Published online 6 December 2017

Nucleic Acids Research, 2018, Vol. 46, No. 2 ***765–781***
doi: 10.1093/nar/gkx1202

Trypanosomatid mitochondrial RNA editing: dramatically complex transcript repertoires revealed with a dedicated mapping tool

Evgeny S. Gerasimov[1,2], Anna A. Gasparyan[1], Iosif Kaurov[3,4], Boris Tichý[5], Maria D. Logacheva[6,7,8], Alexander A. Kolesnikov[1], Julius Lukeš[3,4], Vyacheslav Yurchenko[3,9,10], Sara L. Zimmer[11,*] and Pavel Flegontov[3,6,9,*]

[1]Faculty of Biology, M.V. Lomonosov Moscow State University, Moscow 119991, Russia, [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127051, Russia, [3]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice 370 05, Czech Republic, [4]Faculty of Science, University of South Bohemia, České Budějovice 370 05, Czech Republic, [5]Central European Institute of Technology, Masaryk University, Brno 625 00, Czech Republic, [6]Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow 119991, Russia, [7]Russia Extreme Biology Laboratory, Institute of Fundamental Medicine and Biology, Kazan Federal University, Kazan, 420008, Russia, [8]Skolkovo Institute of Science and Technology, Moscow, 14326, Russia, [9]Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava 710 00, Czech Republic, [10]Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava 710 00, Czech Republic and [11]Department of Biomedical Sciences, University of Minnesota Medical School, Duluth, MN 55812-3031, USA

Received July 31, 2017; Revised October 23, 2017; Editorial Decision November 16, 2017; Accepted November 20, 2017

ABSTRACT

RNA editing by targeted insertion and deletion of uridine is crucial to generate translatable mRNAs from the cryptogenes of the mitochondrial genome of kinetoplastids. This type of editing consists of a stepwise cascade of reactions generally proceeding from 3′ to 5′ on a transcript, resulting in a population of partially edited as well as pre-edited and completely edited molecules for each mitochondrial cryptogene of these protozoans. Often, the number of uridines inserted and deleted exceed the number of nucleotides that are genome-encoded. Thus, analysis of kinetoplastid mitochondrial transcriptomes has proven frustratingly complex. Here we present our analysis of *Leptomonas pyrrhocoris* mitochondrial cDNA deep sequencing reads using *T-Aligner*, our new tool which allows comprehensive characterization of RNA editing, not relying on targeted transcript amplification and on prior knowledge of final edited products. *T-Aligner* implements a pipeline of read mapping, visualization of all editing states and their coverage, and assembly of canonical and alternative translatable mRNAs. We also assess *T-Aligner* functionality on a more challenging deep sequencing read input from *Trypanosoma cruzi*. The analysis reveals that transcripts of cryptogenes of both species undergo very complex editing that includes the formation of alternative open reading frames and whole categories of truncated editing products.

INTRODUCTION

Bizarre genome architectures, surprisingly large molecular complexes, and intricate RNA processing pathways in living cells have likely evolved via genetic drift and subsequent random walk through a complexity space (1–5). Mitochondrially-encoded RNAs of the euglenozoan clade of protists are processed in some of the most complex ways identified to date. These include *trans*-splicing of multiple transcript modules to forge a complete transcript, insertion of multiple uridines (U) between the modules, and numerous A-to-I and C-to-U deaminations in diplonemids (6–9). In kinetoplastids, which include trypanosomatids, RNA processing includes abundant uridine insertions and deletions in mitochondrially-encoded transcripts (U-indel editing) (10,11).

This targeted insertion and deletion of Us to generate the proper code for translation into protein is required for most protein-coding transcripts of kinetoplastid mitochondrial genomes. Editing even supplies Us for start and stop

*To whom correspondence should be addressed Sara L. Zimmer. Tel: +1 218 726 6741; Fax: +1 218 726 7906; Email: szimmer3@d.umn.edu
To whom correspondence should be addressed Pavel Flegontov. Email: pavel.flegontov@osu.cz

codons for some mRNAs (11). Some transcripts are edited over their entire length at hundreds of sites (pan-edited), while others are edited within limited domains, or not edited at all (12–16). U-indel editing utilizes short antisense RNAs termed guide RNAs (gRNAs) as templates for the editing reactions that initiate when the first gRNA binds to the unedited transcript immediately downstream of the region requiring editing. The middle part of the gRNA directs editing at multiple nucleotide positions or 'sites' just upstream of the gRNA-bound region. The 'RNA Editing Core Complex' or RECC (10,11,17) performs editing, with several other complexes orchestrating RECC activities and additional RNA processing (10,11,18). The edited mRNA region that was directed by the first gRNA serves as the binding sequence for the second gRNA. This process repeats until editing is completed; therefore, the editing process is generally 3′ to 5′ directional. Hundreds of gRNAs are needed to 'decrypt' all mitochondrial transcripts, and kinetoplastid genomes retain far more gRNAs than the minimum required to execute editing (19–21).

U-indel editing results in mRNAs containing U insertions and deletions in patterns deviating from known final editing products. Most copies of a specific transcript are neither the nascent transcript nor the fully-edited and translationally competent form of the mRNA. Instead, most cryptogene mRNAs are incompletely edited (12,22–24) and presumed to be in the process of editing (editing intermediates). Intermediates typically possess 'junction regions' between edited and pre-edited portions of the mRNA where alternative U addition and deletion patterns are observed (22–24). It is hypothesized that junction regions represent temporary editing states of editing intermediates that will resolve to canonical sequences prior to gRNA exchange. Alternatively, or additionally, transcripts containing junctions could represent abortive products of editing.

Conversely, binding of incorrect, non-cognate gRNAs could lead to 'misediting', and even editing guided by cognate gRNAs could be performed imperfectly to generate a portion of unproductive, dead-end intermediates that prevent the binding of the correct subsequent gRNAs (25–28). It is also possible that multiple RNAs coding for different protein sequences could be generated from a single cryptogene (29–31). Deep sequencing of gRNAs (19,20,32) and targeted deep sequencing of editing intermediates (12,18,24) have revealed both non-productive editing events and potentially functional alternative editing.

Given the complexity of U-indel editing, it may seem surprising that such a process exists at all. The constructive neutral evolution model (1,3,4,6,33,34) hypothesizes that complex RNA processing events exist because of an initial gene mutation restoration by then-existing enzymes and short antisense RNAs. This initial restoration event made the involved short RNAs and enzymatic tools indispensable and precipitated fixation of further mutations to be corrected in a similar manner. Other evolutionary explanations that invoke selective advantage have been proposed for the emergence of U-indel RNA editing (35–37). Evolutionary theories can be evaluated by analyzing similarities and differences of RNA processing events in organisms utilizing them. In the case of kinetoplastid mitochondrial U-indel editing, however, the complexity of the edited transcriptome makes such an approach particularly challenging. The major difficulty in establishing the genome-wide range and type (i.e. productive or non-productive) of U-indel editing has been the ability to capture and catalogue editing products. This is particularly true for the majority of kinetoplastids, for which final edited products have not been painstakingly determined one-by-one as they largely have in *Trypanosoma brucei* and *Leishmania* spp. (e.g. 38–41).

The current study builds on our exploration of U-indel RNA editing in the early-branching kinetoplastid *Perkinsela* whose mitochondrial genome encodes only six mRNAs, all edited. We were able to map shotgun U-indel edited reads on *Perkinsela* mitochondrial genes, but only those overlapping a certain 'anchor region' (12). Other teams used a similar approach in *T. brucei* to map short (50 nt) amplicon-derived reads with the software package PARERS (42), or mapped longer amplicons onto entire short cryptogenes with the tool TREAT (18,24). PARERS requires 'anchor sequences'; TREAT relies on completely sequenced amplicons that cover a specific edited region, which precludes analysis of long pan-edited cryptogenes. Additionally, these two approaches require prior knowledge of the fully-edited mRNA sequence. Considering that in the sequencing era, the number of analyzed kinetoplastid protists will undoubtedly grow fast (43–45), it was essential to develop a tool to analyze U-indel RNA editing in non-model species for which no prior data is available except the mitochondrial genome sequence, a commodity easy to obtain.

T-Aligner is the new read mapping and assembly tool with which we tackled these mapping, organizational, and reporting challenges. *T-Aligner* does not rely on PCR amplification of specific transcripts or prior knowledge of canonically edited sequence. Instead, it assembles multiple potential edited open reading frames (ORFs) from shotgun reads mapped to each cryptogene. *T-Aligner* users identify a most likely final edited product(s) in ways we demonstrate here for the *Leptomonas pyrrhocoris* (44) U-indel edited transcriptome, for which fully-edited mRNAs have not yet been identified. *Leptomonas pyrrhocoris* is an emergent model species (46) of the monoxenous trypanosomatids parasitizing insects (45).

We reconstructed RNA editing intermediates and final products (complete mRNAs) for all expressed *L. pyrrhocoris* cryptogenes. Even on a more challenging dataset from *Trypanosoma cruzi* with shorter reads and substantially lower percentages of edited reads, *T-Aligner* successfully reconstructed some long fully-edited mRNAs with up to 219 edited sites. In both these distantly related species, U-indel RNA editing appears to be of low fidelity. Multiple truncated and rare alternative products were observed within each transcriptome. Alternative edited products also appear that are more common; for example, inclusion of two alternative amino acids or an alternative 3′ end among *L. pyrrhocoris* ND8 ORFs. An interesting discovery was that reads covering the 5′ end of the RPS12 transcript from both species could be assembled into translatable products containing N-termini of different sequences and lengths similar to what was found in *T. brucei* (18). Thus, *T-Aligner* has allowed the discovery of potential organism-specific mi-

tochondrial genome products. Our results, placed alongside those from *Perkinsela* (12) and *T. brucei* (18,24,37,42), demonstrate that U-indel RNA editing is universally an extremely chaotic and 'noisy' process with ample possibility for functional alternative products. These alternative products could be informative (see, e.g. (37)) when considering the merits of constructive neutral evolution versus conferral of selective advantage as the reason this complex RNA editing process appears in nature.

MATERIALS AND METHODS

Cell culture

Leptomonas pyrrhocoris, isolate H10 (44), was grown in RPMI-1640 medium with or without glucose (R6504, R1383, Sigma-Aldrich) supplemented with 20% heat inactivated fetal bovine serum (PAA Laboratories), 2.0 g/l sodium bicarbonate and 1% penicillin-streptomycin solution (HyClone) at 27°C. Cell cultures reaching a density 2–5 × 10^7 cells/ml were passaged daily. *Trypanosoma cruzi* strain Sylvio X10 epimastigotes acquired from ATCC® (ATCC 50823) were grown in liver infusion tryptose medium (LIT) (47). Cultures were allowed to grow for 2–3 days in fresh medium to reach the final density of 2–5 × 10^7 cell/ml.

Extraction of mitochondrial vesicles

To enrich for *L. pyrrhocoris* mitochondrial transcripts, we first isolated mitochondrial vesicles in a Percoll gradient as described previously (48) with the following modifications. Cells were grown in 4 L of the standard and 4 L of the glucose-free medium. Flagellates were collected by centrifugation at 3000×g for 15 min at 4°C and the resulting suspensions were homogenized in DTE buffer (1 mM Tris–HCl, pH 8.0; 1 mM EDTA) in a Dounce tissue grinder with five movements of a pestle. Mitochondrial vesicle pellets were resuspended in RNA Blue reagent (Top-Bio) for RNA isolation.

For *T. cruzi*, crude mitochondrial preparations were generated (49), with the following modifications due to intrinsic resistance of *T. cruzi* to hypotonic lysis: cells were incubated 10 min on ice in 0.65× DTE prior to 15 passages through a 25 G needle. TRIzol reagent (Thermo Fisher Scientific) was directly added to pelleted mitochondrial vesicles for RNA isolation.

RNA extraction, cDNA synthesis and qPCR analysis

L. pyrrhocoris mitochondrial RNA in RNA Blue reagent was isolated according to the manufacturer's protocol, and poly(A) enrichment was performed on 50% of each collected sample with the Dynabeads mRNA Purification Kit (Ambion). RNA was quantified using the Qubit RNA BR Assay Kit and instrument (Invitrogen), with RNA integrity confirmed by electrophoresis on a denaturing agarose gel. Enrichment of mitochondrial transcripts was assessed by an RT-qPCR analysis (50) and the following primer sets: 9S small-subunit rRNA (5′-CACCATGAAAAGGCTAAGGAA-3′, 5′-AATTGGTGGGCAACAATACC-3′), ND1 (5′-CATTGTTCGCATAGCCGATA-3′, 5′-CGCGATGTTCATTACCAGTTT-3′), ND5 (5′-GGCTACAAGATATCCGCTGCT-3′, 5′-AAAGCCGCATAAGAACCAAA-3′); and for three abundant nuclear transcripts, 18S small-subunit rRNA (5′-ACCAAGACGAACTACAGCGA-3′, 5′-GTTTGCAGCGTGGACTACAA-3′), β-tubulin (5′-TACTGCTGGTACTCGGACAC-3′, 5′-GTTCATCGGCAACAACACCT-3′) and gGAPDH (5′-CGCTGATCACGACCTTCTTC-3′, 5′-GAGGTGAAGAAGCCGGATGT-3′). Values were normalized to those of 18S rRNA. Comparison of the ratio of selected mitochondrial and nuclear transcripts in obtained fractions with whole cells and isolated mitochondria showed ∼10-fold and 2-fold enrichment with mitochondrial transcripts, for cells grown in media with and without glucose, respectively (data not shown).

In addition to mitochondrially-enriched *T. cruzi* RNA, total RNA samples were prepared from 4 ml of the starting culture using the TRIzol reagent. Ten micrograms were DNase treated in a 50 μl volume using the Ambion DNA-free kit, after which 3 volumes of TRIzol reagent were added and the DNase-treated RNA was purified with the Direct-zol RNA miniprep kit (Zymo Research), while 5 μg of mitochondrially-enriched RNA was DNase treated on the Direct-zol column according to the manufacturer's protocol. RNA quality was verified by the sequencing facility.

Transcriptomic library construction, sequencing, and read processing

L. pyrrhocoris RNA-seq libraries were prepared with the NEXTflex Directional RNA-Seq Kit (dUTP-Based) (Bioo Scientific). Two libraries were generated from poly(A)-enriched mitochondrial RNA (derived from the standard or glucose-free culture), and another two libraries from total mitochondrial RNA. RNAs were treated with DNase (TURBO DNA-free Kit, Ambion) prior to cDNA library synthesis. Libraries were prepared according to the manufacturer's protocol, but the volume ratio of paramagnetic beads vs. sample was adjusted to 0.7:1, to enrich for 300–400 nt RNA fragments. Size distribution of cDNA was checked on a 2% agarose gel, and the libraries were quantified using qPCR with adapter-specific primers. The libraries were sequenced on the MiSeq platform (Illumina), generating 25.5 million 300 nt reads in pairs. As uridine as well as adenine are nucleotide components of the non-templated tail on trypanosomatid mitochondrial mRNAs (51,52), *T-Aligner* users should be cautioned that differences in recovery during poly(A) enrichment between transcripts potentially exist.

L. pyrrhocoris RNA was extracted from both culture variants and sequenced independently. The data for both culture variants were subsequently merged in order to attain maximal coverage as we chose to not focus on differential gene expression patterns. Poly(A)-enriched and total RNA-derived libraries were treated separately. Each read set underwent trimming of low-quality regions and adapters using CLC Genomic Workbench v. 6.5.1 with the following settings: error probability threshold, 0.01; no more than one undetermined position (N) per read; minimal read length of 50 nt. Then reads in pairs were merged using CLC Ge-

nomic Workbench default settings into longer pseudoreads with modal length of 250 nt, and ranging from about 100 to 500 nt in length. Finally, merged and non-merged reads mapping on a high-quality assembly of the *L. pyrrhocoris* nuclear genome (46) were removed, with the mapping performed using CLC Genomics Workbench v. 6.5.1 and these settings: at least 90% of the read aligned, with identity of at least 80%. Performance of *T-Aligner* was tested on the following datasets, prepared for both poly(A)-enriched and total RNA libraries: paired reads prior to the merging step; merged reads with nuclear genomic reads filtered out; merged reads combined with non-merged paired and orphan reads (Supplementary Table S1).

One half of one μg of total and mitochondrial RNA from *T. cruzi*, in biological duplicate, was used by the University of Minnesota Genomics Center for generation of cDNA libraries following the TruSeq® Stranded mRNA Sample Preparation including the poly(A) mRNA enrichment step. The library was gel-purified in order to size select fragments from 270 to 340 nt for collection of cDNAs containing inserts of a 200 nt target size. 2 × 125 cycle sequencing was performed in High Output Mode on a HiSeq 2500 (Illumina). Read yields for the four samples (two total RNAs and two mitochondrial RNA-enriched RNAs) ranged from 19 million to 22.5 million. All 4 samples were pooled for *T-Aligner* analysis and processed by CLC Genomics Workbench v. 9.5.2. using the same approaches as for *L. pyrrhocoris*; however, trimming of poor-quality regions was not performed prior to pairwise merging. *T. cruzi* Sylvio X10 genome assembly (TriTrypDB (53)) was used for depleting reads derived from the nuclear genome. Thus, one read set was prepared for *T. cruzi*: all libraries combined, merged reads with nuclear genomic reads filtered out, modal length of 150 nt, length range from 125 to about 250 nt.

Reconstruction of edited products using *T-Aligner*

T-Aligner is written in C++ and uses the Qt library. The *T-Aligner* alignment algorithm is based on T-less reads and reference sequences where all Ts are masked. A T-less profile consists of a string in the 3-letter alphabet of length **m**, where **m** is the number of A/G/C in a given sequence, and an array of the same length, storing the number of Ts following each position. For mapping, T-less read profiles are broken into short *k*-mers called seeds (the seed length and seed step can be adjusted via command line options). Seed locations are determined using a hash-table search. Extension of matched seeds is done either by connecting neighboring seeds or via letter-by-letter elongation of the alignment in both directions until the number of mismatches exceeds a threshold (set by another command line option, one mismatch per alignment by default). The longest alignments with fewest mismatches are reported for each sequence. Each mismatch gives an alignment length penalty equal to seed length.

Mapping reads on a given cryptogene results in an alignment matrix with rows representing reads mapped and columns representing the numbers of Ts inserted or deleted at each cryptogene position. *T-Aligner* generates a graphical representation of the alignments: a dot matrix showing editing states (number of Ts) at each site. The reference sequence in the dot matrix corresponds to a row of **m** dots, where **m** is the length of the T-less reference. If at the position *i* the read has an insertion of two Ts, a dot is put in the *i*-th column of the dot matrix two rows above the reference row. If at the *j*-th position the read has one T deleted—a dot is put in the *j*-th column one row below the reference row. Thus, each dot represents an editing event supported by at least two reads. An individual read alignment can be traced by connecting the dots, but drawing all mapped reads is prohibitively complex. Instead, relative read coverage (normalized on a site by site basis) is shown with a grey-black-blue gradient for editing states: a blue dot represents the most frequent insertion/deletion; a black dot, the next most frequent modification; and grey dots represent modifications of lesser frequency (Figures 3A, 4A and 5A). To depict the reference-wide coverage of editing states, a 'cloud' diagram (Figures 3D, 4D and 5D) is plotted without normalization: this plot is constructed by scattering **n** dots around position (x, y) randomly within a circle, where **n** is the total number of reads passing through the (x, y) cell of the matrix.

To reduce time and space complexity of ORFs assembly, *T-Aligner* merges into super-alignments either exactly matching read alignments or alignments which are included in some longer alignments as exact substrings. Building super-alignments is especially advantageous when there are many reads that exactly match the reference sequence. Also super-alignments can help to reduce possible sequencing errors by construction of consensus sequence (like OLC assembler does) supported by most of the reads joined.

For the purpose of ORF reconstruction, an overlap graph is built, with nodes corresponding to super-alignments, and edges corresponding to read overlaps of minimal length **g**. The parameter **g** can be adjusted to obtain more reliable overlaps or to process shorter reads. An algorithm based on the breadth-first search (BFS) is used to find ORFs in the graph. All nodes corresponding to reads with possible start codons are used to initiate BFS for the longest possible ORF starting at the node. A general overview of *T-Aligner*'s workflow provided in Supplementary Figure S1, and its default settings are summarized in Supplementary Table S2.

We use 'distance', 'support', and 'relative abundance' metrics for reconstructed products, and output two variants of each metric, an mRNA-based and an ORF-based variant, and mRNA-based variants are described below. The distance metric (abbreviated 'mRNA-ES_Dist') is calculated for mRNA **x** in relation to the canonical mRNA **y** (set with a dedicated option or the mRNA containing the longest ORF, by default) as the number of sites where **x** has editing states different from those in **y**. Another version of this metric (abbreviated 'mRNA-Dist') considers only sites where **x** has editing states different from those in **y** and from those in the cryptogene reference (non-reference editing states). Support represents the number of reads that can be aligned on the mRNA with an exact match search algorithm. There are two versions of this metric: based on all reads (abbreviated 'sup_total'), or on edited reads only ('sup_edited'). Relative abundance, abbreviated as 'mRNA-ES_ratio' is calculated in the following way: (i) take two mRNAs, the canonical one and an alternative one; (ii) take a set of overlapping sites with different editing states; (iii) in this subset of sites, find the least supported editing state in the

alternative mRNA and the least supported editing state in the canonical mRNA, denote coverage of these states as A and B, respectively; 4) divide A by B.

Since conditions for ORF reconstruction are often far from optimal on real data due to low percentage of fully-edited reads, short read length or extremely high coverage overwhelming the algorithm, all three *L. pyrrhocoris* read sets (paired reads prior to the merging step; merged reads with nuclear genomic reads filtered out; merged reads combined with non-merged paired and orphan reads) were alternatively used as input for *T-Aligner*, and a range of seed step and seed length settings was tested for both species. The following settings were optimal for most *L. pyrrhocoris* maxicircle loci: seed length, 10; seed step, 10. The following settings were optimal for the *T. cruzi* loci analyzed: seed length, 24; seed step, 12 (Supplementary Table S2). The best dataset and *T-Aligner* setting were chosen on a gene by gene basis, guided in part by BLASTP hits to homologous translated mRNAs. Since *L. pyrrhocoris* is related to *Leishmania* spp. (46), *Leishmania* predicted proteins were used for constructing a BLAST database, and proteins across all kinetoplastids were used for optimizing ORF reconstruction in *T. cruzi*. Hit coverage and bit score were crucial parameters we tried to maximize. For the ND8 and RPS12 cryptogenes, 'gold standard' translated mRNA sequences were chosen among the published ones (54,55), and BLASTP results were verified on those sequences.

T-Aligner was also used for assessing expression of maxicircle loci across libraries and datasets (Supplementary Table S1). Expression values were calculated as reads per one kbp of fully edited mRNA sequence per one million reads mapped on the maxicircle coding region (RPKM).

Mitochondrial genome assembly and annotation

The coding region of the circular mRNA- and rRNA-encoding *L. pyrrhocoris* mitochondrial genome was assembled by CLC Genomics Workbench v. 6.5.1 from paired-end Illumina HiSeq reads (two libraries, 430 nt and 900 nt average insert size, 100 nt raw read length) generated elsewhere (46). Annotation of genes and cryptogenes was performed manually according to similarity with various *Leishmania* and *Leptomonas* species. In the case of *T. cruzi* Sylvio X10, we used a published annotation of the maxicircle (56).

Read mapping

Cryptogene sequences with flanking regions of 100 nt were used as references for read mapping with *bowtie2* v.2.10 (57), *bowtie2-mod* v.2.0.2 (12), and *T-Aligner*. Run options for *bowtie2-mod* were as follows: '--score-min L,0,-2 -D 20 -R 3 -N 1 -L 14 -i S,1,0.50 --rfg 10,10 --rdg 10,10 --mp 18 --rfg-T 1,1 --rdg-T 1,1 --end-to-end --gbar 0 --dpad 50'. *Bowtie2* was run with the '--sensitive' and '--end-to-end' options. For mapping performance tests (Figure 1B), *T-Aligner* was run with the '--mf 0.75 --mr 16 --sl 20 --ss 10 --mm 1 --xi 16' options and poly(A)-enriched paired reads as input (library 1A, Supplementary Table S1); for ORF reconstruction parameters were adjusted individually to get most reliable results and are summarized in Supplementary Table S3.

RESULTS

T-Aligner effectively maps *Leptomonas pyrrhocoris* maxicircle transcripts

The coding portion of the *L. pyrrhocoris* maxicircle was assembled from reads generated previously (46), and is shown in Figure 1A. Eighteen protein-coding genes and 2 rRNA genes are clustered on 17 kb of the ∼27 kb maxicircle molecule. Twelve of the mRNAs undergo editing, with six of them pan-edited. This maxicircle cryptogene editing pattern is almost identical to that found in some related *Leishmania* spp. (21) and is similar to that of the most extensively studied *T. brucei*.

Reliable mapping of cDNA sequencing reads with multiple closely spaced insertions/deletions is not possible with general-purpose read mappers, which are not optimized to take into account the complexity of the RNA editing process. Therefore, we used *T-Aligner* (see Supplementary Figure S1 for a workflow) to undertake an overview of *L. pyrrhocoris* maxicircle mRNA expression. We used paired-end Illumina MiSeq reads (read length 300 nt before trimming) from poly(A)-enriched and total cDNA libraries derived from purified mitochondrial vesicles. Various read processing protocols are described in Methods and statistics for read datasets provided in Supplementary Table S1. Expression of maxicircle loci measured in reads per 1 kbp of fully-edited mRNA sequence per 1 million reads mapped on the maxicircle (RPKM) was similar in total and poly(A)-enriched libraries, but COII, CYb, ND2, ND4 and RPS12 expression was over 2-fold higher in the poly(A) fraction, and expression of A6 and ND5 was over two-fold higher in the total library (Supplementary Table S1).

These differences potentially reflect a bias in the determination of relative abundances of mRNAs in the poly(A) selected population. Uridine as well as adenine are nucleotide components of the non-templated tail on mitochondrial mRNAs, and percentage of uridine as a component of tails can vary by transcript (51,52,58) which could affect relative transcript recovery. Non-templated tailing of kinetoplastid mitochondrial mRNA is mainly unstudied outside of *T. brucei*, so it is possible that this potential bias may be species-specific and it does not play a role here. As absolute maxicircle transcript coverage was on average ∼20 times lower in the total library (Supplementary Table S1), only the poly(A)-enriched library was used in *T-Aligner* editing reconstructions.

Reads derived from all 20 maxicircle loci were detected in the poly(A) library (see read per nucleotide expression values in Figure 1B and RPKM values in Supplementary Table S1). These include exact matches to the *L. pyrrhocoris* maxicircle loci as well as reads containing U insertions and/or deletions. Maxicircle genes G3, MURF2 and MURF5 can be considered nearly silent, as few reads were aligned on those loci, while ND8, COI, COII, ND9 and COIII reads were highly abundant (Figure 1B, Supplementary Table S1).

The number of reads mapped on a reference gene may be used as a measure of mapping algorithm performance. Figure 1B shows the expression level of maxicircle genes estimated with three algorithms. We compared *T-Aligner* performance to that of *bowtie2,* a general-purpose mapping

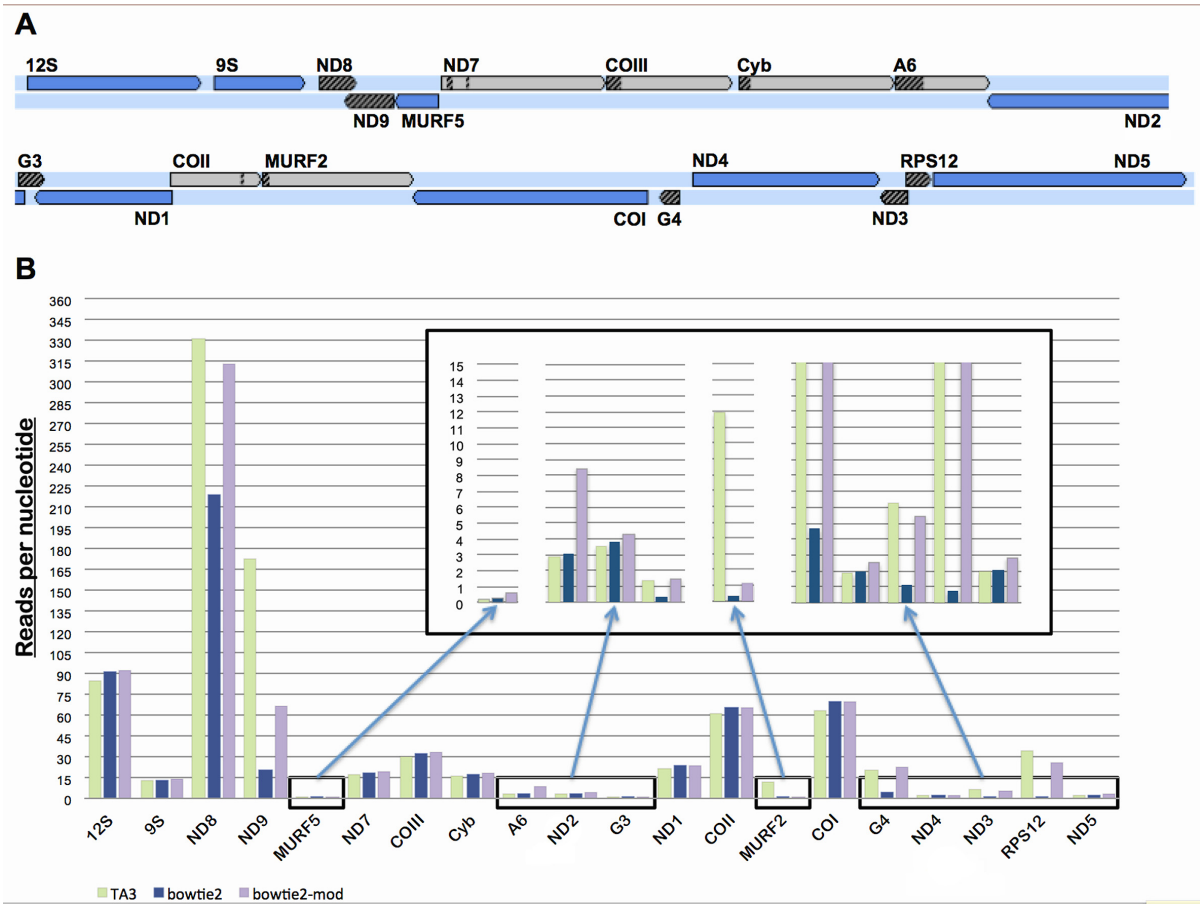

Figure 1. (**A**) A gene map of the coding region of the *Leptomonas pyrrhocoris* mitochondrial maxicircle. Edited transcript regions are hatched, and never-edited transcripts are shown in blue. (**B**) Read mapping performance compared for *bowtie2* (blue bars), *bowtie2-mod* (purple bars) and *T-Aligner* (green bars) on all genes. Read mapping performance was quantified in reads per one nucleotide of the genomic sequence (100 nt flanks on both sides were also counted). The inset shows genes expressed at a low level.

algorithm, and *bowtie2-mod* that was optimized for mapping reads containing multiple T-indels (12), which are U-indels reflected in the cDNA read population. Manual inspection of alignment quality showed that for recognizing reads of the six pan-edited transcripts, *T-Aligner* is more accurate (produces more correct alignments) and more sensitive (reports more alignments) than both *bowtie2* versions. *Bowtie2* performs well on the never-edited genes (e.g. ND4, COI, 12S large-subunit rRNA), but fails on pan-edited transcripts even with relaxed settings. While *bowtie2-mod* can detect pan-edited transcripts, it produces many incorrect alignments containing excessive mismatches, T/N mismatches instead of T-indels, and A/G/C-gaps. In summary, *T-Aligner* is best-equipped to accurately detect maxicircle products of various abundances within a shotgun read population.

T-Aligner reconstructs the limited edited domains of transcripts A6, COII, COIII, CYb, MURF2 and ND7

Overall transcript abundance is only an initial foray into the understanding of a U-indel edited transcriptome. A complete maxicircle transcription analysis must include the whole spectrum of editing intermediates. We accomplished this with *T-Aligner's* algorithms and its gene-by-gene visualization engine. The *T-Aligner* ORF tracing algorithm generates ORFs meeting length thresholds and provides multiple statistics that allow evaluation of whether an assembled product is likely to be a mature mRNAs, or rather an unproductively edited mRNA or unfinished intermediate.

Analyzing *L. pyrrhocoris* and *T. cruzi* cryptogene editing with *T-Aligner* requires defining new terms in order to convey findings. We define an edited site as the interval between adjacent A/G/C bases that contains a U insertion/deletion in at least one read; therefore, an edited site can only be definitely established post-analysis. Edited sites achieve differing frequencies with which they are edited within the read population. Sites can be classified into minor and major sites based on the degree to which they are edited. While a single edited read percentage cut-off to classify sites as major or minor may seem desirable for inter-transcript comparisons, editing levels differ too widely to use a single percentage. We use distinction between major and minor edited sites when comparing multiple editing pathways within transcripts, so we define cut-offs for major and minor site edited read percentages for each transcript separately.

We define the assembled ORF that, when translated, is most similar to putative protein products of homologous

fully-edited mRNAs as 'canonical', created by the canonical editing cascade (pathway). All other reconstructed products are named 'alternative'. An edited site is termed canonical if it is edited in the canonical product, and termed non-canonical if it is not. In contrast to edited *site*, the editing *state* is defined as the number of Us at an edited site. Both insertion/deletion editing states at non-canonical sites and editing states at canonical sites that do not match those of the canonical product are considered 'alternative editing'. A site is termed 'alternatively edited' if its state is different from that in the canonical product. Importantly, use of the term 'alternatively edited' in this study carries no connotation of either potential to be translated or editing error. While earlier studies using the term 'alternative editing' did not confine its use to translatable mRNAs only, by and large they used them mainly in reference to such products (24,29–31). As the ultimate fate of almost all 'alternatively edited mRNAs' are empirically unknown, our definitions of 'alternative' make us less prone to bias and assumptions. We also note that only some alternative products are seemingly incompatible with the canonical editing pathway, i.e. at one or more sites contain a longer insertion/deletion as compared to the canonical product, or a deletion instead of insertion and *vice versa*.

Six *L. pyrrhocoris* maxicircle transcripts (A6, COII, COIII, CYb, MURF2, ND7) have short editing domains, making them ideal targets to initially test *T-Aligner* ORF assembly. ND7 in the Leishmaniinae subfamily (e.g. genera *Leishmania*, *Crithidia*, *Leptomonas* (59)) possesses two edited domains near its 5′-end (Figure 1A). Editing of each domain is directed by a single gRNA, with spacing such that each gRNA can potentially bind and act independently (21). Figure 2A displays *T-Aligner* editing reconstruction of both the ten-site upstream and three-site downstream domains of *L. pyrrhocoris* ND7 (full gene reconstructions for all edited cryptogenes containing limited edited domains and for never-edited maxicircle loci are available in Supplementary Figure S2).

Putative translation of the reconstructed canonical fully-edited (mature) ND7 ORF (depicted by the red line in Figure 2A) shares 96% identity across 97% of its length (Supplementary Table S3) with the *Leishmania tarentolae* edited ND7 mRNA translation, strongly suggesting an accurate reconstruction. *T-Aligner* reconstructed additional ORFs including that represented by the blue line in Figure 2A that is edited only in the second domain and includes three non-canonical editing sites (Figure 2A; arrows). As just one read 'passes through', or supports, these three non-canonical editing sites, we interpret these events yielding this ORF as editing errors or 'noise'. Other minor non-canonical editing sites are scattered along the entire ND7 transcript (Supplementary Figure S2A) including those used to reconstruct the ORF in green in Figure 2A. Minor sites are also present in reconstructions of never-edited mRNAs and other cryptogenes with limited regions of editing. In summary, editing of *L. pyrrhocoris* ND7 appears as expected.

T-Aligner's ability to reveal evolutionary and/or mechanistic details of editing is exhibited in its output for COIII. Putative translation of the reconstructed canonical edited product (in red, Figure 2B) shares 88% identity with the *L. tarentolae* translated COIII mRNA over 99% of its length (Supplementary Table S3), yet editing patterns themselves differ between *L. tarentolae* (40) and *L. pyrrhocoris*. This finding supports a scenario where maxicircle gene product sequences are largely preserved by U-indel editing in the face of a rapidly-evolving maxicircle gene sequence (13,60), and lends support to the hypothesis of coevolution of minicircle-encoded gRNA and maxicircle-encoded edited domains.

Alternative minor COIII ORFs were also reconstructed, for example, the two highly divergent ones visualized in Figure 2B and Supplementary Figure S2B (blue and green lines). These ORFs require editing at non-canonical sites (sites #19, 22, 24 in the 'green' product; sites #19, 22, 24, 25 in the 'blue' product) and alternative editing at canonical sites. For instance, in the 'green' product, five rather than four Us are deleted at site #21 and six Us are inserted at site #27. Both COIII alternative products are represented by just one or two reads over the edited region, thus, editing that would result in alternative proteins for this cryptogene is extremely rare. Conversely, as seen in the 'coverage cloud' panel (Figure 2B), the canonical edited product incorporates major editing states at each site excepting site #8. Similar to COIII and ND7, *T-Aligner* reconstructions of the other transcripts with limited regions of editing: CYb (Figure 2C, Supplementary Figure S2C), MURF2 (Figure 2D, Supplementary Figure S2D), and COII (Supplementary Figure S2E) were consistently as expected.

The edited domain of the *L. pyrrhocoris* canonical A6 *T-Aligner* reconstruction is larger than those of the other limited domain-containing cryptogenes. It contains 30 insertion sites and one deletion site (Supplementary Figure S3) and is similar in size and position within the cryptogene to that of *C. fasciculata* (61). Read support for the fully-edited canonical sequence is good (Supplementary Figure S4). Interestingly, a few reads support an ORF containing an additional edited domain of 11 closely spaced minor non-canonical sites rather than the canonical domain (in black, Supplementary Figure S3C, D and Supplementary Figure S4). It could conceivably be encoded by a single minor gRNA species, given that a gRNA template region in *L. tarentolae* covers 12 edited and non-edited sites on average with a standard deviation of three sites, as calculated based on complete minicircle sequences from GenBank (mostly obtained in (21)).

We also investigated the never-edited transcripts (COI, ND1, ND2, ND4 and ND5, see Supplementary Figure S2F–J), to determine whether they exhibited the same low-level editing 'noise' as in never-edited regions of the minimally-edited transcripts. As depicted for COI and ND1 there is potential for alternative ORF generation of very low abundance even on 'never-edited' templates (Supplementary Figure S2F and G); presumably as a result of minor non-cognate gRNA binding. Overall, *T-Aligner* performed as expected in recreating ORFs for minimally-edited and never-edited transcripts, and provided evidence supporting prior hypotheses (13,60).

Editing errors and alternative cascades are revealed in pan-edited transcripts

The *L. pyrrhocoris* transcripts requiring pan-editing are

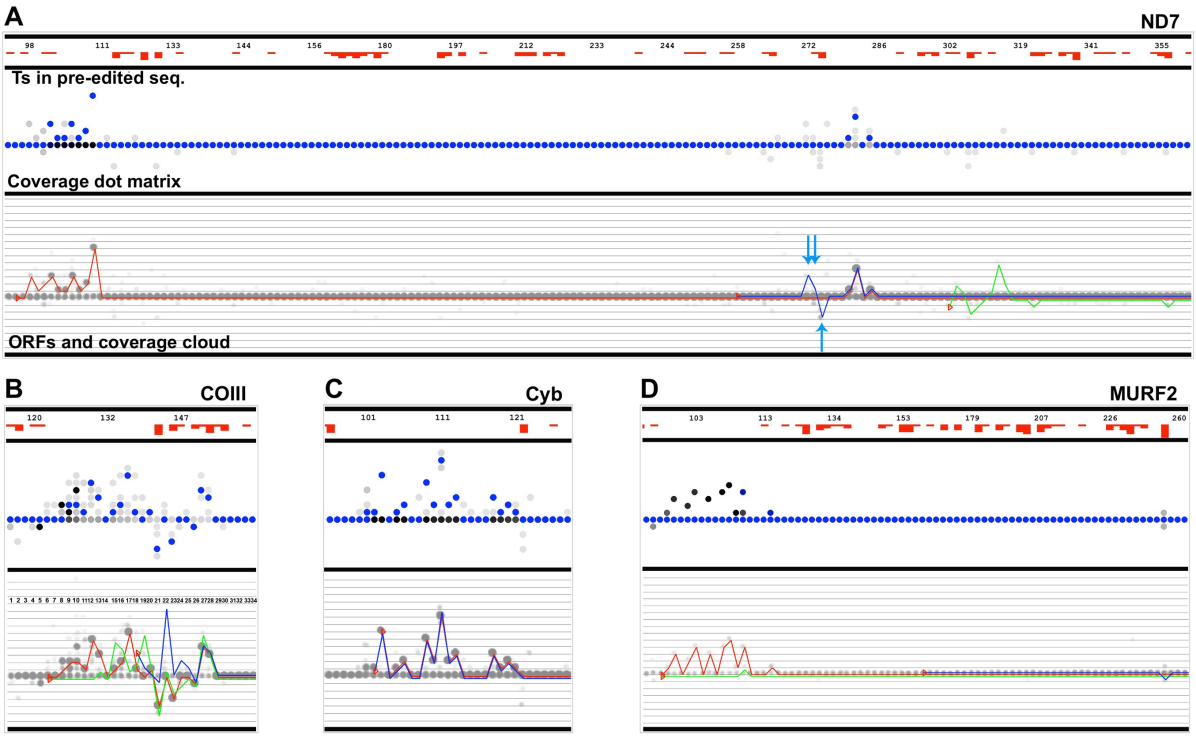

Figure 2. Reconstruction of limited edited domains of the *L. pyrrhocoris* ND7 (**A**), COIII (**B**), CYb (**C**) and MURF2 (**D**) maxicircle cryptogenes. Only 5′ parts of transcripts are shown; for full reconstructions see Supplementary Figure S2. The number of Ts in the genomic sequence is shown with stacked red bars at top. Each center panel shows editing states at each site found in at least two reads (light-grey dots). Proportion of reads supporting an editing state is color-coded, with the blue dot being the most supported and black, the next most supported. Cloud coverage diagrams of editing states at each edited site are shown in the bottom panel. Each read is plotted as a semi-transparent circle. ORFs visualized as a path through editing states are shown, with the red line representing the canonical ORF in each figure. Minor alternative ORFs are represented in blue and green. Red triangles, start codons. For ND7 (A), non-canonical edited sites within the blue ORF are marked with arrows. Editing sites are numbered in COIII (B).

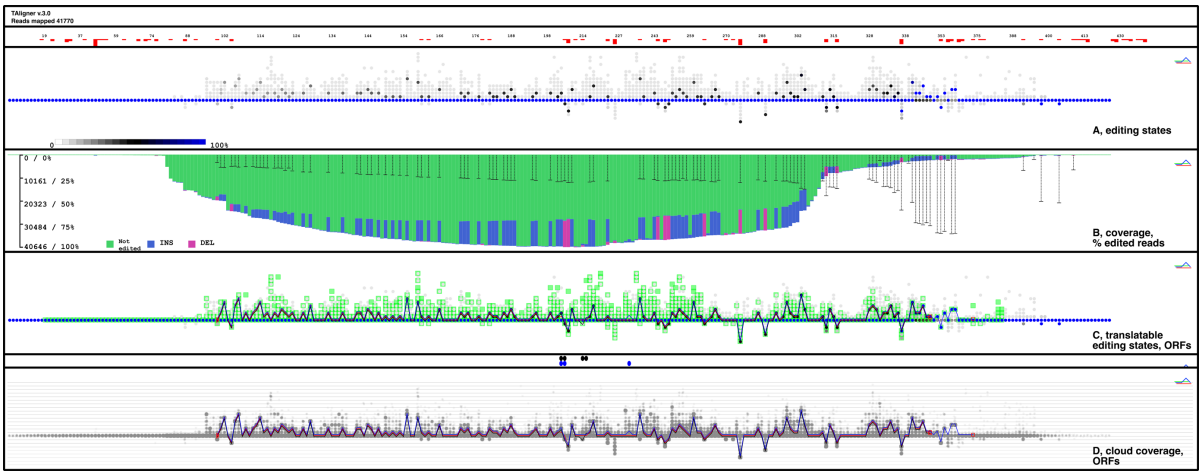

Figure 3. Overview of *L. pyrrhocoris* ND8 transcript editing and reconstruction of the canonical edited product and two abundant alternative products. The number of Ts in the genomic sequence is shown with stacked red bars at top. (**A**) Visualization of editing states found in at least two reads (light-gray dots). Proportion of reads supporting an editing state is calculated for each site separately and color-coded with the blue dot being the most supported and black, the next most supported. (**B**) Absolute coverage bar graph with proportions of insertion edited (blue), deletion edited (pink), or never-edited reads (green). The proportion of edited reads at each site is also shown with black vertical lines. Y-axis values for both metrics are provided. (**C**) ORFs visualized as paths through editing states. Translatable editing states (those included into at least one ORF >60 aa in length) are boxed in green and overlaid with the canonical ORF (red) and alternative ORFs (black and blue). Where editing differs in the black and blue ORF compared with the canonical ORF, it is indicated by dots between panels C and D. Red triangles, start codons; red squares, stop codons. (**D**) Cloud coverage diagrams of editing states at each edited site. Each read is plotted as a semi-transparent dot. The canonical and alternative ORFs shown in C are also plotted.

NADH dehydrogenase subunits ND3, ND8 and ND9; short G-rich transcripts G3 and G4 (most likely coding for other less conserved subunits of the NADH dehydrogenase complex (62,63)), and small subunit ribosomal protein RPS12. The ND8 and ND9 transcripts were highly expressed (>150 reads per one reference sequence nucleotide), RPS12 and G4 were moderately expressed, and ND3 and G3 reads were rare (<10 reads per one reference sequence nucleotide, Figure 1B, Supplementary Table S1). Compared to analysis of cryptogenes with limited editing domains, evaluation of reconstructed ORFs for pan-edited transcripts required additional *T-Aligner* output. *T-Aligner* can display a read coverage map for reconstructed products selected by the user, and these are provided for pan-edited transcripts in the Supplementary Material. Additionally, *T-Aligner* ranks abundance of the multiple generated ORFs relative to the canonical product chosen by the user. The relative abundance metric is based on edited sites shared by the alternative and canonical mRNAs that are edited differently between them. The editing state with the lowest read coverage is found within this edited site subset in the alternative product, the same is done for the canonical product, and from this is derived the alternative/canonical coverage ratio. Such an approach is necessary due to the extremely uneven coverage of some assembled products. Two versions of the relative abundance metric were used, either analyzing sites within full mRNAs, or only within their ORFs. Below we describe reconstruction of fully-edited products for each pan-edited transcript except for G3, for which low coverage precluded ORF assembly.

ND8 – four edited product classes emerge for a single cryptogene. The second longest *L. pyrrhocoris* maxicircle cryptogene, ND8, had by far the highest expression level (Figure 1B). Most ND8 reads are pre-edited, as the proportion of edited reads per canonical site was 31% on average. Based on other studies (e.g. (18)), and given the mechanism of editing progression, we expect the fraction of reads edited at each canonical site (we term this statistic 'editing level') to decrease in the 3′ to 5′ direction. We observe this to varying degrees for nearly all *L. pyrrhocoris* cryptogenes. The level of ND8 editing dropped especially abruptly after the first few sites, and then decreased slowly (Figure 3B).

Edited sites in the ND8 transcript were classified as follows. Major sites were defined as those where >10% of reads were edited. The minor sites were edited in only 1% of reads on average. *T-Aligner* assembled 1389 different ORFs encoding proteins of at least 60 amino acids (aa) from the aligned reads. The canonical edited product (in red in Figure 3C and D), in this case most homologous to translated fully-edited mRNAs in *T. brucei* (54) and *L. amazonensis* (64), was assembled from 502 reads and aligns to the *L. amazonensis* ortholog end-to-end with sequence identity of 74% (Supplementary Table S3). It possesses fairly even coverage across the entire sequence (Supplementary Figure S5) and always utilizes major edited sites. However, it is edited to a state with relatively low read coverage at one of the major sites, and not edited at all at the very next major site (see the two adjacent sites marked with stacked black and blue dots between ORF panels in Figure 3). The one rather than two U insertions followed by lack of a deletion preserves the reading frame and merely results in a single amino acid difference from what would be coded if the major editing state and site had been used (Cys rather than Arg; Supplementary Table S4). Editing that does code for Arg is present in an alternative ORF (in black in Figure 3C and D) with both ORF-based and mRNA-based relative abundance metrics of 0.68. The only other divergence of this abundant alternative ORF from the canonical ORF is the use of two adjacent minor edited sites (rightmost black dots between ORF panels in Figure 3), again conserving the reading frame and resulting in a Ser–Cys substitution.

There were five alternative edited products having both abundance metrics of over 0.3 of the canonical product. One of these products displayed abortive or unfinished editing, and happened to encode a start codon near the 5′ end (encoded protein labeled 'truncated product', Supplementary Table S4). However, a product with an mRNA-based abundance metric of 2.38 and an ORF-based metric of 0.82 (in blue in Figure 3C and D) utilized a single minor edited site near the middle of the transcript (the single blue dot between ORF panels in Figure 3C,D). This shifts the reading frame, resulting in an encoded protein with a C-terminal half that differs entirely from the canonical product and ND8 identified in other species. Additionally, this edited product utilized five major edited sites that are downstream of the canonical product's stop codon. Variants of this alternative edited product that encode N-termini differing in only a few amino acids are also abundant (Supplementary Table S4).

Overview inspection of the ∼200 *T-Aligner* assembled ORFs edited at 90 or more sites (editing at 104 sites is required to generate the canonical product) reveals four product classes based on amino acid alignment (Supplementary Table S4). The canonical class includes the canonical product (in red in Figure 3C and D) and the first alternative (in black in Figure 3C and D). One of the three other classes includes the second abundant alternative product with the divergent C-terminus (in blue in Figure 3C and D). Encoded proteins of these four product classes possess various short indels and amino acid substitutions when compared to other members of the same class. They either possess alternative edits in the middle of the sequence—'bubbles' in the graph of editing pathways, or contain junctions (18,22,24) of aberrant editing between short alternatively edited regions and never-edited regions upstream—'branches' in the graph. Branches of the editing cascade can form at almost any point, and combinations of branches and bubbles are possible. These general types of alternatively edited regions were previously observed in *Perkinsela* (12) and in *T. brucei* (18,24). As we now report these patterns in both *L. pyrrhocoris* and *T. cruzi*, they are apparently common to all U-indel RNA editing systems. Most edited products of each class have extremely low abundance, i.e. they have only single read coverage spanning regions longer than 50 nt. Some of these rare products are alternatively edited at up to 29 sites (only sites with non-reference alternative states were taken into account). In summary, the most abundant putative ND8 edited products differ in editing of only a few sites. Results of these editing divergences span from single amino acid differences to a completely different C-terminus. *T-Aligner* also reconstructs additional products where editing patterns are much different, but these products are rare.

ND9 and ND3—reliable reconstruction with classic patterns of 5′ editing drop-off and truncated editing cascades. The longest *L. pyrrhocoris* cryptogene, ND9, is situated on the maxicircle overlapping with ND8 on the opposite strand (Figure 1). Abundance of reads aligning to ND9 was high (Figure 1B). Very conspicuously, the percentage of edited reads per site fell rapidly in the 3′ to 5′ direction, from 92% at the 3′ end to ∼1% at the 5′ end (Supplementary Figure S6B). Strong drop-offs in editing were also observed for ND3, G3, G4 pan-edited cryptogenes and A6 in *L. pyrrhocoris*, and are even more conspicuous among pan-edited *T. cruzi* cryptogenes. Abrupt drop-offs in editing activity at specific transcript positions ('bottlenecks') can be caused by a rare gRNA (20) or by a difficult-to-resolve mRNA-gRNA pairing. However, for the aforementioned *L. pyrrhocoris* transcripts, the degree of editing within the reads falls off gradually, possibly due to a more general inefficiency inherent in editing. This presents a technical challenge: while the distinction between major and minor edited sites is clear in the 3′ portion of the transcripts, it becomes less obvious as the fraction of edited reads falls.

T-Aligner generated 143 ND9 ORFs that are at least 60 aa in length. The translated canonical ORF shares approximately 70% sequence identity to *Leishmania* spp. ND9 (in red in Supplementary Figure S6C and D). It is assembled from 186 reads (Supplementary Figure S7), and is coterminous with its *Leishmania* ortholog (Supplementary Table S3), a possible indicator of reconstruction reliability. Alternative editing patterns are similar to those observed for ND8. Interestingly, a few non-canonical sites (arrows) approach the canonical ones in their editing levels (Supplementary Figure S6B). These non-canonical editing sites in the middle of the transcript may be edited in short truncated cascades, or may represent prominent junction regions at editing pause sites (22,24). A few reads were assembled into ORFs that possessed a high number of alternatively edited sites; up to 33 alternatively edited sites (possessing non-reference alternative editing states) spanning a total of 52 sites in the reference cryptogene. Editing of a span of this many sites would be expected to require the direction of about four gRNAs, given that a trypanosomatid gRNA is predicted to cover 12 ± 3 edited and non-edited sites. Two examples of such products are shown in black and blue in Supplementary Figure S6C,D, with their read coverage shown in Supplementary Figure S7.

For ND3, the 3′ to 5′ drop-off in editing (Supplementary Figure S8B) is even more of a challenge for *T-Aligner* because the ND3 transcript is present at much lower levels (Figure 1B; Supplementary Table S1). Nevertheless, *T-Aligner* assembled 45 different ND3 ORFs of at least 60 aa in length, most of which again had very low read coverage over domains of divergent editing. The canonical ORF (in red in Supplementary Figure S8C and D and Supplementary Figure S9), when translated, aligned to the *L. amazonensis* ortholog (125 aa long) over its entire length with 42% sequence identity (Supplementary Table S3). The abundant alternative editing states near its 3′ end do not assemble into alternative ORFs; generated reconstructions instead diverge from the canonical product near the middle of the transcript (examples shown in black and blue are provided in Supplementary Figure S8C,D; Supplementary Figure S9). Some of these would require editing by up to six successive alternative gRNAs covering up to 70 sites in the reference.

RPS12—an alternative ORF with a radically different 5′ half. Unusually, we see no noticeable drop in degree of editing from 3′ to 5′ for RPS12 (Figure 4B), which is expressed at moderate levels. This allowed *T-Aligner* to easily reconstruct 295 ORFs of at least 60 aa in length, including the canonical ORF (in red in Figure 4C and D; Supplementary Figure S10) that aligned to and was coterminous with the thoroughly-sequenced fully-edited RPS12 translation of *T. brucei* (18,22,24,55), and its *L. amazonensis* ortholog, with which it shares 62% sequence identity (Supplementary Table S3). The RPS12 canonical *T-Aligner* reconstruction passed through major sites (defined as those edited in >33% of reads) except the 5′-terminal site, which had a much lower editing level, and contained the most common editing states at these sites. However, alternative RPS12 ORFs abound. Again, as with ND8, reconstructed ORFs coalesce into one main and three alternative product classes plus other miscellaneous low-coverage products. A major alternative product (in black in Figure 4C and D) has as much coverage support as the canonical product (Supplementary Table S5 and Supplementary Figure S10), and shares the same product class. It differs only in that it utilizes an alternative editing state near the 5′ end of the editing region, and is not edited at any site upstream of that. The resulting encoded product is N-terminally truncated by five amino acids, and initiates with a different five amino acids before lining up with the canonical product.

A second abundant RPS12 product of interest (in blue in Figure 4C and D; Supplementary Figure S10) results when the very first canonical editing site is not used and two closely spaced major editing sites are not edited (blue dots between panels C and D in Figure 4). The resulting encoded protein is far more divergent. The loss of the single U insertion in the first site completely shifts the reading frame until additional loss of the two downstream insertions half-way through the transcript restores the reading frame. This results in the encoded amino acid sequence shown in Supplementary Table S5 that is the most abundant product of the 'Alternative 1' class of products. Products in other classes were far less abundant than those of the main and 'Alternative 1' class. However, products within these other classes are reconstructed from reads with lengthy portions of alternative editing. Based on the lengths of the alternatively edited regions found in these products (up to 32 alternatively edited sites with non-reference states, spanning up to 58 sites), the divergent cascades may require up to five unique gRNAs each. In summary, the RPS12 product population consists of many very rare products, a few more abundant products of an alternative class that totally diverge in the N-terminal half due to reading frame shifts, and a class of abundant canonical products with alternatives as to the length and composition of the extreme N-terminus.

G4—an assembly challenge for T-Aligner remains in L. pyrrhocoris. Reconstruction of a final edited product was not possible for G4 due to a combination of low read coverage (Figure 1B) and low edited read percentages at the 5′

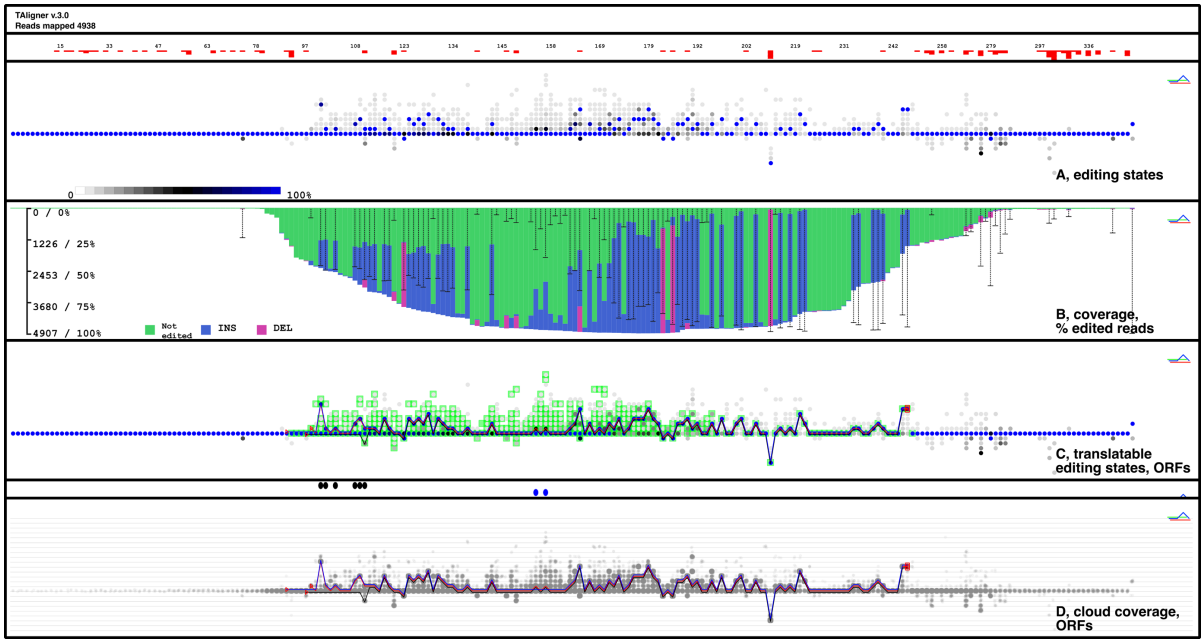

Figure 4. Overview of *L. pyrrhocoris* RPS12 transcript editing and reconstruction of the canonical edited product and two abundant alternative products. The number of Ts in the genomic sequence is shown with stacked red bars at top. (**A**) Visualization of editing states found in at least two reads (light-gray dots). Proportion of reads supporting an editing state is calculated for each site separately and color-coded, with the blue dot being the most supported and black, the next most supported. (**B**) Absolute coverage bar graph with proportions of U-indel insertion edited (blue), deletion edited (pink), or never-edited reads (green). The proportion of edited reads at each site is also shown with black vertical lines. Y-axis values for both metrics are provided. (**C**) ORFs visualized as paths through editing states. Translatable editing states (those included into at least one ORF >60 aa in length) are boxed in green and overlaid with the canonical ORF (red) and alternative ORFs (black and blue). Where editing differs in the black and blue ORF compared with the canonical ORF, it is indicated by dots between panels C and D. Red triangles, start codons; red squares, stop codons. (**D**) Cloud coverage diagrams of editing states at each edited site. Each read is plotted as a semi-transparent dot. The canonical and alternative ORFs shown in C are also plotted.

end. However, even in this challenging situation *T-Aligner* assembled a 5′ truncated ORF encoding a protein of 132 aa (in red in Supplementary Figures S11 andS12) that aligned to *L. amazonensis* 166 aa edited G4 (Supplementary Table S3). As with all other pan-edited cryptogenes, single reads containing alternatively edited sequence for which it was possible to assemble alternative ORFs were found (in black and blue in Supplementary Figure S11C and D and Supplementary Figure S12). Again, some of these were long enough (up to 29 alternatively edited sites with non-reference states spanning up to 47 sites) to require multiple gRNAs for their generation.

T-Aligner reconstructs open reading frames from a more challenging *T. cruzi* read population

Use of *T-Aligner* allowed us to describe the U-indel transcriptome of *L. pyrrhocoris* that is considerably more complex than the 6-gene *Perkinsela* mitochondrial transcriptome analyzed by the modified *bowtie2* and the *T-Aligner* precursor (12). We wanted to test *T-Aligner* performance on a yet more challenging transcriptome: that of the *T. cruzi* maxicircle for which we already had deep sequencing reads available from a different project. Although *T. cruzi* and *L. pyrrhocoris* mitochondrial genomes encode the same products and are completely syntenic, the encoded *T. cruzi* mRNAs are presumably more extensively edited (56,65). Furthermore, we discovered that read coverage for the longer *T. cruzi* pan-edited cryptogenes was uneven, and extremely low fractions of reads were edited near the 5′ end. Finally, a different library preparation and sequencing protocol for *T. cruzi* RNA resulted in much shorter and narrow read range to use as *T-Aligner* input (merged paired reads 125–250 nt versus 100–500 nt).

While some *T. cruzi* editing reconstructions were possible, it was clear that due to the complications described above we would not define the full complement of *T. cruzi* canonical edited mRNAs with the available reads. Therefore, we chose to limit analysis to two *T. cruzi* mRNAs that each satisfy a different experimental goal. *T. cruzi* COIII reconstruction provided us an example with which to verify that the most supported *T-Aligner* product matched a very long product which had been determined independently by traditional means (56). The frequently-analyzed RPS12 (18,22,24) was the choice that would illustrate some of the differences in editing between *L. pyrrhocoris* and *T. cruzi*.

T. cruzi COIII—a very long cryptogene reconstructed by T-Aligner. We took advantage of the published *T. cruzi* edited mRNA sequence for COIII (66) as a standard with which to compare *T-Aligner T. cruzi* COIII assemblies. Remarkably, in spite of editing levels as low as one editing modification per 10 000 reads per site near the 5′ end, the fully-edited canonical ORF was reconstructed for *T. cruzi* COIII, the longest cryptogene in this study (in red in Supplementary Figure S13C and D and Supplementary Figure S14). The 296 aa canonical ORF aligned to the published putative *T. cruzi* COIII protein of 288 aa

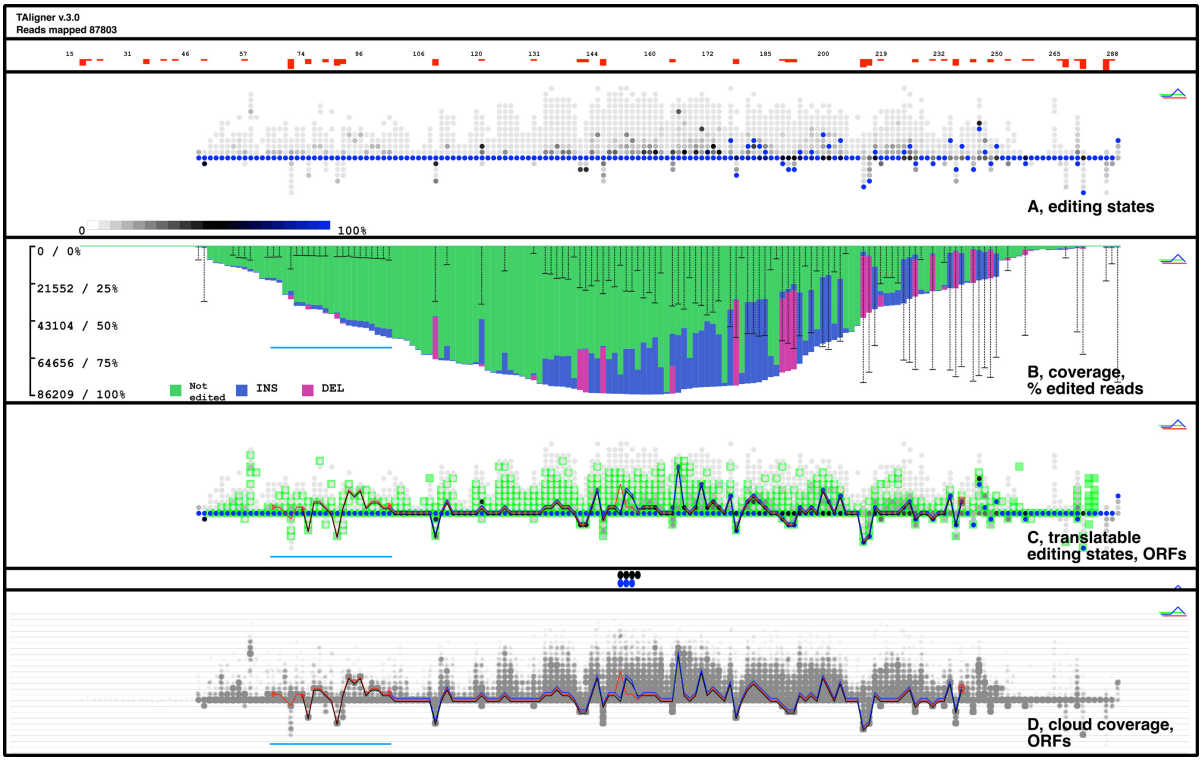

Figure 5. Overview of *T. cruzi* RPS12 transcript editing and reconstruction of the canonical edited product, a well-supported 5′ truncated alternative ORF, and a full-length alternative ORF with an N-terminal shifted reading frame. The number of Ts in the genomic sequence is shown with stacked red bars at top. (**A**) Visualization of editing states found in at least two reads (light-grey dots). Proportion of reads supporting an editing state is calculated for each site separately and color-coded, with the blue dot being the most supported and black, the next most supported. (**B**) Absolute coverage bar graph with proportions of U-indel insertion edited (blue), deletion edited (pink), or never-edited reads (green). The proportion of edited reads at each site is also shown with black vertical lines. Y-axis values for both metrics are provided. (**C**) ORFs visualized as paths through editing states. Translatable editing states (those included into at least one ORF >60 aa in length) are boxed in green and overlaid with the canonical ORF (red) and alternative ORFs (black and blue). Where editing differs in the black and blue ORF compared with the canonical ORF, it is indicated by dots between panels C and D. Red triangles, start codons; red squares, stop codons. (**D**) Cloud coverage diagrams of editing states at each edited site. Each read is plotted as a semi-transparent dot. The canonical and alternative ORFs shown in C are also plotted. Blue horizontal lines in B, C and D indicate a domain of editing that remains unedited in the 'blue' truncated alternative product.

with identity of 99.7% (Supplementary Table S3). The small length difference could be due to strain variability or limited read/sequence input for reconstruction for either this or the previously published sequence. Alternative editing was primarily confined to the 3′ half of the transcripts. However, 10 isolated sites in the 5′ half of the transcript demonstrate unusually high rates of alternative editing, and at some sites in the 3′ half, alternative editing states are more abundant than canonical ones, which was seldom found in *L. pyrrhocoris* read alignments (Supplementary Figure S13A and D). Additionally, ORFs reconstructed of 2–4 overlapping reads with up to 48 alternatively edited sites spanning up to 80 sites in the reference were revealed (in black and blue in Supplementary Figure S13C and D, and Supplementary Figure S14). These are the longest alternative editing cascades observed in this study, potentially encoded by up to seven sequential gRNAs.

T. cruzi RPS12—evidence for an N-terminally truncated protein as a major product. Like RPS12 analysis in *L. pyrrhocoris*, *T. cruzi* RPS12 translated ORF reconstructions also clustered into sequence classes. Although there were many translatable products in these clusters, the vast majority of them again had very low read support (Supplementary Table S6). Unlike the *L. pyrrhocoris* RPS12 assemblies, however, separate sequence clusters were found for the N- and C-terminal portions of *T. cruzi* RPS12, as well as classes with N-terminal truncations.

The reconstructed canonical *T. cruzi* RPS12 putative protein product of 82 aa (in red in Figure 5C and D) is coterminous with other RPS12 proteins and was assembled from only 125 reads (Supplementary Figure S15). It requires utilization of minor editing states at three successive sites in the middle of the transcript (black and blue stacked dots between panels of Figure 5C,D). When aligned to the *T. vivax* (67) or *T. brucei* (55) orthologs, it displays 85% sequence identity (Supplementary Table S3). However, utilization of the successive minor editing states results in a non-conserved Ile at the position occupied by Gln in the orthologous proteins, while use of the major editing states would result in a more conserved Ser at that position. A shorter product just over 60 aa long (in blue in Figure 5C and D) would initiate translation with a downstream Met start codon present in the canonical product. *T-Aligner* reconstructs this product, having astoundingly high abundance metrics (ORF-based, 484; mRNA-based, 258) rela-

tive to the canonical sequence, from 5450 reads (Supplementary Figure S15). The truncated product exhibits major editing state utilization at the three alternatively edited sites in the middle of the transcript (Figure 5). Thus, this ORF encodes an N-terminal truncated version of the canonical protein except that it contains the more conserved Ser amino acid (Supplementary Table S6).

One explanation for this result is that the 5′ region (underlined in Figure 5) of RPS12 is edited in only a small fraction of potentially functional products, while most RPS12 transcripts lack this upstream editing. The *T-Aligner* ORF reconstruction demonstrates that these 5′ truncated products assemble into something that is translatable. That a majority of RPS12 ORFs are 5′ truncated is consistent with our experiences attempting to clone full-length edited RPS12 using full-gene RT-PCR in *T. cruzi*; we were only able to obtain clones in which canonical editing has been accomplished in the nucleotides corresponding to the last 25 aa of the encoded protein. The site beyond which we no longer see *any* editing among the 14 individual RPS12 clones corresponds to the 5′ most edited site of the truncated ORF, which was edited in two of our clones. Additionally, we note that a longer version of the canonical product was also reconstructed, with six additional amino acids at the N-terminus (Supplementary Table S6). This version is as abundant as the shorter one, but the N-terminus of the latter matches the N-terminus of the thoroughly reconstructed *T. brucei* protein perfectly, and for that reason the shorter version was preferred.

Finally, we made the surprising discovery of abundant alternatively edited products derived from the *T. cruzi* RPS12 cryptogene transcript that have either the N- or the C-terminus encoded in a different frame. One of 11 such products having both abundance metrics >1 (N-terminal alternative 1/C-terminal canonical, in Supplementary Table S6) is shown in black in Figure 5C and D and in Supplementary Figure S15. It has an ORF-based abundance metric of 2 and an mRNA-based metric of 5.3. This specific product uses a start codon that is few sites downstream of the canonical start codon, and the reading frame is restored by a non-canonical insertion of one U in the middle of the transcript: this site is marked with the rightmost black dot between panels C and D in Figure 5.

Further tests confirm the validity and low degree of bias of *Leptomonas T-Aligner* results

A high stringency requirement of zero or one A/G/C mismatches in the entire read, and other quality control procedures (Materials and Methods), makes mis-mapping of non-cognate transcripts highly unlikely. Nevertheless, the rarity of reads covering each unique long region of alternative editing within many ORFs prompted us to test for the likelihood of such errors. We hypothesized that if alternatively edited reads are the result of incorrect cryptogene assignment or sequencing errors, they should also appear in read populations from never-edited mRNAs (or never-edited portions of minimally-edited transcripts). To that end, we searched for reads with one or more U-indels that do not match the editing states in the canonical product (in the case of never-edited mRNAs this is the genomic sequence). Thus, for each gene we binned reads according to the number of alternative U-indel sites in them (see Figure 6). We found that the RPS12 and especially ND3 pan-edited loci stand out according to the extent of alternative editing, with the longest strings of alternatively edited sites being rather abundant: e.g. 42% reads mapped on ND3 and 30% reads mapped on RPS12 contain 10 or more alternatively edited sites. On the other end of the spectrum, reads with two or more alternative U-indels have negligible frequency (0.4–1.6%) in read pools derived from never-edited transcripts. Similar frequencies (0.1–3.9%) were obtained for reads with three or more alternative U-indels on minimally edited transcripts. Conversely, reads containing 3 or more alternatively edited sites are clearly abundant (25–70%) in read pools derived from pan-edited transcripts. Furthermore, of all the reads from all never-edited mRNAs, only 163 possessed three or more U-indels; that number drops to 39 for those having five or more. Therefore, ORFs of pan-edited mRNAs that were generated utilizing reads with extensive stretches of alternative editing are likely a legitimate but rare portion of the edited transcript population.

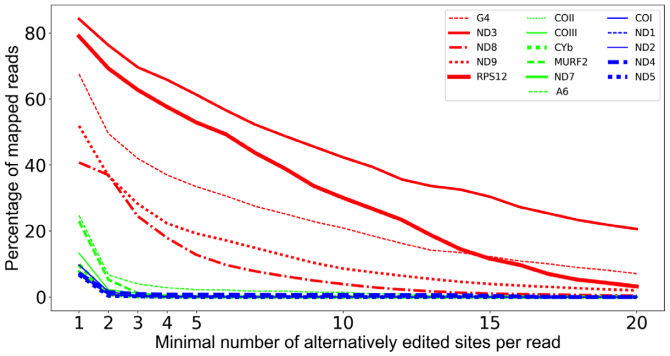

Figure 6. Percentages of *L. pyrrhocoris* mapped reads with at least 1 to at least 20 alternatively edited sites (i.e. sites with editing states different from those within the canonical open reading frame and different from the genomic sequence). All maxicircle protein-coding genes except for G3 and MURF5 (due to low coverage; Supplementary Table S1) were analyzed. Non-edited transcripts are shown in blue, minimally edited in green, and pan-edited in red. Several transcript groups can be distinguished based on the relative abundance and extent of alternatively edited reads: (i) non-edited and minimally edited transcripts with very short edited domains; (ii) MURF2, and A6 with a longer edited domain; (iii) pan-edited transcripts.

In our read coverage analysis (Supplementary Table S1), we noted that expression values for several maxicircle genes were somewhat dependent on library preparation. Therefore, we wanted to also check whether editing patterns are different in poly(A)-selected and total RNA read pools. A *T-Aligner* code extension allows comparison of relative coverage across editing states at all sites between two libraries. The coverage ratio is:

$$\left(\frac{number\ of\ reads\ covering\ the\ editing\ state}{total\ number\ of\ reads\ mapped\ on\ the\ gene}\right)_{poly\,A} \Big/ \left(\frac{number\ of\ reads\ covering\ the\ editing\ state}{total\ number\ of\ reads\ mapped\ on\ the\ gene}\right)_{total}$$

In this way, we counted editing states overrepresented in any of the two libraries, i.e. those with ratios >3 or <1/3. Only sites covered by canonical ORFs were considered. For 5 of 16 *L. pyrrhocoris* transcripts analyzed, there were no differences or only marginal differences (one or two sites with ratios >3 or <1/3) between the two libraries (Supple-

mentary Table S7). Relative coverage of non-edited states was >3 times higher in the total RNA library across the whole edited domain of A6, in the short edited domains of COII, CYb and ND7, and at both ends of the pan-edited RPS12 transcript. The only case where edited states compared to non-edited states were better represented in the total RNA library was in ND9, in which the canonical editing pathway near the 5′ end was better represented. In A6, COI, G4 and ND8, more than 10 alternative editing states per transcript had >3× higher relative coverage in the total RNA library (Supplementary Table S7). In A6, those corresponded to a minor alternative product edited outside of the canonical editing domain at the 5′ end (the 'black' ORF in Supplementary Figure S3). However, in general, the differences in editing patterns between the poly(A) and total read pools were minor, possibly skewing toward a higher representation of edited products in the poly(A) selected library for some mRNAs. The latter result could reflect differences in tail composition of the populations analyzed (51,52,68), and this also makes the poly(A)-selected library more suitable for ORF reconstruction. On the other hand, we cannot rule out that some of these differences in relative coverage might be due to the very low sample size of the read population of the total RNA library (Supplementary Table S1).

DISCUSSION

T-Aligner can probe U-indel editing of multiple species to an unprecedented degree

The dissection of the molecular oddities of protists such as kinetoplastids and the related diplonemids may have profound implications for the fields of evolution, ecology and veterinary and human medicine. The neglected diplonemids have recently emerged as the most species-rich eukaryotic group in oceanic plankton worldwide (69,70). Similarly, kinetoplastids of the group Neobodonida also represent a notable component of marine ecosystems (71). Trypanosomatids, another kinetoplastid clade, are widespread and diverse endoparasites of insects (45), with those of the genera *Leishmania* and *Trypanosoma* also adapted to parasitism in vertebrates and cause serious diseases in humans and domestic animals (72,73). One of the most fascinating and prominent features of kinetoplastids is the unique structure and expression of their mitochondrial genomes, including the U-indel editing that is necessary to make most or all encoded mRNAs translatable. This study presents a tool to probe form and function of U-indel RNA editing process across these diverse kinetoplastids.

Our U-indel transcriptome dedicated tool, *T-Aligner,* assigns deep sequencing reads to their specific cryptogenes of origin, after which it assembles them into all possible ORFs of a specific length. *T-Aligner* essentially generates sequences of edited products in the same manner that was used decades ago to identify the fully-edited forms of cryptogene-encoded mRNAs: assembling a fully-edited sequence from several overlapping partially edited clones. Sequences derived in this fashion (54,55) are widely depended upon even now (18,22). *T-Aligner*'s ability to infer all possible ORFs and sort them by read coverage will put inferences of edited products on a higher-evidence footing.

In this capacity, *T-Aligner* is different from other current software tools to investigate U-indel-edited transcriptomes which require having the identity of the fully-edited mRNA in hand prior to analysis (24,18,42). These approaches have proven highly informative in the trypanosomatid *T. brucei* where fully-edited sequences are identified, but *T-Aligner* is essential for investigations of RNA editing in virtually all the other hundreds of kinetoplastid species. Additionally, in the instances of high maxicircle variability between strains of a single kinetoplastid species, such as *T. cruzi* (56), identification of the edited sequences of one strain may not be sufficient for work in another strain. Importantly, the identification of fully-edited cryptogene sequences in various kinetoplastid species or strains by *T-Aligner* would make it then possible to use the existing TREAT and PARER tools, with their own inherent strengths, to probe editing in these new contexts.

In addition to reconstructed ORFs, the program also generates multiple graphical outputs that show read coverage and percentage of edited reads across a cryptogene, editing states at each potential edited site, and up to three pathways through editing states that were used to generate ORFs of interest. Perusal of this output allowed us to grasp unique aspects of each cryptogene analyzed, such as the demonstration that an *L. pyrrhocoris* RPS12 ORF could be reconstructed that shares sequence identity with the canonical product in the second half of the ORF but fully diverges for most of the first half. *T-Aligner* graphics also amply illustrates when and how levels of editing drop off during progression from 3′ to 5′ on the transcript. Furthermore, the low numbers of reads that can be assembled into canonical ORFs, even given the abundance of partially edited transcripts, illustrate how small a proportion of mitochondrial transcripts are productively edited (i.e. that the edited sequence encodes a translatable product).

Although *T-Aligner* has given us a view of U-indel editing that is not possible with any other tool, probing editing still remains challenging. Primarily, use of *T-Aligner* requires familiarity with both the biology of editing and deep sequencing protocols. Understanding aspects of editing such as its directionality, processivity, and the typical number of edited sites directed by a gRNA are necessary in order to evaluate *T-Aligner* output, particularly the degree to which reconstructed edited products are likely to actually exist in the transcriptome as complete, mature mRNAs. In comparing *T-Aligner* results on the *L. pyrrhocoris* and *T. cruzi* datasets, where library generation and sequencing were performed differently, it is likely that the length of reads used as input plays a role in the probability that *T-Aligner* will reconstruct translatable products. Finally, the previously mentioned potential bias of poly(A) tail selection should always be considered.

Potential functional impacts of alternative editing

T-Aligner has revealed the extent to which editing produces open reading frames that differ from the canonical ones that are the expected translated products of the mitochondrial genome. If these potential alternative mature mRNAs are indeed translated, it is important to consider whether or not they may impact cellular function. The products we re-

constructed for all pan-edited transcripts that possess long stretches of alternatively edited sites are present at such a low abundance that it is difficult to envision their translation being impactful. However, it is interesting that such patterns, probably requiring cascades of up to seven alternative gRNAs, are retained at all. Possessing an inherent possibility for the emergence of radically new protein sequences (currently at background levels) may confer an evolutionary advantage. A mitochondrial gene coding for two radically different transcripts (due to a frameshift introduced by RNA editing, for example) might by advantageous for another reason (37). If the canonical mitochondrial protein is dispensable at a certain life cycle stage (for instance, bloodstream *T. brucei*), the gene would be easily lost due to mutations and genetic drift, but the loss would be prevented if the frameshifted transcript version is functional and indispensable at this life cycle stage.

We also reconstructed products that differ from the canonical product by one or a few internal amino acids. Even small changes such as these could alter the function of the translated protein. As these smaller deviations from the canonical product have a higher read support metric, they are more likely to be functionally relevant than the highly different but very rare alternative sequences. Another group of abundant products resulting from one or few alternatively edited sites are extensions, truncations, and alternative sequences of the 5′ ends and 3′ ends of pan-edited mRNAs, such as with *L. pyrrhocoris* ND8 or *L. pyrrhocoris* and *T. cruzi* RPS12. Others' analyses identified alternative 5′ ends of *T. brucei* RPS12 (24) and CR3 and ND7 (37), the latter even describing the gRNAs likely responsible for such differences. If such alternative products were translated, they could retain at least some function of the canonical product. However, barring any evidence of translation of more than one product from cryptogene loci, lacking entirely in the organisms we investigated, a more detailed analysis of possible function of our edited product reconstructions would be premature.

Questions about U-indel editing triggered by *T-Aligner* analysis

The fact that many editing patterns found in the collected reads do not conform to a functional product and/or cannot be assembled into an ORF brings up a major question. Does the kinetoplastid mitochondrion contain a large population of transcripts that are dead-end products of aberrant editing, a useless byproduct of a sloppy process that is slated for degradation? Alternatively, are these multiple early-terminating editing cascades or isolated alternatively edited sites components of a functional yet so far overlooked process? We might have captured an enzymatic process requiring flexible editing patterns in the junction regions that will eventually resolve into functional product(s) as proposed by Simpson et al. (18,24). However, in many cases editing patterns in the main and alternative products diverge too much to imagine that continued modifications to the existing editing pattern of the alternative products would eventually resolve into the canonical sequence.

In this case, another key question is whether these complex editing patterns that deviate from the main ORF pathway require alternative gRNAs. Due to the intrinsic capacity of kinetoplast DNA to encode extremely diverse gRNAs (19,20), binding of non-cognate (alternative) gRNAs may be responsible for these patterns. For example, utilization of multiple different gRNAs in alternative editing cascades can aptly account for the very low coverage (and thus presumably rare) background of products that have long stretches of alternatively edited sites. It is difficult to explain the existence of long alternative editing cascades without alternative gRNAs, but most other alternative patterns can be explained by either temporary or abortive mis-editing with cognate gRNAs.

Perhaps the most exciting question is whether the multiple editing patterns that assemble into alternative translatable products are actually translated? In the case of a transcript harboring only a few internal editing modifications, such as those that result in a single amino acid substitution, insertion, or deletion, it is hard to envision a mechanism to prevent its translation. However, its translation is currently difficult to prove. Emerging technologies such as mitochondrial ribosomal profiling (74) and highly sensitive mass spectrometry may prove useful in probing it. In conclusion, we expect that *T-Aligner* will continue to be instrumental in solving these and similar questions regarding the molecular process of U-indel editing. Moreover, there is an area in which it could potentially be even more useful, and that is comparative biology. The analysis of multiple kinetoplastid mitochondrial transcriptomes using *T-Aligner* will allow us to identify ways that U-indel editing has been utilized and/or has evolved differently, and perhaps may even shed light on the origins of this fascinating molecular phenomenon.

AVAILABILITY

T-Aligner's code is available at GitHub [https://github.com/jalgard/T-Aligner3]. Transcriptome sequencing data are deposited in NCBI SRA with BioProject IDs PRJNA392757 for *L. pyrrhocoris* and PRJNA395140 for *T. cruzi*. Sequences of various alternative products are shown in Supplementary Tables S4–S6.

Canonical edited mRNA sequences reconstructed with *T-Aligner* are deposited in GenBank [accession numbers MF409180-MF409198]; see Supplementary Table S3 for details.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

European Research Council CZ [LL1601 to J.L.]; Czech Science Foundation [15-21974S to J.L., 17-10656S to V.Y., 16-18699S to J.L. and V.Y.]; University of Ostrava IRP project "New research directions in the Life Science Research Centre"; Moravian–Silesian Region [research programs 2013–2014 and 2015, DT01-021358 to V.Y. and P.F.]; Czech Ministry of Education, Youth and Sports [LO1208 "TEWEP" to V.Y.]; Russian Science Foundation [transcriptomic library sequencing was supported by 14-50-00029 to

M.D.L. and P.F.]; Russian Foundation for Basic Research [14-04-01717 to A.A.K. and E.S.G.]; American Heart Association [16SDG26420019 to S.L.Z.]; University of Minnesota Genomics Center Pilot Project Award to S.L.Z.; S.L.Z. received an Alexander Dubček Fund travel fellowship (University of Minnesota). Funding for open access charge: American Heart Association [16SDG26420019 to S.L.Z.]; University of Ostrava (Rector's Award to P.F.).
Conflict of interest statement. None declared.

REFERENCES

1. Gray,M.W., Lukeš,J., Archibald,J.M., Keeling,P.J. and Doolittle,W.F. (2010) Cell biology. Irremediable complexity? *Science*, **330**, 920–921.
2. Koonin,E.V. (2016) Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biol.*, **14**, 114.
3. Lukeš,J., Archibald,J.M., Keeling,P.J., Doolittle,W.F. and Gray,M.W. (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life*, **63**, 528–537
4. Stoltzfus,A. (2012) Constructive neutral evolution: exploring evolutionary theory's curious disconnect. *Biol Direct*, **7**, 35.
5. Smith,D.R. and Keeling,P.J. (2016) Protists and the Wild, Wild West of gene expression: new frontiers, Lawlessness, and Misfits. *Annu. Rev. Microbiol.*, **70**, 161–178.
6. Burger,G., Moreira,S. and Valach,M. (2016) Genes in Hiding. *Trends Genet.*, **32**, 553–565.
7. Faktorová,D., Dobáková,E., Peña-Diaz,P. and Lukeš,J. (2016) From simple to supercomplex: mitochondrial genomes of euglenozoan protists. *F1000Research*, **5**, 392.
8. Valach,M., Moreira,S., Faktorová,D., Lukeš,J. and Burger,G. (2016) Post-transcriptional mending of gene sequences: looking under the hood of mitochondrial gene expression in diplonemids. *RNA Biol.*, **13**, 1204–1211.
9. Yabuki,A., Tanifuji,G., Kusaka,C., Takishita,K. and Fujikura,K. (2016) Hyper-eccentric structural genes in the mitochondrial genome of the algal parasite *Hemistasia phaeocysticola*. *Genome Biol Evol.*, **8**, 2870–2878.
10. Aphasizheva,I. and Aphasizhev,R. (2016) U-insertion/deletion mRNA-editing holoenzyme: definition in sight. *Trends Parasitol.*, **32**, 144–156.
11. Read,L.K., Lukeš,J. and Hashimi,H. (2016) Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip. Rev. RNA*, **7**, 33–51.
12. David,V., Flegontov,P., Gerasimov,E., Tanifuji,G., Hashimi,H., Logacheva,M.D., Maruyama,S., Onodera,N.T., Gray,M.W., Archibald,J.M. *et al.* (2015) Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsela*, an endosymbiotic kinetoplastid. *mBio.*, **6**, e01498–15
13. Simpson,L., Thiemann,O.H., Savill,N.J., Alfonzo,J.D. and Maslov,D.A. (2000) Evolution of RNA editing in trypanosome mitochondria. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 6986–6993.
14. Simpson,L. and Maslov,D.A. (1999) Evolution of the U-insertion/deletion RNA editing in mitochondria of kinetoplastid protozoa. *Ann. N. Y. Acad. Sci.*, **870**, 190–205.
15. Aravin,A.A., Yurchenko,V., Merzlyak,E.M. and Kolesnikov,A.A. (1998) The mitochondrial ND8 gene from *Crithidia oncopelti* is not pan-edited. *FEBS Lett.*, **431**, 457–460.
16. Gerasimov,E.S., Kostygov,A.Y., Yan,S. and Kolesnikov,A.A. (2012) From cryptogene to gene? ND8 editing domain reduction in insect trypanosomatids. *Eur. J. Protistol.*, **48**, 185–193.
17. McDermott,S.M., Luo,J., Carnes,J., Ranish,J.A. and Stuart,K. (2016) The architecture of *Trypanosoma brucei* editosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, e6476–e6485.
18. Simpson,R.M., Bruno,A.E., Chen,R., Lott,K., Tylec,B.L., Bard,J.E., Sun,Y., Buck,M.J. and Read,L.K. (2017) Trypanosome RNA Editing Mediator Complex proteins have distinct functions in gRNA utilization. *Nucleic Acids Res.*, **45**, 7965–7983.
19. Kirby,L.E., Sun,Y., Judah,D., Nowak,S. and Koslowsky,D. (2016) Analysis of the *Trypanosoma brucei* EATRO 164 bloodstream guide RNA transcriptome. *PLoS Negl. Trop. Dis.*, **10**, e0004793.
20. Koslowsky,D., Sun,Y., Hindenach,J., Theisen,T. and Lucas,J. (2014) The insect-phase gRNA transcriptome in *Trypanosoma brucei*. *Nucleic Acids Res.*, **42**, 1873–1886.
21. Simpson,L., Douglass,S.M., Lake,J.A., Pellegrini,M. and Li,F. (2015) Comparison of the mitochondrial genomes and steady state transcriptomes of two strains of the trypanosomatid parasite, *Leishmania tarentolae*. *PLoS Negl. Trop. Dis.*, **9**, e0003841.
22. Ammerman,M.L., Presnyak,V., Fisk,J.C., Foda,B.M. and Read,L.K. (2010) TbRGG2 facilitates kinetoplastid RNA editing initiation and progression past intrinsic pause sites. *RNA*, **16**, 2239–2251.
23. Koslowsky,D.J., Bhat,G.J., Read,L.K. and Stuart,K. (1991) Cycles of progressive realignment of gRNA with mRNA in RNA editing. *Cell*, **67**, 537–546.
24. Simpson,R.M., Bruno,A.E., Bard,J.E., Buck,M.J. and Read,L.K. (2016) High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing. *RNA*, **22**, 677–695.
25. Arts,G.J., van der Spek,H., Speijer,D., van den Burg,J., van Steeg,H., Sloof,P. and Benne,R. (1993) Implications of novel guide RNA features for the mechanism of RNA editing in *Crithidia fasciculata*. *EMBO J.*, **12**, 1523–1532.
26. Maslov,D.A. and Simpson,L. (1992) The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. *Cell*, **70**, 459–467.
27. Maslov,D.A., Thiemann,O. and Simpson,L. (1994) Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of *Leishmania tarentolae*. *Mol. Biochem. Parasitol.*, **68**, 155–159.
28. Sturm,N.R., Maslov,D.A., Blum,B. and Simpson,L. (1992) Generation of unexpected editing patterns in *Leishmania tarentolae* mitochondrial mRNAs: misediting produced by misguiding. *Cell*, **70**, 469–476.
29. Ochsenreiter,T. and Hajduk,S.L. (2006) Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. *EMBO Rep.*, **7**, 1128–1133.
30. Ochsenreiter,T., Anderson,S., Wood,Z.A. and Hajduk,S.L. (2008) Alternative RNA editing produces a novel protein involved in mitochondrial DNA maintenance in trypanosomes. *Mol. Cell. Biol.*, **28**, 5595–5604.
31. Ochsenreiter,T., Cipriano,M. and Hajduk,S.L. (2008) Alternative mRNA editing in trypanosomes is extensive and may contribute to mitochondrial protein diversity. *PLoS One.*, **3**, e1566.
32. Madina,B.R., Kumar,V., Metz,R., Mooers,B.H., Bundschuh,R. and Cruz-Reyes,J. (2014) Native mitochondrial RNA-binding complexes in kinetoplastid RNA editing differ in guide RNA composition. *RNA*, **20**, 1142–1152.
33. Flegontov,P., Gray,M.W., Burger,G. and Lukeš,J. (2011) Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Curr. Genet.*, **57**, 225–232.
34. Gray,M.W. (2012) Evolutionary origin of RNA editing. *Biochemistry*, **51**, 5235–5242.
35. Speijer,D. (2006) Is kinetoplastid pan-editing the result of an evolutionary balancing act? *IUBMB Life*, **58**, 91–96.
36. Speijer,D. (2011) Does constructive neutral evolution play an important role in the origin of cellular complexity? Making sense of the origins and uses of biological complexity. *Bioessays*, **33**, 344–349.
37. Kirby,L.E. and Koslowsky,D. (2017) Mitochondrial dual-coding genes in *Trypanosome brucei*. *PLoS Negl. Trop. Dis.*, **11**, e0005989.
38. Maslov,D.A., Sturm,N.R., Niner,B.M., Gruszynski,E.S., Peris,M. and Simpson,L. (1992) An intergenic G-rich region in *Leishmania tarentolae* kinetoplast maxicircle DNA is a pan-edited cryptogene encoding ribosomal protein S12. *Mol. Cell. Biol.*, **12**, 56–67.
39. Feagin,J.E., Abraham,J.M. and Stuart,K. (1988) Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell*, **53**, 413–422.
40. Shaw,J.M., Feagin,J.E., Stuart,K. and Simpson,L. (1988) Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. *Cell*, **53**, 401–411.
41. Feagin,J.E., Shaw,J.M., Simpson,L. and Stuart,K. (1988) Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 539–543.

42. Carnes,J., McDermott,S., Anupama,A., Oliver,B.G., Sather,D.N. and Stuart,K. (2017) *In vivo* cleavage specificity of *Trypanosoma brucei* editosome endonucleases. *Nucleic Acids Res.*, **45**, 4667–4686.
43. del Campo,J., Sieracki,M.E., Molestina,R., Keeling,P., Massana,R. and Ruiz-Trillo,I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.*, **29**, 252–259.
44. Votýpka,J., Klepetková,H., Yurchenko,V.Y., Horák,A., Lukeš,J. and Maslov,D.A. (2012) Cosmopolitan distribution of a trypanosomatid *Leptomonas pyrrhocoris*. *Protist.*, **163**, 616–631.
45. Maslov,D.A., Votýpka,J., Yurchenko,V. and Lukeš,J. (2013) Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol.*, **29**, 43–52.
46. Flegontov,P., Butenko,A., Firsov,S., Kraeva,N., Eliáš,M., Field,M.C., Filatov,D., Flegontova,O., Gerasimov,E.S., Hlaváčová,J. *et al.* (2016) Genome of *Leptomonas pyrrhocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. *Sci. Rep.*, **6**, 23704.
47. Castellani,O., Ribeiro,L.V. and Fernandes,J.F. (1967) Differentiation of *Trypanosoma cruzi* in culture. *J. Protozool.*, **14**, 447–451.
48. Pelletier,M., Read,L.K. and Aphasizhev,R. (2007) Isolation of RNA binding proteins involved in insertion/deletion editing. *Methods Enzymol.*, **424**, 75–105.
49. Horváth,A., Horáková,E., Dunajcíková,P., Verner,Z., Pravdová,E., Slapetová,I., Cuninková,L. and Lukes,J. (2005) Downregulation of the nuclear-encoded subunits of the complexes III and IV disrupts their respective complexes but not complex I in procyclic *Trypanosoma brucei*. *Mol. Microbiol.*, **58**, 116–130.
50. Záhonová,K., Hadariová,L., Vacula,R., Yurchenko,V., Eliáš,M., Krajčovič,J. and Vesteg,M. (2014) A small portion of plastid transcripts is polyadenylated in the flagellate *Euglena gracilis*. *FEBS Lett.*, **588**, 783–788.
51. Zhang,L., Sement,F.M., Suematsu,T., Yu,T., Monti,S., Huang,L., Aphasizhev,R. and Aphasizheva,I. (2017) PPR polyadenylation factor defines mitochondrial mRNA identity and stability in trypanosomes. *EMBO J.*, **36**, 2435–2454.
52. Gazestani,V.H., Hampton,M., Shaw,A.K., Liggett,C., Salavati,R. and Zimmer,S.L. (2017) Tail characteristics of *Trypanosoma brucei* mitochondrial transcripts are developmentally altered in a transcript-specific manner. *Int. J. Parasitol.*, doi:10.1016/j.ijpara.2017.08.012.
53. Aslett,M., Aurrecoechea,C., Berriman,M., Brestelli,J., Brunk,B.P., Carrington,M., Depledge,D.P., Fischer,S., Gajria,B., Gao,X. *et al.* (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.*, **38**, D457–D462.
54. Souza,A.E., Myler,P.J. and Stuart,K. (1992) Maxicircle CR1 transcripts of *Trypanosoma brucei* are edited and developmentally regulated and encode a putative iron-sulfur protein homologous to an NADH dehydrogenase subunit. *Mol. Cell. Biol.*, **12**, 2100–2107.
55. Read,L.K., Myler,P.J. and Stuart,K. (1992) Extensive editing of both processed and preprocessed maxicircle CR6 transcripts in *Trypanosoma brucei*. *J. Biol. Chem.*, **267**, 1123–1128.
56. Ruvalcaba-Trejo,L.I. and Sturm,N.R. (2011) The *Trypanosoma cruzi* Sylvio X10 strain maxicircle sequence: the third musketeer. *BMC Genomics*, **12**, 58.
57. Langmead,B. and Salzberg,S. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
58. Gazestani,V.H., Hampton,M., Abrahante,J.E., Salavati,R. and Zimmer,S.L. (2016) circTAIL-seq, a targeted method for deep analysis of RNA 3′ tails, reveals transcript-specific differences by multiple metrics. *RNA*, **22**, 477–486.
59. Jirků,M., Yurchenko,V.Y., Lukeš,J. and Maslov,D.A. (2012) New species of insect trypanosomatids from Costa Rica and the proposal for a new subfamily within the Trypanosomatidae. *J. Eukaryot. Microbiol.*, **59**, 537–547.
60. Landweber,L.F. and Gilbert,W. (1993) RNA editing as a source of genetic variation. *Nature*, **363**, 179–182.
61. Yasuhira,S. and Simpson,L. (1995) Minicircle-encoded guide RNAs from *Crithidia fasciculata*. *RNA*, **1**, 634–643
62. Nebohácová,M., Kim,C.E., Simpson,L. and Maslov,D.A. (2009) RNA editing and mitochondrial activity in promastigotes and amastigotes of *Leishmania donovani*. *Int. J. Parasitol.*, **39**, 635–644.
63. Read,L.K., Wilson,K.D., Myler,P.J. and Stuart,K. (1994) Editing of *Trypanosoma brucei* maxicircle CR5 mRNA generates variable carboxy terminal predicted protein sequences. *Nucleic Acids Res.*, **22**, 1489–1495.
64. Maslov,D.A. (2010) Complete set of mitochondrial pan-edited mRNAs in *Leishmania mexicana amazonensis* LV78. *Mol. Biochem. Parasitol.*, **173**, 107–114.
65. Westenberger,S.J., Cerqueira,G.C., El-Sayed,N.M., Zingales,B., Campbell,D.A. and Sturm,N.R. (2006) *Trypanosoma cruzi* mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. *BMC Genomics*, **7**, 60.
66. Thomas,S., Martinez,L.L., Westenberger,S.J. and Sturm,N.R. (2007) A population study of the minicircles in *Trypanosoma cruzi*: predicting guide RNAs in the absence of empirical RNA editing. *BMC Genomics*, **8**, 133.
67. Greif,G., Rodriguez,M., Reyna-Bello,A., Robello,C. and Alvarez-Valin,F. (2015) Kinetoplast adaptations in American strains from *Trypanosoma vivax*. *Mutat. Res.*, **773**, 69–82.
68. Etheridge,R.D., Aphasizheva,I., Gershon,P.D. and Aphasizhev,R. (2008) 3′ adenylation determines mRNA abundance and monitors completion of RNA editing in *T. brucei* mitochondria. *EMBO J.*, **27**, 1596–1608.
69. Flegontova,O., Flegontov,P., Malviya,S., Audic,S., Wincker,P., de Vargas,C., Bowler,C., Lukeš,J. and Horák,A. (2016) Extreme diversity of diplonemid eukaryotes in the ocean. *Curr. Biol.*, **26**, 3060–3065.
70. Gawryluk,R.M., Del Campo,J., Okamoto,N., Strassert,J.F., Lukeš,J., Richards,T.A., Worden,A.Z., Santoro,A.E. and Keeling,P.J. (2016) Morphological identification and single-cell genomics of marine diplonemids. *Curr. Biol.*, **26**, 3053–3059.
71. de Vargas,C., Audic,S., Henry,N., Decelle,J., Mahé,F., Logares,R., Lara,E., Berney,C., Le Bescot,N., Probert,I. *et al.* (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.
72. Lukeš,J., Skalický,T., Týč,J., Votýpka,J. and Yurchenko,V. (2014) Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.*, **195**, 115–122.
73. Rodrigues,J.C., Godinho,J.L. and de Souza,W. (2014) Biology of human pathogenic trypanosomatids: epidemiology, lifecycle and ultrastructure. *Subcell. Biochem.*, **74**, 1–42.
74. Parsons,M., Ramasamy,G., Vasconcelos,E.J., Jensen,B.C. and Myler,P.J. (2015) Advancing *Trypanosoma brucei* genome annotation through ribosome profiling and spliced leader mapping. *Mol. Biochem. Parasitol.*, **202**, 1–10.