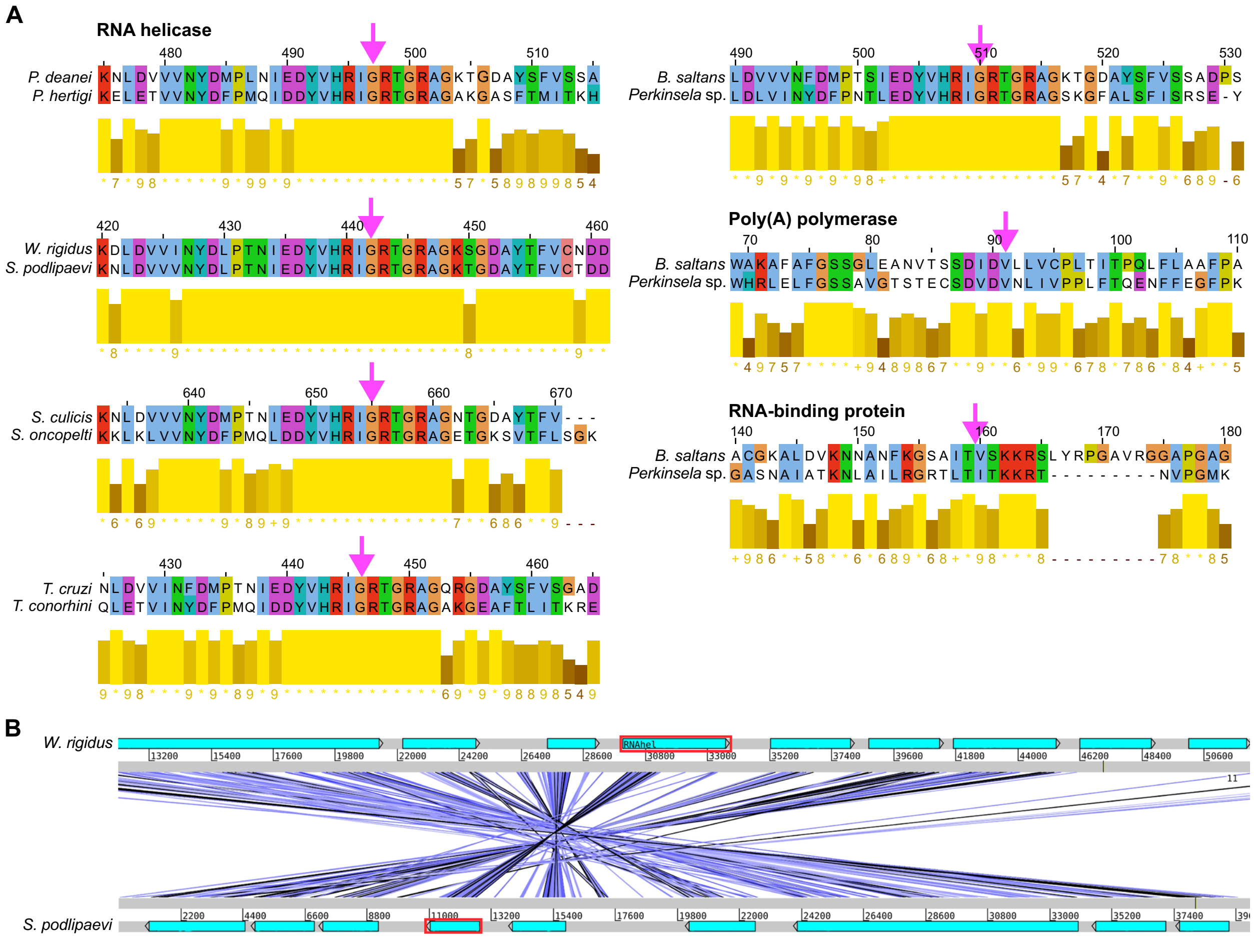
A
RNA helicase
P. deanei KNLDVVVNYDMPLNIEDYVHRIGRTGRAGKTGDAYSFVSSA
P. hertigi KELETVVNYDFPMQIDDYVHRIGRTGRAGAKGASFTMITKH
W. rigidus KDLDVVINYDLPTNIEDYVHRIGRTGRAGKSGDAYTFVCNDD
S. podlipaevi KNLDVVVNYDLPTNIEDYVHRIGRTGRAGKTGDAYTFVCTDD
S. culicis KNLDVVVNYDMPTNIEDYVHRIGRTGRAGNTGDAYTFV---
S. oncopelti KKLKLVVNYDFPMQLDDYVHRIGRTGRAGETGKSVTFLSGK
T. cruzi NLDVVINFDMPTNIEDYVHRIGRTGRAGQRGDAYSFVSGAD
T. conorhini QLETVINYDFPMQIDDYVHRIGRTGRAGAKGEAFTLITKRE
B. saltans LDVVVNFDMPTSIEDYVHRIGRTGRAGKTGDAYSFVSSADPS
Perkinsela sp. LDLVINYDFPNTLEDYVHRIGRTGRAGSKGFALSFISRSE-Y
Poly(A) polymerase
B. saltans WAKAFAFGSSGLEANVTSSDIDVLLVCPLTITPQLFLAAFPA
Perkinsela sp. WHRLELFGSSAVGTSTECSDVDVNLIVPPLFTQENFFEGFPK
RNA-binding protein
B. saltans ACGKALDVKNNANFKGSAITVSKKRSLYRPGAVRGGAPGAG
Perkinsela sp. GASNAIATKNLAILRGRTLTITKKRT---------NVPGMK
B
W. rigidus
RNAhel
S. podlipaevi

Fig. 4 Footprints of reverse transcriptase-mediated intron loss in Kinetoplastea genomes. **A** Protein alignment around the exon–exon junctions. The upper protein in the alignment is derived from the intron-containing gene, whereas the lower one is from the intron-less orthologue. The magenta arrow indicates the intron position, placed between amino acids in case of intron phase 0 (RBP20) and above the respective amino acid for introns of phases 1 and 2 (RNA helicase and PAP1, respectively). Sequence conservation level for each alignment position is depicted with bars (short brown and high yellow bars indicate low and high conservation, respectively). Amino acids are colored according to the Clustal scheme implemented in Jalview. **B** A snapshot of the genomes of *Wallacemonas rigidus* and *Sergeia podlipaevi* demonstrating synteny at the RNA helicase locus. Genes are shown as blue rectangles; the rectangle corresponding to RNA helicase is shown within red box. Blue lines link conserved regions within the two genomes. Numbers indicate gene position within scaffolds

Thus, we conducted an analysis of synteny around the genes of interest in the examined pairs of species (Fig. 4A), facilitated by the high levels of synteny generally observed in trypanosomatids [49]. The intron-less *DBP2B* RNA helicase in *Sergeia podlipaevi* is syntenic to the intron-containing orthologue from *Wallacemonas rigidus* (Fig. 4B), whereas orthologues in the genus *Trypanosoma* are non-syntenic. We were unable to perform synteny analysis for *Perkinsela* sp. and *B. saltans* due to significant divergence between their genomes. For *Porcisia* and *Strigomonas* spp., analysis was impeded by assembly fragmentation in the region of interest. Given that extant kinetoplastids possess only a single intron per gene and the respective introns are absent from the closest outgroup species, *P. papillatum*, we cannot determine whether intron loss is biased towards the 3′ end of the gene or if there is a tendency to lose adjacent introns [47].

Next, we investigated the hypothesis that the RT-mediated intron loss had contributed to the elimination of introns in the kinetoplastid common ancestor. To this end, we clustered predicted proteins from three trypanosomatids, as well as the diplonemid *P. papillatum* and the heterolobosean *Naegleria gruberi* (Additional file 9: Fig. S7), into orthologous groups (OGs). We considered an intron as ancestral if it was identified in two latter species and its position in their genomes did not differ by more than 10 codons. From a set of 643 OGs, each containing a single protein per species, we selected 107 groups containing highly conserved proteins (with an average protein identity of ≥ 50%). Among these, seven genes contained presumably "ancestral" introns lost in the kinetoplastid common ancestor (Additional file 9: Fig. S7). Almost all of these introns were phase 0 supporting their ancient origin [50]. In each case, intron loss in Kinetoplastea appeared to be precise, as the amino acid sequence in the region of the putative ancestral intron was highly conserved, with no observed gaps within five amino acids in the vicinity of the intron positions in *P. papillatum* and *N. gruberi* (Additional file 9: Fig. S7).

We identified RT domain-containing proteins in the genomes of nearly all analyzed euglenozoans, including those that have lost the RNA helicase intron. Finding over 2000 such proteins in the *P. papillatum* genome suggests that RT domains with a potential of mediating intron loss were present in the common ancestor of Glycomonada, likely indicative of the increased activity of type I transposable elements and/or retroviruses [48]. Taken together, these observations suggest that the RT-mediated intron loss played an important role in the intron elimination throughout kinetoplastid evolution.

Discussion

Kinetoplastid genomes exhibit a striking scarcity of introns compared to their closest euglenozoan relatives, the diplonemids and euglenids [10, 14, 15]. For example, in the diplonemid *P. papillatum* genome, ~ 40% of protein-coding genes contain introns, with an average of 1.6 introns per gene [14]. The *E. gracilis* genome harbors both conventional and non-conventional introns, and similar to *P. papillatum*, there is evidence for alternative splicing [15]. Conversely, within Kinetoplastea, only two gene transcripts have been previously reported to undergo *cis*-splicing, those encoding poly(A) polymerase PAP1 involved in polyadenylation of certain noncoding RNAs and an RNA helicase DBP2B, whose function remains unknown [20–22]. However, our view on intron distribution in kinetoplastids was derived from the medically and veterinary relevant yet phylogenetically restricted members of the genera *Leishmania* and *Trypanosoma* [17, 20], with scarce reports from their free-living relatives [51]. Another obstacle to a comprehensive understanding of intron evolution within this group arises from a common method for intron identification involving analysis of transcriptomic split reads, derived from mature mRNAs [30]. This method may not be ideal for repetitive genomes such as those of trypanosomatids, which contain large multigenic families [52], because it may lead to predictions of spurious introns and mask the presence of real ones. Indeed, mapping of sequenced cDNA fragments from the transcriptomes of the procyclic and bloodstream stages of *T. brucei* to the genome assembly did not reveal any new *cis*-spliceosomal introns [17].

Using a new approach for intron identification relying on the occurrence of a highly conserved sequence present at the 5′ end of introns (Additional file 5: Fig. S4), we identified a *bona fide* novel intron-containing gene across Kinetoplastea encoding the RBP20 protein (Fig. 1). The homologs of the three intron-containing genes *RBP20*, *PAP1*, and *DBP2B* are invariably present in kinetoplastids with a few exceptions likely due to incompleteness of the respective genome assemblies (Fig. 2; Additional file 1: Table S1). During their evolution, these genes experienced duplications, after which one copy lost the intron. Moreover, in some taxa, intron loss was observed also in the "main" copy, but this has not happened if the gene had not been duplicated. There are just two exceptions from this rule: the intracellular endosymbiont of an amoebozoan, the prokinetoplastid *Perkinsela*, whose genome is highly reduced due to the simplification of its life strategy [33], and the fish blood parasite, the parabodonid *T. borreli*, for which the assembly is of low quality, and therefore, artifacts are likely. Interestingly, compared to Kinetoplastea, the genome of diplonemid

P. papillatum has experienced more duplications in the *PAP1* and *DBP2B* genes, which correlates with its significantly larger size and may reflect a more complex regulation of RNA metabolism.

The general scarcity of introns in kinetoplastid genomes may offer an evolutionary advantage to these *r*-selected organisms by reducing the time required for gene transcription and genome replication, thereby allowing faster cell proliferation [7]. The preservation of introns within the *RBP20*, *PAP1*, and *DBP2B* genes (Fig. 2) implies a corresponding selective pressure. One of the factors determining intron retention, at least in some trypanosomatids, may be the presence of ncRNA genes within them (Fig. 1C; Additional file 1: Table S4). This suggests a contemporaneous functional significance, contrary to the notion that introns in Kinetoplastea are non-functional evolutionary relics [46]. However, the scenario differs for the RNA helicase gene, as its respective intron, which is the longest of all, has been independently lost at least five times within Trypanosomatidae (Fig. 2). We were unable to trace the ancestry of the *RBP20*, *PAP1*, and *DBP2B* introns beyond the common ancestor of eubodonids and trypanosomatids, as homologous introns could not be reliably identified in more divergent *T. borreli*, *Perkinsela* sp., and *P. papillatum*. While *P. papillatum* does contain introns within the putative orthologues of these genes, their positions significantly differ from those found in the kinetoplastids.

A plausible scenario for the origin of the three introns is that they were acquired by the common ancestor of *B. saltans* and trypanosomatids. An alternate model posits that these introns were already present in the kinetoplastid common ancestor and subsequently lost in *Perkinsela* sp. and *T. borreli* (the scenario depicted in Fig. 2 and considered above). To resolve these scenarios, more genomic sequences of Kinetoplastea are needed. Regardless, our analysis of intron distribution strongly suggests that *Perkinsela* sp. completely lacks the *cis*-spliceosomal introns (Fig. 2). However, contrary to our expectations, we identified its putative U1 snRNA gene, which is presumed to recognize the 5′ splice site based on complementarity during *cis*-splicing (Additional file 1: Table S1). It cannot be excluded that U1 snRNA and the reduced (compared to other kinetoplastids) repertoire of proteins binding it function in *trans*-splicing [46]. Notably, *Perkinsela* sp. and *T. borreli* are the only euglenozoans in our dataset lacking a homolog of U1-C, a highly conserved U1-associated protein that stabilizes the interaction between U1 small nuclear ribonucleoprotein and the 5′ splice site in humans [53]. In *T. brucei*, knockdown of the corresponding gene reduced the efficiency of *cis*-splicing for the *PAP1* gene, while *trans*-splicing remained unaffected [46]. While complete loss of spliceosomal introns is observed in the nucleomorph of the cryptophyte *Hemiselmis andersenii* and certain microsporidia with highly reduced genomes [7–9], *Perkinsela* sp. stands out as the first known lineage that has lost *cis*-splicing while retaining the *trans*-splicing. A comparative study involving the spliceosome of *Perkinsela* and that of its relatives, which carry out both *cis*- and *trans*-splicing, could elucidate the minimal spliceosome necessary for the latter process in Kinetoplastea and its specific components, an unresolved aspect of kinetoplastid biology [16].

The presence of prolonged conserved sequences at the 5′ splice site of kinetoplastid introns (Additional file 5: Fig. S4) mirrors the situation in rare U12-type introns, which are excised by a minor spliceosome [54]. In addition, unusually long 5′ splicing signals are observed in comparably intron-poor organisms, such as the diplomonads *Giardia* and *Spironucleus* [4]. In the only gene experimentally studied from this point of view in kinetoplastids, namely *PAP1* in *T. brucei*, the extended conserved sequence at the 5′ splice site of the intron proved to be crucial for its removal [20]. The conserved sequence, which is nearly complementary to the 5′ end of the U1 snRNA, may be important for recognition of the 5′ splice site by the spliceosome [20]. Our data suggests that the presence of an extended conserved motif cannot be a universal requirement for intron removal in kinetoplastids since the extent of complementarity varies across taxa and genes potentially contributing to different splicing efficiency and, therefore, expression levels. In addition to the canonical conserved intron elements, such as the polypyrimidine tract, kinetoplastid introns contain additional motifs (Additional file 6: Fig. S5), which might function as splicing regulators [55]. The position of these elements within introns varies across Kinetoplastea, suggesting that their function (if any) is mainly defined by their sequence.

Establishing scenarios of intron loss presents a considerable challenge since multiple mechanisms can operate within the same genome simultaneously or substitute each other during the evolution of the same lineage [7, 47]. We evaluated the potential role of the RT-mediated intron loss in Kinetoplastea, as it has been convincingly shown to cause an extensive intron loss in several eukaryotic groups, including microsporidia, fungi, angiosperms, and mammals [47, 48, 56]. Our findings suggest that the elimination of introns by this mechanism followed the radiation of Kinetoplastea from their common ancestor. Our analysis indicates that the same mechanism likely played a role in the massive intron loss in the kinetoplastid common ancestor, as suggested by the “exact” intron excision in the set of highly conserved genes in the extant kinetoplastids compared to the orthologues in the outgroups, for which the high-quality genome assemblies are available (Additional file 9: Fig. S7). However, the extent of this mechanistic contribution awaits further investigation, particularly as more high-quality genomes become available for the closest euglenozoan relatives of Kinetoplastea, such as diplonemids and euglenids.

Most functional information about the products of the three intron-containing genes is available for PAP1 in *T. brucei*. This is not the major poly(A) polymerase for mRNAs, this role is apparently carried out by PAP2 [57]. The main substrates for PAP1 are snoRNAs and lncRNAs, of which the former apparently undergo dual polyadenylation by PAP1 and PAP2 [18]. The RNA helicase encoded by the intron-containing gene has been localized to the nucleus in *T. brucei* [58], and RNAi-mediated depletion has not led to a severe impact on the culture growth [59]. While the exact function of RBP20 in trypanosomatids remains unknown, several studies suggest its nuclear localization in *T. brucei* [58, 60]. Additionally, in *T. brucei*, there is evidence of the interaction between RBP20 and RNA-binding zinc finger CCCH-type containing protein 11, which stabilizes several chaperone mRNAs in both bloodstream and procyclic stages of the parasite and represents a component of the post-transcriptional regulatory network [61].

The analysis of transcript isoforms in *T. brucei* [35, 36] and *L. mexicana* suggests the existence of sequences corresponding to SL-5′ UTR-exon1-exon2, and SL-exon2. Functional implications for the existence of such transcript isoforms (given they are successfully translated to proteins) might differ for the three intron-containing genes. The *PAP1* intron interrupts the catalytic domain, with some critical aspartic acid residues located in both exon 1 and exon 2 [57], likely making a protein encoded solely by exon 1 not functional. In contrast, in the *DBP2B* and *RBP20* genes, the intron separates the bulk of the functional domains localized in the exon 1 from the exon 2 containing low complexity glycine/arginine- and glycine/proline-rich sequences in RBP20 and RNA helicase, respectively (Additional file 3: Fig. S2). The presence of glycine/arginine-rich domain was shown to regulate the subcellular localization of proteins [62]. This raises the possibility that alternative *trans*-splicing of *RBP20* might lead to the generation of a protein encoded solely by exon 1 and localized elsewhere. Although our analysis of processed transcripts using long-read transcriptomic data did not reveal such isoforms, the low coverage does not allow us to confidently exclude such a possibility.

The two genes previously known to contain *cis*-spliceosomal introns in Kinetoplastea, poly(A) polymerase and RNA helicase, encode proteins potentially capable of interacting with multiple RNAs. Since PAP1 depletion leads to a decrease in the levels of mature snoRNAs in *T. brucei* [18], and many snoRNAs in this organism have been shown to play crucial roles in rRNA processing and modifications [63], it is tempting to speculate that the regulation of PAP1 splicing may serve as a mechanism for controlling ribosome biogenesis and/or function, ultimately enhancing adaptation to various environmental conditions. The identification of the intron in the gene coding for RBP20, protein potentially impacting transcription, strengthens the idea that regulation of *cis*-splicing represents an additional mechanism of fine-tuning gene expression in Kinetoplastea [64]. These introns might affect the speed of transcription, contain splicing enhancers and suppressors, and recognition sites for RNA binding proteins [65, 66]. Hence, the decision whether to *cis*- or *trans*-splice might affect protein localization and function. We propose that *cis*-splicing in Kinetoplastea, which essentially lack mechanisms for regulating gene expression at the transcriptional level, serves as a means of "regulating the regulators"—genes whose products may influence the expression of numerous other genes. This hypothesis awaits a comprehensive experimental validation.

Conclusions

In this work, we scrutinized *cis*-spliceosomal introns in the available genome assemblies of kinetoplastids and their free-living relatives. We observed extended splice site conservation and identified putative ncRNAs within kinetoplastid introns. In addition, we discovered a novel *cis*-spliceosomal intron-containing gene encoding an RNA-binding protein, bringing the total number of these genes in trypanosomatids to three and demonstrated the complete absence of such introns in the genome of the early-branching kinetoplastid *Perkinsela*, which is endowed with *trans*-splicing. All three *cis*-spliced genes are involved in key processes of RNA metabolism; therefore, we propose that Kinetoplastea, which lost regulation of individual gene expression at the transcriptional level, preserve *cis*-splicing for genes having a global impact on the transcriptome.

Methods

Lafontella sp. genome sequencing and assembly

The genome assemblies are available for virtually all trypanosomatid genera, except *Lafontella*, for which only the transcriptome has been published [28]. Therefore, we sequenced and assembled its genome for the purpose of intron identification. *Lafontella* sp. isolate GMO-01 was cultivated as described previously [28]. The species identity was confirmed as in [67]. DNA was isolated using DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) and sequenced on Illumina NovaSeq 6000 platform at Macrogen (Amsterdam, Netherlands) yielding approximately 22.8 million 150-nt-long reads. The reads were adapter- and quality-trimmed using Fastp v.0.23.2 with default settings [68], and around 21.9 million paired-end reads were retained for further analysis. The quality of raw and trimmed reads was evaluated using FastQC v.0.11.8 [69]. These trimmed genomic reads were assembled using SPAdes v.3.13.0 [70] with default settings. The scaffolds shorter than 200 bp were discarded and the final assembly contained 32,943 scaffolds with a total length of approximately 34.2 Mb and N_{50} of 3473 base pairs. The same procedure was used for the reassembly of *P. confusum* genome using published reads [71] with the aim of U1 snRNA identification (see below).

Analysis of *cis*-spliceosomal intron-containing genes

The high-quality genome assemblies and available genome-derived proteomes for 78 kinetoplastid species (of which 74 belong to Trypanosomatidae) and the diplonemid *P. papillatum* were obtained from the sources specified in Table S1 (Additional file 1). Twenty-five of these genomes (one species per genus; marked by asterisks in Table S1) comprised a representative dataset used for the analysis. The homologs of the two genes known to contain *cis*-spliceosomal introns in Trypanosomatidae, PAP1 and DBP2B, as well as the newly identified gene encoding RPB20 were searched for in the reference Euglenozoa dataset (Additional file 1: Table S1). BLASTp and tBLASTn searches [72] were performed with an *e*-value thresholds of e^{-10} (PAP1 and RBP20) and e^{-100} (RNA helicase) using *T. brucei* proteins as queries (Tb927.3.3160 – PAP1; Tb927.8.1510 – DBP2B; Tb927.8.6440 – RBP20).

For phylogenetic analysis, the respective protein sequences after the exclusion of incomplete sequences (Additional file 10) were aligned with MAFFT v.7.520 [73] using L-INS-i algorithm, BLOSUM45 substitution matrix, and gap open penalty of 1.25. The resulting alignment was trimmed with trimAl v.1.2 [74] in “gappyout” mode. Maximum likelihood phylogenetic trees were inferred in IQ-TREE v.2.2.2.6 [75] with automatic selection of the amino acid substitution model by the built-in ModelFinder and edge support estimated by ultrafast bootstrap method with 1000 replicates.

Functional domains in *T. brucei* proteins were predicted with the online SMART tool [76]. Low-complexity regions in RBP20 protein were predicted using the PlaToLoCo server [77] with default settings.

Intron identification

When available, the transcriptomic reads for the reference species were downloaded from the NCBI Sequence Read Archive and/or European Nucleotide Archive (Additional file 1: Table S1) [78]. Illumina paired-end and single transcriptomic reads underwent adapter and quality trimming using Trimmomatic v.0.39 [79]. Paired and single reads were processed in paired- and single-end mode, respectively. The trimming parameters applied were [TruSeq3-PE-2.fa/Nextera]:2:20:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15. For paired-end reads, the “MINLEN” parameter was set to 50 and for single reads to 30. In the case of paired-end reads, only those reads that remained paired after trimming were included in further analyses. Read quality was assessed with FastQC v.0.11.8 before and after trimming. Trimmed Illumina reads were mapped to the genome assemblies using a splice-aware aligner HISAT2 v.2.2.1 [80] with default settings. Long reads were mapped using Minimap2 v.2.26-r1175 [81] with “map-pb” and “map-ont” for PacBio and Oxford Nanopore reads, respectively, and other options at their default values. The read mappings were sorted using Samtools v.1.13 [82] and visualized in Artemis v.18.2.0 [83].

The identification of intron borders in the genes found as described above relied on the following criteria: (i) the presence of canonical GT and AG dinucleotides at the 5′ and 3′ ends, respectively and (ii) for species with available transcriptomic data, at least one split read covering the intron was required (a few exceptions are specified in Additional file 1: Table S1, where we inferred the presence of an intron without transcriptomic evidence). By a split read, we understand a read originating from a mature mRNA and exhibiting a gapped alignment to the reference genome, where the gap corresponds to the intron, while the mapped regions represent the exons. For species lacking transcriptomic data, intron borders were predicted by alignment of protein sequences from closely related species and considering the conserved GT/AG dinucleotides at the intron borders. The correctness of intron prediction was verified by translating the respective exon sequences and aligning predicted proteins to their homologs.

The search for new *cis*-spliceosomal intron-containing genes was conducted using two independent approaches. In the frame of the first approach, 15-nt-long sequences located at the 5′ splice site in poly(A) polymerase- and RNA helicase-encoding genes were collected from 25 species in the representative dataset (Additional file 1: Table S1; Additional file 10). The sequences were used as an input for MEME suite v.5.5.5 [84] with the maximum motif length set to 5 and then 15 nt with other settings left at default values. Both retrieved motifs were used as the input for FIMO program of the MEME suite v.5.5.5, which scanned genome assemblies of *Blastocrithidia nonstop*, *Bodo saltans*, *Leishmania major*, *Leptomonas pyrrhocoris*, *Perkinsela* sp., and *T. brucei* for the presence of potential intronic motifs. The top 50 hits in each genome were manually inspected using transcriptomic read mappings visualized in Artemis v.18.2.0. The second approach focused on analyzing split reads mapping and did not depend on sequence conservation at the splice sites (enabling the identification of nonconventional introns). For this, the genomes of *B. saltans* and *Perkinsela* sp., along with the corresponding read mappings generated as described above, were used as input for Portcullis v.1.1.2 with the default settings [85]. The set of high-confidence splice junctions for each species was further filtered by discarding junctions supported by fewer than 5 unique reads and those located within 50 nucleotides of the scaffold ends. The remaining candidates were manually inspected using Artemis.

Analysis of transcript isoforms using long reads and PCR

Publicly available long-read transcriptomic data produced using MinION and PacBio platforms were downloaded from NCBI (PRJEB38965 [86] and PRJEB60500 [87] – *Leishmania infantum*; PRJEB60502 [88] – *L. donovani*; PRJEB60504 [89] – *L. braziliensis*; PRJEB60505 [90] – *L. major*; PRJEB39255 [91] and PRJEB60364 [35] – *T. brucei*). The read quality was assessed using FastQC [69], and the reads with a Phred quality score lower than 8 and/or shorter than 500 nt were discarded. Reads containing spliced leader sequence were selected using Cutadapt v.1.18 [92] with the following settings: -g SL -m 20 -O 8 -e 0.3, where "SL" corresponds to a 15-nt-long conserved sequence at the 3′ end of the mini-exon (TTTCTGTACTTTATT for *T. brucei*). Only those reads that contained SL sequence within the first 50 nt at the 5′ end were retained. Complete gene sequences of *PAP1*, *DBP2B*, and *RBP20* were used as queries in BLASTn searches with the filtered reads as the database, an *e*-value threshold of e^{-50}, and other settings left as default [72]. The gene sequences were aligned to the respective BLAST hits using MAFFT with default settings and visually inspected in Jalview v.2.11.3.2 [93].

For the analysis of transcript isoforms using PCR, *Leishmania mexicana* promastigotes were cultured in M199 medium (Cayman Chemical, Ann Arbor, USA) supplemented with 2 μg/mL biopterin (Cayman Chemical, Ann Arbor, USA), 2 μg/mL hemin (Jena Bioscience, Jena, Germany), 25 mM HEPES, 50 units/mL of penicillin/streptomycin (both from Biowest, Nuaillé, France), and 10% fetal bovine serum (BioSera, Cholet, France) at 23 °C. Metacyclic promastigotes and amastigotes were differentiated by adjusting pH and temperature according to an established protocol [94]. Total RNA was isolated using TRIzol Reagent (MRC, Cincinnati, USA). Transcriptor First Strand cDNA synthesis kit (Roche Life Science, Penzberg, Germany) was used for cDNA synthesis with oligo (dT) primer.

All PCR amplification reactions were performed using PCRBIO Taq Mix Red (PCR Biosystems, London, UK). A single-step conventional PCR assay was performed for exon-exon junction amplification. Amplification of the SL-exon1 and SL-exon2 regions in all three genes was achieved by a semi-nested PCR approach with the use of SL_F and internal reverse primers in the second run (Additional file 1: Table S2). In the case of *PAP1*, amplicons produced in the first PCR run were purified by the QIAquick PCR & Gel Cleanup Kit (Qiagen, Hilden, Germany) before the second run. Conventional PCRs and the first run of semi-nested PCRs were performed with an annealing temperature of 55 °C, which was increased to 58 °C in the second-step reactions. The PCR products were resolved in a 1% agarose gel that was post-stained with Midori Green dye (Nippon Genetics Europe, Düren, Germany), and the major bands were extracted from the gel using the QIAquick PCR & Gel Cleanup Kit and sequenced at Eurofins Genomics (Ebersberg, Germany). The SL-exon1 amplicon of *RBP20* was purified from the gel and cloned into pJET1.2 (Thermo Fisher Scientific, Carlsbad, USA) prior to sequencing as above.

Inference of intron features

Sequence logos were produced using the online tool WebLogo [95] for single-gene alignments of sequences from a representative dataset (Additional file 1: Table S1; Additional file 10) focusing on two regions: (i) the first 50 nt of the intron and 10 nt upstream of the splice site (i.e., belonging to exon 1) and (ii) the last 50 nt of the intron and 10 nt downstream (i.e., belonging to exon 2). The same analysis was done for alignments of the region (i) in all available *cis*-spliceosomal intron-containing genes of a single species.

Intronic motifs in kinetoplastids were identified using the GLAM2 program of the MEME suite v.5.5.5 with the search performed only on the sense strand and other settings left as default. For each motif, the replicate with the highest score was collected.

For the identification of putative ncRNAs, intron sequences were used as queries for online searches in the RNAcentral database (release 23) [39]. The hits were filtered according to the following criteria: *e*-value threshold of e^{-10}, target sequence coverage higher than 20%, sequence identity higher than 35%, and only the best hit (based on the lowest *e*-value) was retained for each query sequence. In addition, available RNA-seq data for *L. donovani* promastigotes enriched for the fraction of non-polysomal RNAs bound to proteins were downloaded from the NCBI [40, 41]. Transcriptome reads were processed and mapped to the genome as described in the section "Intron identification," except for the "MINLEN" parameter of Trimmomatic that was set to 30.

For establishing if introns in the genes encoding PAP1, DBP2B, and RBP20 are ancestral for Glycomonada (the clade consisting of Kinetoplastea and Diplonemea [96]), protein sequences for each gene identified in reference species (Additional file 10) were aligned to their orthologues from *P. papillatum* using MAFFT with default settings in Jalview v.2.11.3.2. We considered that the intron position is shared between kinetoplastids and diplonemids (i.e., the intron is ancestral to Glycomonada) if the intron positions differed by no more than five codons between at least 50% of species in the reference dataset and *P. papillatum*.

Analysis of intron loss in Kinetoplastea

For analysis of intron loss in kinetoplastids, we aligned homologous proteins from closely related pairs of species differing in intron content, such as *B. saltans* and *Perkinsela* sp., *Porcisia deanei* and *P. hertigi*, *Wallacemonas rigidus* and *Sergeia podlipaevi*, *Strigomonas culicis* and *S. oncopelti*, *T. cruzi* and *T. conorhini*, using MAFFT with default settings in Jalview v.2.11.3.2. Intron loss was considered "exact" if no gaps were observed within five amino acids (five codons) to either side of putative intron position. Synteny analysis was performed and data visualized in Artemis Comparison Tool v.18.2.0 [97] with the homology regions identified with tBLASTx at NCBI with the following parameters: max target sequences: 5000, expected threshold: e^{-20}, and other settings at default values.

For the analysis of intron loss in the kinetoplastid common ancestor, annotated proteins of the trypanosomatids *L. major*, *P. confusum*, and *T. brucei* along with the eubodonid *B. saltans* were downloaded from TriTrypDB release 63 [98], and data for the diplonemid *P. papillatum* and a heterolobosean relative of Euglenozoa, *N. gruberi*, were obtained from the NCBI [99, 100]. Due to a high fragmentation of the genome of *Euglena gracilis* and the presence of nonconventional introns, only *N. gruberi* and *P. papillatum* were used as references for this analysis. Proteins were clustered into OGs using OrthoFinder v.2.5.5 [101] with BLAST as a sequence search program and with other parameters at default values. The OGs containing a single protein per species were retained for further analysis. The proteins within each OG were aligned using MAFFT v.7.490 [73] with L-INS-i algorithm. The average sequence identity within each OG was calculated using esl-alistat script from the HMMER package v.3.3.2 [102]. OGs with average identity ≥ 50% were manually checked for the presence of ancestral introns, i.e., those with a shared position (separated by an arbitrary threshold of no more than 10 codons) between *P. papillatum* and *N. gruberi*. Please note that we applied a more relaxed threshold compared to that used for comparing kinetoplastid genes to *P. papillatum* (five codons), acknowledging a larger evolutionary distance between Euglenozoa and Heterolobosea. The identification of these introns relied on transcriptomic data mapped to the respective genomes, as described in the section "Intron identification". The alignments of the respective OGs were visually inspected in Jalview v.2.11.3.2.

For the identification of RT domain-containing proteins, Pfam model PF00078 [103] was used as a query for hidden Markov model-based searches with an *e*-value e^{-5} against the database of annotated euglenozoan proteins.

Identification and analysis of U1 snRNA orthologues and U1-associated proteins

For searching the U1 snRNAs in the representative genome dataset (Additional file 1: Table S1), annotated U1 snRNAs were collected from TriTrypDB release 67: *L. braziliensis* (LbrM.23.snRNA1), *L. donovani* (LdCL_230012650), *L. major* (LmjF.23.snRNA.0), *T. brucei* (Tb927.8.2855), and *T. cruzi* (C3747_21nc9). The sequences were aligned using LocARNA v.1.9.1 [104], and the alignments in the Stockholm format served as an input to build and calibrate a covariance model using the "cmbuild" and "cmcalibrate" functions of Infernal v.1.1.4 [105]. Next, we searched for snRNA sequences in genomes of representative species using a calibrated model with "cmsearch" function of Infernal v.1.1.4 with default settings. Since we were unable to obtain hits for all species using this approach, the reference dataset of snRNAs was expanded by combining additional hits from the first round of searches and the publicly available U1 snRNA sequence of *E. gracilis* (GenBank accession U57366.1). The searches were repeated after incorporating the hits obtained in the previous rounds until no new U1 snRNA sequences could be identified in the dataset.

Using the previously published information on U1 snRNAs of *Crithidia fasciculata*, *Leishmania tarentolae*, and *T. brucei* [43, 45], we manually identified the termini that had been incomplete predominantly at the 3′ end due to a relatively low sequence conservation, functional motifs, and predicted secondary structures for the sequences analyzed in this work. The IPknot web server [106] was used to facilitate the identification of potentially pairing subsequences. Out of the four genomic copies of U1 snRNA for *P. papillatum* differing by one to three nucleotides, the one localized in the scaffold JAPJBO010001534.1 (positions 29,166–29,329) was used for the secondary structure prediction based on a previously published inference [14]. Visualization of RNA secondary structures was made using RnaViz v.2.0.3 [107].

For each species from the representative dataset, potential interactions of the 5′ end of U1 snRNA with that of the introns of each gene were predicted based on the previously published inference [43] and those made using the RNAcofold web server [108].

For the identification of U1-associated proteins, the sequences of *T. brucei* U1-70 K (Tb927.8.4830), U1A (Tb927.10.8280/8300), U1-C (Tb927.10.2120), and U1-24 K (Tb927.3.1090) were downloaded from TriTrypDB release 63. They were used as BLAST (BLASTp and tBLASTn) queries with an *e*-value of 1 and all euglenozoan protein/genome sequences as a database. To detect divergent homologs, BLASTp hits with an *e*-value $\leq e^{-20}$ were retrieved, aligned using MAFFT v.7.520 with the L-INS-i algorithm and subsequently used for several rounds of hidden Markov model-based searches with hmmsearch from the HMMER package v.3.3.2 using default settings and euglenozoan proteins as the database. The identities of the hits were confirmed using HHpred and InterProScan web servers [109, 110].

Abbreviations

cDNA	Complementary DNA
DBP2	DEAD-box ATPase 2
DBP2B	ATP-dependent RNA helicase (DEAD-box protein 2B)
lncRNA	Long noncoding RNA
Mb	Megabase
ncRNA	Noncoding RNA
nt	Nucleotide
OG	Orthologous group
PAP1	Poly(A) polymerase 1
PCR	Polymerase chain reaction
pre-mRNA	Precursor mRNA
RBP20	RNA-binding protein 20
RT	Reverse transcriptase
SL	Spliced leader
Sm protein	Smith protein
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
U1-24K	U1 small nuclear ribonucleoprotein 24 kDa
U1-70K	U1 small nuclear ribonucleoprotein 70 kDa
U1A	U1 small nuclear ribonucleoprotein A
U1-C	U1 small nuclear ribonucleoprotein C
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-02080-z.

Additional file 1: Table S1-S4. Table S1. Dataset used in this study. Some genome assemblies contained gaps within the sequences of the studied genes, which is mentioned in the columns for introns' presence. Representative genome assemblies are marked with asterisks. Table S2. PCR primers used in this study. Table S3. Position of common motifs identified within the intron of the three studied genes. Start and end positions of a motif were estimated relative to the 5' end of the intron (positive numbers, blue bars) and to the 3' end (negative numbers, red bars). Table S4. Regions of kinetoplastid introns demonstrating similarity to noncoding RNAs.

Additional file 2: Fig. S1. Maximum likelihood phylogenetic trees of the three studied genes reconstructed using amino acid sequences. A. Poly(A) polymerase. B. RNA helicase. C. RNA-binding protein 20. The clades of poly(A) polymerases and RNA helicases containing exclusively intron-less genes are collapsed and designated as PAP2 and DBP2A, respectively. Numbers at branches are ultrafast bootstrap supports, maximal (100) values are shown as circles. Double-crossed branches have 50% of their actual length. The scale bar denotes the number of substitutions per site.

Additional file 3: Fig. S2. Domain architecture of proteins encoded by intron-containing genes in *T. brucei*. The scale bar shows protein length. Intron position is indicated by a vertical dashed line with a number indicating the position of the first amino acid encoded by the second exon. Protein domains are indicated with colored shapes: black rectangle – poly(A) polymerase; red rectangle and orange rhombus – RNA helicase; green square – RNA recognition motif. Coiled coils and low complexity regions are indicated with green and magenta boxes, respectively. G – glycine; P – proline; Q – glutamine; R – arginine.

Additional file 4: Fig. S3. Transcript isoforms of the studied genes identified by PCR in *Leishmania mexicana*. A-C. *PAP1*. D-E. *RBP20*. F-G. *DBP2B*. Sequences of PCR products with accompanying chromatograms are aligned to predicted mature transcripts for each gene (except panel C where a part of intron is depicted). Base calling quality is shown as a blue bar graph; chromatogram peaks for individual bases are depicted in red, blue, yellow, and green for A, C, G, and T, respectively. PCR primers are shown as green bars.

Additional file 5: Fig. S4. Conservation of splice sites in the genes encoding poly(A) polymerase, RNA helicase and RNA binding protein at the interspecific (A) and intraspecific (B) levels. The analysis at the interspecific level is based on the representative sequences from genomes of 24 kinetoplastid species: *Angomonas deanei*, *Blastocrithidia nonstop*, *Blechomonas ayalai*, *Bodo saltans*, *Borovskyia barvae*, *Crithidia fasciculata*, *Endotrypanum monterogeii*, *Herpetomonas samuelpessoai*, *Jaenimonas drosophilae*, *Kentomonas sorsogonicus*, *Leishmania major*, *Leptomonas pyrrhocoris*, *Lotmaria passim*, *Novymonas esmeraldas*, *Obscuromonas modryi*, *Paratrypanosoma confusum*, *Phytomonas francai*, *Porcisia deanei*, *Sergeia podlipaevi*, *Strigomonas culicis*, *Trypanosoma brucei*, *Vickermania ingenoplastis*, *Wallacemonas rigidus*, *Zelonia costaricensis*. For each gene, 60 nucleotides around 5 and 3' splice sites (10 nucleotides of the exon and 50 nucleotides of the intron sequence) were extracted and analyzed with WebLogo. The conservation within the exon upstream of the intron border was not always associated with preservation of the amino acid sequence (for example, Phe/Cys change in the last position of *PAP1* exon 1 or Leu/Met alteration in the penultimate codon of the first exon of the *RBP20* gene). Within the five last nucleotides A or T were predominant. The 5' end of the second exon demonstrated different patterns between the three genes, of which two could not be explained by amino acid conservation. In each of the first three codons of exon 2, the *PAP1* gene had predominant pyrimidines at the third position, while the first and third position nucleotides in the same codons were invariant in the *DBP2B* gene. Exon 2 of the *RBP20* gene represented an exception: the absolute conservation of the first two nucleotides (GT) was associated with encoding an invariant Val. For the analysis at the intraspecific level (B), only the representative species, for which the sequences of the three *cis*-spliced intron containing genes were available, are included. Black dotted vertical lines indicate splice sites.

Additional file 6: Fig. S5. Nucleotide motifs identified in the introns of *PAP1*, *DBP2B* and *RBP20* genes by GLAM2. Except for polypyrimidine tracts, no further motifs were detected in the *PAP1* gene intron. The *RBP20* introns also featured an adenine-rich motif. RNA helicase introns contained GT- and AC-rich motifs as well as a distinct polypyrimidine tract, where Ts were predominant.

Additional file 7: Fig. S6. Alignment of U1 snRNAs in Euglenozoa. A. Trypanosomatidae and *Bodo saltans*. B. Alignment of U1 snRNA candidates of *Paratrypanosoma confusum* and *Perkinsela* sp. with the sequences of *Bodo saltans, C. fasciculata and T. brucei* as references. The three characteristic conservative motifs are boxed. The purple box-line-box signs designate hairpins (stems and loops, respectively). The majority-rule consensus of aligned sequences is shown on top.

Additional file 8: Predicted interactions between the 5' ends of U1 snRNA and the three studied protein-coding genes. Multiple competing variants are shown in some cases.

Additional file 9: Fig. S7. Protein alignment around the exon-exon junctions corresponding to presumably ancestral euglenozoan introns. Intron positions are marked by black rectangles. Species abbreviations: *B. saltans – Bodo saltans*; *L. major – Leishmania major*; *N. gruberi – Naegleria gruberi*; *P. confusum – Paratrypanosoma confusum*; *P. papillatum – Paradiplonema papillatum*; *T. brucei – Trypanosoma brucei*.

Additional file 10. Intron and protein sequences of *PAP1*, *DBP2B*, and *RBP20*.

Acknowledgements
Not applicable.

Authors' contributions
AB conceived the study and collected the data. AB, AYK, KS, and CM performed the analyses and interpreted the data. EK and NK performed the experiments. AB and AYK were major contributors in writing the manuscript. JL, MCF and VY edited the manuscript. All authors read and approved the final manuscript.

Funding
This work was primarily supported by a grant from the Grant Agency of Czech Republic (project 23-07695S to A.B. and A.Y.K; analyses of ***cis*** -spliced genes, analysis of intron loss, analysis of U1 snRNAs). Additional funding was provided by the EU's Operational Program "Just Transition" (LERCO CZ.10.03.01/00/22_003/0000003 to V.Y., A.Y.K., and N.K.; ***Lafontella*** sp. genome sequencing and assembly), Ministry of Education, Youth and Sports of the Czech Republic (INTER-EXCELLENCE-LUASK22033 to V.Y. and N.K.; phylogenetic analysis), and the University of Ostrava (SGS/2024 to E.K.; experimental validation of transcript isoforms).

Data availability
The *Lafontella* sp. genome assembly was deposited in NCBI under the BioProject accession PRJNA1102159 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1102159) [111]. The data that support the findings of this study are openly available in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) and TriTrypDB (https://tritrypdb.org/tritrypdb/app/). The genome and transcriptome accession numbers can be found in this publication and its supplementary information files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava 710 00, Czech Republic. [2]Zoological Institute of the Russian Academy of Sciences, St. Petersburg 199034, Russia. [3]School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK. [4]Institute of Parasitology, Czech Academy of Sciences, České Budějovice 370 05, Czech Republic. [5]Faculty of Science, University of South Bohemia, České Budějovice 370 05, Czech Republic.

Received: 14 September 2024 Accepted: 26 November 2024
Published online: 03 December 2024

References

1. Petrillo E. Do not panic: an intron-centric guide to alternative splicing. Plant Cell. 2023;35(6):1752–61.
2. Wilkinson ME, Charenton C, Nagai K. RNA splicing by the spliceosome. Annu Rev Biochem. 2020;89:359–88.
3. Lei Q, Li C, Zuo ZX, Huang CH, Cheng HH, Zhou RJ. Evolutionary insights into RNA *trans*-splicing in vertebrates. Genome Biol Evol. 2016;8(3):562–77.
4. Hudson AJ, McWatters DC, Bowser BA, Moore AN, Larue GE, Roy SW, et al. Patterns of conservation of spliceosomal intron structures and spliceosome divergence in representatives of the diplomonad and parabasalid lineages. BMC Evol Biol. 2019;19(1):162.
5. Jo BS, Choi SS. Introns: the functional benefits of introns in genomes. Genomics Inform. 2015;13(4):112–8.
6. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7:11.
7. Jeffares DC, Mourier T, Penny D. The biology of intron gain and loss. Trends Genet. 2006;22(1):16–22.
8. Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, et al. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci U S A. 2007;104(50):19908–13.
9. Cuomo CA, Desjardins CA, Bakowski MA, Goldberg J, Ma AT, Becnel JJ, et al. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. 2012;22(12):2478–88.
10. Liang XH, Haritan A, Uliel S, Michaeli S. *Trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. Eukaryot Cell. 2003;2(5):830–40.
11. Kostygov AY, Karnkowska A, Votýpka J, Tashyreva D, Maciszewski K, Yurchenko V, et al. Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. Open Biol. 2021;11(3): 200407.
12. Gawryluk RMR, Del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, et al. Morphological identification and single-cell genomics of marine diplonemids. Curr Biol. 2016;26(22):3053–9.
13. Milanowski R, Karnkowska A, Ishikawa T, Zakrys B. Distribution of conventional and nonconventional introns in tubulin (alpha and beta) genes of euglenids. Mol Biol Evol. 2014;31(3):584–93.
14. Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, et al. Recent expansion of metabolic versatility in *Diplonema papillatum*, the model species of a highly speciose group of marine eukaryotes. BMC Biol. 2023;21(1):99.
15. Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, et al. Transcriptome, proteome and draft genome of *Euglena gracilis*. BMC Biol. 2019;17(1):11.
16. Günzl A. The pre-mRNA splicing machinery of trypanosomes: complex or simplified? Eukaryot Cell. 2010;9(8):1159–70.
17. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res. 2010;38(15):4946–57.
18. Chikne V, Gupta SK, Doniger T, K SR, Cohen-Chalamish S, Waldman Ben-Asher H, et al. The canonical poly(A) polymerase PAP1 polyadenylates non-coding RNAs and is essential for snoRNA biogenesis in *Trypanosoma brucei*. J Mol Biol. 2017;429(21):3301–18.
19. Rajan KS, Madmoni H, Bashan A, Taoka M, Aryal S, Nobe Y, et al. A single pseudouridine on rRNA regulates ribosome structure and function in the mammalian parasite *Trypanosoma brucei*. Nat Commun. 2023;14(1):7462.
20. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, et al. A new twist in trypanosome RNA metabolism: *cis*-splicing of pre-mRNA. RNA. 2000;6(2):163–9.
21. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 2005;309(5733):436–42.
22. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome *Trypanosoma brucei*. Science. 2005;309(5733):416–22.
23. Ma WK, Paudel BP, Xing Z, Sabath IG, Rueda D, Tran EJ. Recruitment, duplex unwinding and protein-mediated inhibition of the DEAD-box RNA helicase Dbp2 at actively transcribed chromatin. J Mol Biol. 2016;428(6):1091–106.
24. Lai YH, Choudhary K, Cloutier SC, Xing Z, Aviran S, Tran EJ. Genome-wide discovery of DEAD-box RNA helicase targets reveals RNA structural remodeling in transcription termination. Genetics. 2019;212(1):153–74.

25. Bond AT, Mangus DA, He F, Jacobson A. Absence of Dbp2p alters both nonsense-mediated mRNA decay and rRNA processing. Mol Cell Biol. 2001;21(21):7366–79.
26. Song QX, Lai CW, Liu NN, Hou XM, Xi XG. DEAD-box RNA helicase Dbp2 binds to G-quadruplex nucleic acids and regulates different conformation of G-quadruplex DNA. Biochem Biophys Res Commun. 2022;634:182–8.
27. Kovalev N, Barajas D, Nagy PD. Similar roles for yeast Dbp2 and RH20 DEAD-box RNA helicases to Ded1 helicase in tombusvirus plus-strand synthesis. Virology. 2012;432(2):470–84.
28. Albanaz ATS, Carrington M, Frolov AO, Ganyukova AI, Gerasimov ES, Kostygov AY, et al. Shining the spotlight on the neglected: new high-quality genome assemblies as a gateway to understanding the evolution of Trypanosomatidae. BMC Genom. 2023;24(1):471.
29. Kostygov AY, Albanaz ATS, Butenko A, Gerasimov ES, Lukeš J, Yurchenko V. Phylogenetic framework to explore trait evolution in Trypanosomatidae. Trends Parasitol. 2024;40(2):96–9.
30. Mertes C, Scheller IF, Yepez VA, Celik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. Nat Commun. 2021;12(1):529.
31. De Gaudenzi J, Frasch AC, Clayton C. RNA-binding domain proteins in Kinetoplastids: a comparative analysis. Eukaryot Cell. 2005;4(12):2106–14.
32. Butenko A, Opperdoes FR, Flegontova O, Horak A, Hampl V, Keeling P, et al. Evolution of metabolic capabilities and molecular features of diplonemids, kinetoplastids, and euglenids. BMC Biol. 2020;18(1):23.
33. Tanifuji G, Cenci U, Moog D, Dean S, Nakayama T, David V, et al. Genome sequencing reveals metabolic and cellular interdependence in an amoeba-kinetoplastid symbiosis. Sci Rep. 2017;7:11688.
34. *Perkinsela* sp. pyNocScin1, whole genome shotgun sequencing project. 2024. GenBank https://www.ncbi.nlm.nih.gov/nuccore/CAXHTS000000000.1.
35. Kruse E, Goringer HU. Nanopore-based direct RNA sequencing of the *Trypanosoma brucei* transcriptome identifies novel lncRNAs. 2023. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB60364.
36. Kruse E, Goringer HU. Nanopore-based direct RNA sequencing of the *Trypanosoma brucei* transcriptome identifies novel lncRNAs. Genes. 2023;14(3):610.
37. Lucke S, Jurchott K, Hung LH, Bindereif A. mRNA splicing in *Trypanosoma brucei*: branch-point mapping reveals differences from the canonical U2 snRNA-mediated recognition. Mol Biochem Parasitol. 2005;142(2):248–51.
38. Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L, Fedorov A. Critical association of ncRNA with introns. Nucleic Acids Res. 2011;39(6):2357–66.
39. RNAcentral Consortium. RNAcentral: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 2021;2021(49):D212–20.
40. Freitas Castro F, Ruy PC, Nogueira Zeviani K, Freitas Santos R, Simoes Toledo J, Kaysel CA. Evidence of putative non-coding RNAs from *Leishmania* untranslated regions. Mol Biochem Parasitol. 2017;214:69–74.
41. Freitas Castro F, Ruy PC, Nogueira Zeviani K, Freitas Santos R, Simoes Toledo J, Kaysel Cruz A. Evidence of putative non-coding RNAs from *Leishmania* untranslated regions. 2016. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA343249.
42. Stark H, Dube P, Luhrmann R, Kastner B. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. Nature. 2001;409(6819):539–42.
43. Djikeng A, Ferreira L, D'Angelo M, Dolezal P, Lamb T, Murta S, et al. Characterization of a candidate *Trypanosoma brucei* U1 small nuclear RNA gene. Mol Biochem Parasitol. 2001;113(1):109–15.
44. Breckenridge DG, Watanabe Y, Greenwood SJ, Gray MW, Schnare MN. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. Proc Natl Acad Sci U S A. 1999;96(3):852–6.
45. Schnare MN, Gray MW. Spliced leader-associated RNA from *Crithidia fasciculata* contains a structure resembling stem/loop II of U1 snRNA. FEBS Lett. 1999;459(2):215–7.
46. Preußer C, Rossbach O, Hung LH, Li D, Bindereif A. Genome-wide RNA-binding analysis of the trypanosome U1 snRNP proteins U1C and U1–70K reveals *cis/trans*-spliceosomal network. Nucleic Acids Res. 2014;42(10):6603–15.
47. Cohen NE, Shen R, Carmel L. The role of reverse transcriptase in intron gain and loss mechanisms. Mol Biol Evol. 2012;29(1):179–86.
48. Wang H, Devos KM, Bennetzen JL. Recurrent loss of specific introns during angiosperm evolution. PLoS Genet. 2014;10(12): e1004843.
49. Yurchenko V, Butenko A, Kostygov AY. Genomics of Trypanosomatidae: where we stand and what needs to be done? Pathogens. 2021;10(9):1124.
50. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 2006;7(3):211–21.
51. Tikhonenkov DV, Gawryluk RMR, Mylnikov AP, Keeling PJ. First finding of free-living representatives of Prokinetoplastina and their nuclear and mitochondrial genomes. Sci Rep. 2021;11(1):2946.
52. Pita S, Diaz-Viraque F, Iraola G, Robello C. The Tritryps comparative repeatome: insights on repetitive element evolution in trypanosomatid pathogens. Genome Biol Evol. 2019;11(2):546–51.
53. Kondo Y, Oubridge C, van Roon AM, Nagai K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. eLife. 2015;4:e04986.
54. Akinyi MV, Frilander MJ. At the intersection of major and minor spliceosomes: crosstalk mechanisms and their impact on gene expression. Front Genet. 2021;12: 700744.
55. Murray JI, Voelker RB, Henscheid KL, BryanWarf M, Berglund JA. Identification of motifs that function in the splicing of non-canonical introns. Genome Biol. 2008;9(6):R97.
56. Zhu T, Niu DK. Frequency of intron loss correlates with processed pseudogene abundance: a novel strategy to test the reverse transcriptase model of intron loss. BMC Biol. 2013;11:23.
57. Koch H, Raabe M, Urlaub H, Bindereif A, Preusser C. The polyadenylation complex of *Trypanosoma brucei*: characterization of the functional poly(A) polymerase. RNA Biol. 2016;13(2):221–31.
58. Billington K, Halliday C, Madden R, Dyer P, Barker AR, Moreira-Leite FF, et al. Genome-wide subcellular protein map for the flagellate parasite *Trypanosoma brucei*. Nat Microbiol. 2023;8(3):533–47.
59. Alsford S, Turner DJ, Obado SO, Sanchez-Flores A, Glover L, Berriman M, et al. High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. Genome Res. 2011;21(6):915–24.
60. Wurst M, Robles A, Po J, Luu VD, Brems S, Marentije M, et al. An RNAi screen of the RRM-domain proteins of *Trypanosoma brucei*. Mol Biochem Parasitol. 2009;163(1):61–5.
61. Singh A, Minia I, Droll D, Fadda A, Clayton C, Erben E. Trypanosome MKT1 and the RNA-binding protein ZC3H11: interactions and potential roles in post-transcriptional regulatory networks. Nucleic Acids Res. 2014;42(7):4652–68.
62. Doron-Mandel E, Koppel I, Abraham O, Rishal I, Smith TP, Buchanan CN, et al. The glycine arginine-rich domain of the RNA-binding protein nucleolin regulates its subcellular localization. EMBO J. 2021;40(20): e107158.
63. Chikne V, Shanmugha Rajan K, Shalev-Benami M, Decker K, Cohen-Chalamish S, Madmoni H, Biswas VK, Kumar Gupta S, Doniger T, Unger R, Tschudi C, Ullu E, Michaeli S. Small nucleolar RNAs controlling rRNA processing in *Trypanosoma brucei*. Nucleic Acids Res. 2019;18;47(5):2609–2629.
64. Kramer S. Nuclear mRNA maturation and mRNA export control: from trypanosomes to opisthokonts. Parasitology. 2021;148(10):1196–218.
65. Parenteau J, Abou ES. Introns: good day junk is bad day treasure. Trends Genet. 2019;35(12):923–34.
66. Girardini KN, Olthof AM, Kanadia RN. Introns: the "dark matter" of the eukaryotic genome. Front Genet. 2023;14:1150212.
67. Yurchenko V, Kostygov A, Havlová J, Grybchuk-Ieremenko A, Ševčíková T, Lukeš J, et al. Diversity of trypanosomatids in cockroaches and the description of *Herpetomonas tarakana* sp. n. J Eukaryot Microbiol. 2016;63(2):198–209.
68. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
69. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010. Online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

70. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
71. Skalický T, Dobáková E, Wheeler RJ, Tesařová M, Flegontov P, Jirsová D, Votýpka J, Yurchenko V, Ayala FJ, Lukeš J. Extensive flagellar remodeling during the complex life cycle of *Paratrypanosoma*, an early-branching trypanosomatid. 2018. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/PRJNA414522.
72. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
73. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
74. Capella-Gutiérrez S, Silla-Martinez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
75. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
76. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. 2021;49:D458–60.
77. Jarnot P, Ziemska-Legiecka J, Dobson L, Merski M, Mier P, Andrade-Navarro MA, et al. PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. Nucleic Acids Res. 2020;48:W77–84.
78. Yuan D, Ahamed A, Burgin J, Cummins C, Devraj R, Gueye K, et al. The European Nucleotide Archive in 2023. Nucleic Acids Res. 2024;52:D92–7.
79. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
80. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
81. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
82. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
83. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28(4):464–9.
84. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
85. Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. Gigascience. 2018;7(12):giy131.
86. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB38965 (2023).
87. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB60500 (2023).
88. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB60502 (2023).
89. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB60504 (2023).
90. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB60505 (2023).
91. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB39255 (2020).
92. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011;17(1):3.
93. Procter JB, Carstairs GM, Soares B, Mourao K, Ofoegbu TC, Barton D, et al. Alignment of biological sequences with Jalview. Methods Mol Biol. 2021;2231:203–24.
94. Bates PA, Tetley L. *Leishmania mexicana*: induction of metacyclogenesis by cultivation of promastigotes at acidic pH. Exp Parasitol. 1993;76(4):412–23.
95. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.
96. Cavalier-Smith T. Higher classification and phylogeny of Euglenozoa. Eur J Protistol. 2016;56:250–76.
97. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005;21(16):3422–3.
98. Shanmugasundram A, Starns D, Bohme U, Amos B, Wilkinson PA, Harb OS, et al. TriTrypDB: an integrated functional genomics resource for Kinetoplastida. PLoS Negl Trop Dis. 2023;17(1): e0011058.
99. Valach M, Moreira S, Petitjean C, Benz C, Butenko A, Flegontova O, et al. *Diplonema papillatum*, whole genome shotgun sequencing project. 2023. GenBank https://www.ncbi.nlm.nih.gov/nuccore/JAPJBO000000000.1.
100. *Naegleria gruberi* strain NEG-M, whole genome shotgun sequencing project. 2014. GenBank https://www.ncbi.nlm.nih.gov/nuccore/ACER00000000.1.
101. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
102. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7(10): e1002195.
103. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51:D418–27.
104. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 2007;3(4): e65.
105. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.
106. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011;27(13):i85–93.
107. De Rijk P, Wuyts J, De Wachter R. RnaViz 2: an improved representation of RNA secondary structure. Bioinformatics. 2003;19(2):299–300.
108. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
109. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Soding J, et al. Protein sequence analysis using the MPI bioinformatics toolkit. Curr Protoc Bioinformatics. 2020;72(1): e108.
110. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–54.
111. Kostygov A, Skýpalová K, Kraeva N, Kalita E, McLeod C, Yurchenko V, Field M, Lukeš J, Butenko A. *Lafontella* sp. isolate GMO-01 genome sequencing and assembly. 2024. NCBI BioProject https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1102159.

Publisher's Note