Molecular & Biochemical Parasitology 202 (2015) 1–10

Contents lists available at ScienceDirect

Molecular & Biochemical Parasitology

Advancing *Trypanosoma brucei* genome annotation through ribosome profiling and spliced leader mapping

Marilyn Parsons[a,b,*], Gowthaman Ramasamy[a], Elton J.R. Vasconcelos[a,1], Bryan C. Jensen[a], Peter J. Myler[a,b,c]

[a] The Center for Infectious Disease Research (formerly Seattle Biomedical Research Institute), 307 Westlake Ave. N., Seattle, WA 98109, USA
[b] Dept of Global Health, University of Washington, Seattle, WA 98195, USA
[c] Dept of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 6 August 2015
Received in revised form
18 September 2015
Accepted 18 September 2015
Available online 21 September 2015

Keywords:
De novo gene evolution
Genome annotation
Ribosome profiling
Translation
Trypanosomes

ABSTRACT

Since the initial publication of the trypanosomatid genomes, curation has been ongoing. Here we make use of existing *Trypanosoma brucei* ribosome profiling data to provide evidence of ribosome occupancy (and likely translation) of mRNAs from 225 currently unannotated coding sequences (CDSs). A small number of these putative genes correspond to extra copies of previously annotated genes, but 85% are novel. The median size of these novels CDSs is small (81 aa), indicating that past annotation work has excelled at detecting large CDSs. Of the unique CDSs confirmed here, over half have candidate orthologues in other trypanosomatid genomes, most of which were not yet annotated as protein-coding genes. Nonetheless, approximately one-third of the new CDSs were found only in *T. brucei* subspecies. Using ribosome footprints, RNA-Seq and spliced leader mapping data, we updated previous work to definitively revise the start sites for 414 CDSs as compared to the current gene models. The data pointed to several regions of the genome that had sequence errors that altered coding region boundaries. Finally, we consolidated this data with our previous work to propose elimination of 683 putative genes as protein-coding and arrive at a view of the translatome of slender bloodstream and procyclic culture form *T. brucei*.

1. Introduction

The protozoan parasite species *Trypanosoma brucei* (which includes the subspecies *T. brucei brucei, T. brucei rhodesiense*, and *T. brucei gambiense*) infects a wide variety of mammals including humans, domestic livestock and wild animals in sub-Saharan Africa. Human African trypanosomiasis (also known as African sleeping sickness) is a fatal disease in the absence of treatment, and currently used drugs are poorly tolerated and facing resistance. In the mammalian host the parasite exists as extracellular bloodstream forms (BF), with actively dividing slender BF transitioning to non-dividing stumpy BF that are poised for transmission via the tsetse fly. In the tsetse fly, the parasite transforms into procyclic midgut stages; similar forms can be cultured *in vitro* (procyclic cultured forms, PCF). Reflecting their vastly different environments, BF and PCF differ widely in gene expression, particularly with respect to metabolism and surface antigen expression [1–5].

The 35 megabase haploid genome of *T. b. brucei* (hereafter *T. brucei*) includes eleven large chromosomes that are thought to encode all active genes [6]. Additionally, the organism harbors a set of about 100 small chromosomes [7] that are comprised mainly of repetitive sequence and whole or partial variant surface glycoprotein genes (variant surface glycoproteins are responsible for antigenic variation). The original annotation of protein coding sequences in *T. brucei* was based on predictions derived from Glimmer, GC bias, and homology to known protein coding sequences in other organisms [6,8,9]. The initially published genome included 9068 protein-coding genes plus 904 pseudogenes and has been revised in multiple versions, in part due to further refinement of the underlying sequence of the large chromosomes [6]. Additionally, mapping of novel transcripts led to the identification of potential coding regions, several of which have been functionally verified via phenotypic analysis following RNAi [10,11]. To validate these potential protein coding sequences (CDSs), as well as discover novel ones, we used a fundamentally different approach, based on the observation

Abbreviations: CDS, coding sequence; BF, bloodstream forms; ORF, open reading frame; PCF, procyclic cultured forms; SL, spliced leader; Tb927, *Trypanosoma brucei* strain TREU927; VSG, variant surface glycoprotein.

* Corresponding author at: The Center for Infectious Disease Research, 307 Westlake Ave. N., Seattle, WA 98109, USA. Fax: +1 2062567229.
E-mail address: marilyn.parsons@cidresearch.org (M. Parsons).
[1] Present address: Dept. of Biochemistry, Institute of Chemistry, University of São Paulo - Av. Prof. Lineu Prestes 748, sala 1200, 05508-000, São Paulo, SP, Brasil.

http://dx.doi.org/10.1016/j.molbiopara.2015.09.002
0166-6851/

of ribosome association with mRNAs that mapped to previously non-annotated open reading frames (ORFs). This approach, called ribosome profiling, exploits the ability of the ribosome to protect a 28 nt fragment of the mRNA from digestion with nuclease [12]. These protected fragments are used to generate a library representing regions of mRNAs undergoing translation. At the same time, a parallel library of mRNA fragments is prepared, allowing a comparison of the relative translation of different CDSs, as well as the boundaries of the mRNA. Ribosome profiling has recently been used to demonstrate widespread translational control during *T. brucei* [13,14] and *Trypanosoma cruzi* [15] development. Using this technology, all regions of the genome can be analyzed for evidence of protein production (or lack thereof), as well as relative translation efficiency. In the case of trypanosomatids, the ribosome profiling approach can be bolstered by spliced leader (SL) mapping [13], which identifies the 5′ ends of all trypanosome mRNAs [11,16,17].

In our previous ribosome profiling studies, which focused on stage-regulation of protein production in *T. brucei*, we observed numerous non-annotated regions with ribosome footprints (via manual inspection), indicating the presence of new CDSs [13] and others have reported a large set of ORFs with footprint reads (*via* bioinformatic analysis) [14]. In this communication we extend those studies by additional analysis of regions previously suggested to have coding potential [10,11,14]. We also utilize the ribosome profiling pipeline together with SL mapping to further assess annotated CDSs, leading to the elimination of 683 gene models. We use the combined data to extend or truncate predicted CDSs and, in some instances, to identify underlying errors in the genome sequence. Here, we present examples that illustrate how ribosome profiling can be used to enhance genome annotation.

2. Methods

2.1. Validation of novel CDSs

We utilized previously obtained ribosome profiling data from *T. brucei* strain TREU927 (Tb927, PCF and *in vivo* derived slender BF) and strain Lister 427 (cultured BF) [13]. A summary of our validation pipeline is shown in Fig. S1. We conducted an initial manual evaluation of ribosome footprint, mRNA and SL reads as reported [13], to identify novel CDSs not present on the eleven major chromosomes in GeneDB version 5.0 of the *T. brucei* genome. For gene-level reviews of protein production, we examined both *in vivo* derived slender BF and PCF data. Of the 182 new CDSs we identified previously [13], 24 were removed from our list because they were subsequently annotated in genome version 5.1. We next extended our analysis to review the 857 ORFs of at least 25 aa in length identified by Vasquez et al. [14] as having ribosome footprints covering at least 70% of the ORF. Additionally we included the 187 predicted CDSs listed by Ericson et al. [10] that had been selected on the basis of mRNA expression and phylogenetic conservation (some of which we had already identified in our initial inspection). Total read counts for each candidate CDS were summed across the nine libraries, after excluding those reads mapping to the first 45 nt of the CDS which have higher ribosome loading [13], a result of using cycloheximide [12]. All ORFs from that dataset that had a total of at least 180 ribosome footprint reads (an average of 20 per library) were visually inspected in Artemis [18]. This cutoff was chosen because visual inspection of all (>300) proposed CDSs on chromosomes 1 and 10 revealed that none with fewer than 180 total reads met our criteria for being translated (see below), but some with fewer than 270 total reads appeared to be translated. To qualify as a CDS, we required that the ribosome footprint reads span the length of the ORF and terminate at the first in-frame stop codon. Evaluating whether translation spans the ORF is more challenging with short CDSs, due to the natural peaks and valleys in signal that follow the higher ribosome loading of the first 15 codons; hence our visual inspection is likely biased against very short CDSs. A region was not considered to be a new CDS if the ribosome footprints mapped to an overlapping CDS without an increase in the density of ribosome footprints or if the mRNA and SL data contradicted the existence of the corresponding transcript. A total of 225 novel CDSs were given a systematic identifier based on genomic localization in consultation with GeneDB curators. To define the unique set of new CDSs (178 in total), CDSs that are closely related copies of known sequences were excluded (as they are not novel) and multicopy families were reduced to a single representative.

Phylogenetic analyses of the unique new CDSs were performed by searching predicted proteomes using Blastp and genomes using tBlastn. The organisms analyzed were *T. b. gambiense* DAL972 (GCA_000210295.1/TriTrypDB V8.0), *Trypanosoma congolense* (GCA_000227395.2/TriTrypDB V8.0), *Trypanosoma vivax* (GCA_000227375.1/TriTrypDB V8.0), *Trypanosoma cruzi* (GCA_000209065.1/TriTrypDB V8.0), *Trypanosoma grayi* (GCA_000691245.1/TriTrypDB V8.0), *Crithidia fasciculata* (GCA_000331325.1/TriTrypDB V8.0), *Leishmania major* Friedlin (GCA_000002725.2/TriTrypDB V7.0), *Bodo saltans* (ftp://ftp.sanger.ac.uk/pub/pathogens/Bodo/saltans), *Sacchoromyces cerevisiae* (GCA_000146045.2), *Caenorhabditis elegans* (GCA_000002985.3), *Drosophila melanogaster* (GCA_000001215.4), *Arabidopsis thaliana* (www.arabidopsis.org; V TAIR10), and *Homo sapiens*(assembly ID: GCA_000001405.15). The cutoff used was an *E*-value ≤ 0.001 combined with 30% identity and 50% coverage of the query sequence.

2.2. Elimination of predicted CDSs

These data represent a refinement of our previously published work [13], with modifications after re-review and adjustments for CDSs eliminated or renamed in Tb927 genome v5.1. In brief, CDS elimination used a structured approach. No CDSs encoding pseudogenes or variant surface glycoproteins (VSGs) were eliminated. No CDSs were eliminated solely due to the lack of mRNA reads and/or ribosome footprints, as they might be expressed in other stages of parasite development. We eliminated all CDSs for which the mRNA reads (and footprints if present) mapped to the opposite strand. We also eliminated those CDSs where the mRNA and SL evidence contradicted the gene model. For example, if the CDS was not encoded on a contiguous mRNA or was in the 3′ UTR of another mRNA and lacked ribosome footprints, the CDS was eliminated.

2.3. Changes in CDS start sites

For re-definition of CDS start sites, we considered the start of the mRNA as demarcated by the SL site that paralleled the increase in mRNA reads, and we also considered ribosome footprints associated with the CDS. By these criteria, almost all CDSs with good ribosome footprint coverage began at the first ATG after the start of the mRNA in both slender bloodstream and procyclic forms. Therefore, we used the first ATG of the mRNA as the start codon, unless there was evidence to the contrary. For example, in some cases small upstream open reading frames (uORFs) were present in alternate reading frames or separated from the main CDS by a stop codon. In these cases, the pattern of ribosome profiling supported a distal ATG as the start codon of the main CDS. We found that systematic computational prediction of the 5′ end of the transcripts was unreliable due to numerous closely spaced, interspersed transcripts that were obvious only upon visual inspection. Hence all boundary changes were visually inspected, with special attention to discrepancies between published datasets [11,13,16,17]. Pseudogenes and VSGs were not evaluated.

Table 1
Summary of changes to CDSs.

Category[a]	Number of CDSs
Deleted	683
New	225
Extended	214
Shortened	200
Unchanged	7791
Total (excluding deleted genes)	8429

[a] As compared to GeneDB annotation of Tb927 genome version 5.1. Pseudogenes are not included.

The complete list of CDSs and coordinates has been provided to GeneDB, where curators are incorporating the changes. Signal sequences and signal anchor sequences were predicted by SignalP [19] and transmembrane domains (TMD) by TMHMM [20]. Domains were identified by searches in Pfam and Interpro. EdgeR-normalized read counts using genome-wide data [21] and ribosome release factors were calculated as described [13]. The re-normalized data for the entire genome have been provided to GEO (GSE72463).

3. Results and discussion

3.1. Verification of new CDSs

The ribosome profiling protocol combines the analysis of the regions of mRNAs protected by ribosomes (ribosome footprints) with the analysis of the mRNA boundaries, as revealed by high throughput sequencing. Thus, for each segment of DNA, the procedure measures the relative abundance of mRNAs and identifies the regions being translated. For detection and verification of putative new CDSs, we used data from our previous ribosome profiling of *T. brucei* strain 927 PCF, *in vivo*-derived slender BF of the same strain, and cultured BF of strain 427 [13], which we analyzed (see decision tree, Fig. S1) as described in the Section 2. We also included data from our SL mapping of PCF (~14.1 million aligned reads) and slender BF (~11.9 million aligned reads), to define the 5′ boundaries of transcripts. The reads from various libraries were mapped to the 11 chromosomes of the *T. brucei* genome and visualized using Artemis. The ribosome profiling data (ribosome footprints and mRNA reads), along with SL mapping, were used to assess the CDS models. Table 1 summarizes the changes to the CDS models (excluding pseudogenes) across the 11 major chromosomes compared to version 5.1 of the *T. brucei* genome. The complete revised list of *T. brucei* CDSs and their coordinates, along with their source, is provided in Table S1.

We visually "walked" through each chromosome to search for regions that showed evidence of protein coding regions at least 25 aa in length in the absence of an annotated gene. Additionally, we inspected ORFs identified in two previous studies [10,14] that were at least 25 aa long and had a total of at least 180 raw reads in our nine ribosome profiling libraries. Together, this led to a set of 225 CDSs. All but one of the 66 previously non-annotated CDSs that had support from mass-spectrometry [10,14] were validated here. The sole exception was Tb11.NT.90 for which the ribosome footprints mapped to an overlapping, annotated CDS. Two of the identified regions (Tb927.8.8215 and Tb927.10.14615) correspond to fragments of other annotated CDSs and hence may or may not be new CDSs. For mRNAs that are truly translated, ribosome occupancy should fall off rapidly at the stop codon. We therefore calculated the ribosome release factor, which compares the ribosome loading before and after the stop codon (Table S3). For the 190CDSs for which a score could be calculated, 89% had at least a 10-fold higher ribosome density in the coding region than in the 3′ UTR (when normalized to mRNA read counts), similar to the genome-wide scores previously obtained (87% > 10-fold, [13]) and supporting the contention that these CDSs are indeed translated.

The new CDSs include unique sequences, CDSs related to known genes, and additional copies of known CDSs (Fig. 1A). Nucleotide and amino acid sequences for each are provided in Table S2. Interestingly, three of the new CDSs corresponded to sequences that had been described in experimental analyses [22,23], but were not yet annotated in the genome. Seventeen CDSs were identical or almost identical to annotated CDSs. These included some that were additional gene copies in a recognized tandem array. For example, two new copies of histone H2B (both the CDSs and intergenic regions are identical) were found, as well as additional copies of the *T. brucei* orthologue of kinetoplastid membrane protein 11 (*KMP11*) and ADP-ribosylation factor 1 (*ARF1*). A few of the new CDSs were nearby or dispersed relatives of another, previously annotated gene. For example, a new CDS on chromosome 11 (Tb927.11.7705) showed 96% nt sequence identity and 100% aa identity to the ribosomal protein S14 gene on chromosome 6 (Tb927.6.4980). Uniquely mapping ribosome footprints indicate that both of these CDSs contribute to RPS14 protein production. Another example is Tb927.11.7707, a copy of the dynein light chain gene Tb927.11.7740 that is located a few kb away, the two being separated by *MSP1* genes. Interestingly, there are two additional copies of this dynein sequence, almost identical (99% nucleotide identity) except for the lack of a start codon, with GTG replacing ATG. Inspection of the sequences of reads spanning the region of the start codon showed that only those copies beginning with canonical start codons are translated. Second copies of ribosomal protein L29, microtubule associated protein CAP15, and cytochrome oxidase subunit X were also detected, as well as several VSGs, a VSG-related CDS and an ESAG.

In addition, on chromosome 9, a single conserved hypothetical CDS annotated in genome version 5 is now joined by eleven distinct, but closely related non-annotated CDSs (plus two that were annotated in version 5.1, Fig. 1B). Of the 11 new CDSs, six were previously identified in unassigned contigs and now can be assigned to this chromosomal region, along with five new CDSs. The CDSs share identical N-termini which commence with a predicted signal sequence, and identical C-termini which bear a predicted transmembrane domain. The central regions of these CDSs contain variable numbers of imperfect hexapeptide repeats yielding products of 410–452 aa, although two genes (Tb927.9.2817 at 811 aa and Tb927.9.2821 at 614 aa) are larger, and have additional, more complex modules of repeats. Interestingly, while this work was in progress, members of this gene family, dubbed invariant glycoprotein 48 (IGP48), were shown to be localized to the endoplasmic reticulum and RNAi studies suggest that as a group they are essential for growth of BF [24].

Most of the novel CDSs are unique. An example shown in Fig. 1C depicts a 75 aa CDS (Tb927.9.1365) that lies on a recently identified transcript on chromosome 9 (Tb9.NT.3) [11]. The ribosome footprints (purple trace) span the region from the start codon (pink bar) just after the SL addition site to the nearest in-frame stop codon (black bar) to encode a conserved hypothetical protein. The mRNA reads continue downstream for about 300 nt. Adjacent to this locus is one that specifies another transcript (Tb9.NT.4) that shows no evidence of protein production in slender BF (Fig. 1C) or in PCF (not shown). In other cases, ribosome profiling allowed discrimination between two or more potential CDSs. As shown in Fig. 1D, a CDS was annotated on reading frame 3. However, the pattern of the ribosome footprints show that translation stops at the stop codon on reading frame 2. Thus the ORF that begins slightly before the annotated CDS represents the sequence that is translated. While the previously annotated CDS had no predicted orthologues in trypanosomatids, the new CDS (Tb927.9.4961) has syntenic, closely related orthologues across the trypanosomatids.

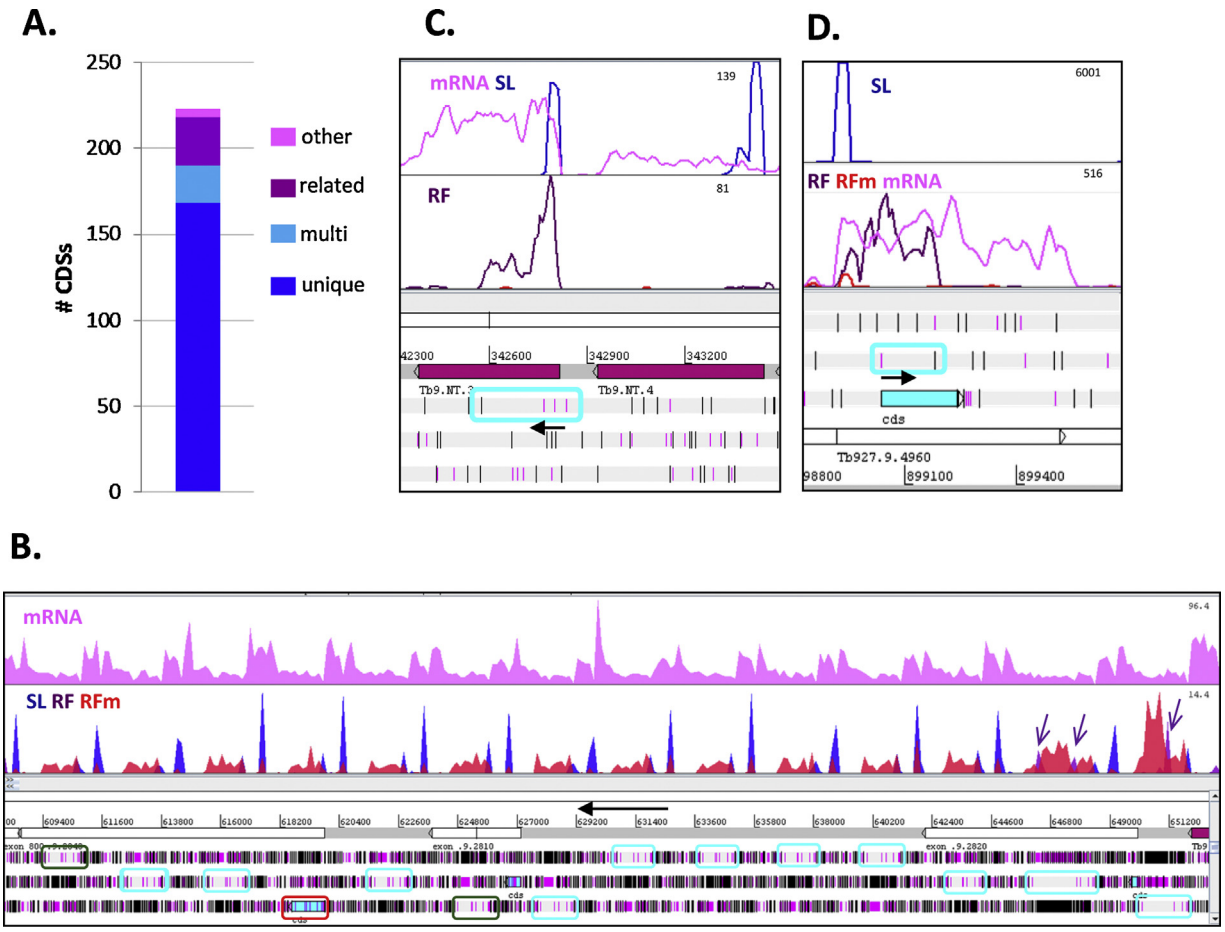

Fig. 1. New CDSs.
(A) The numbers of unique single and multicopy new CDS genes, along with those that are closely related to annotated genes are indicated. The group "other" includes genes that were known but not annotated or that were on unassigned contigs.
(B) A tandem array of related genes on chromosome 9. The upper two panels show the data from ribosome profiling and SL sequencing, while the lower panel shows the map of relevant strand of the Tb927 genome (v. 5.0). There, a single gene was correctly annotated (Tb927.9.2940, red box), but additional CDSs were revealed in this study making a tandem array of 14 genes. Those marked by green were annotated in version 5.1 (Tb927.9.2860, Tb927.9.2853) while those outlined in cyan remained incognito. The two genes at far right have some unique reads (Tb927.9.2817 and Tb927.9.2821, purple arrows). In this and all images of Artemis screen shots, the traces of slender bloodstream form sequencing reads are color-coded: mRNA (pink), SL (blue), ribosome footprint unique reads (RF, purple) and multi-mapping ribosome footprint reads (RFm, red). The numbers in the upper right indicate the peak number of reads in the panel, using a linear scale. In the genome map, black vertical bars indicate stop codons and pink bars indicate methionine codons for each reading frame shown. Black arrows above the gene models indicate the direction of transcription/translation.
(C) New CDS Tb927.9.1635 (outlined in cyan) that lies on the transcript Tb9.NT.3.
(D) Reading frame change. The new CDS Tb927.9.4961 (outlined in cyan) was identified by virtue of its stop codon that paralleled the drop in ribosome footprints.

Similarly, the reading frame of Tb927.9.6540 was corrected and a new identifier provided (Tb927.9.4959).

RNAs that are not considered protein-coding, such as telomerase RNA and the MRP RNA (present in three copies) showed few reads in the ribosome footprint libraries. Moreover, these reads did not map to ORFs, indicating that they do not reflect the presence of translating ribosomes (data not shown). Of the >1100 novel transcripts previously described [11], we found that fewer than 10% showed clear evidence of protein production, suggesting they play other roles in parasite biology. In contrast, only 22 annotated CDSs with mRNA levels above the lowest quartile had negligible ribosome footprints (<180 total summed across all libraries).

3.2. Phylogenetic analysis and functional attributes

We examined the evolutionary conservation of the 178 new unique CDSs. As illustrated in Fig. 2A, left and detailed in Table S2, Blastp searches of genomes from a variety of kinetoplastids and well-studied representatives of higher eukaryotes revealed homologues for 90 of the new CDSs. The large majority of the conservation was within kinetoplastids, with only eight CDSs showing homologues in higher eukaryotes. There were numerous gaps in the pattern for most CDSs, raising the possibility that the lack of hits was a result of incomplete annotation in some species. Hence we used tblastN to search the same genomes. With this strategy, the patchiness in the patterns disappeared and 170 CDSs showed potential homologues in other species or *T. brucei* subspecies (Fig. 2A, right). This finding likely indicates that many CDSs remain to be annotated across these species. Only six new CDSs showed tblastN hits in higher eukaryotes (fewer than recognized by blastp, possibly as a result of introns), but 35 extended beyond the family Trypanosomatidae to the other Kinetoplastid family Bodonidae (*Bodo saltans*). On the other hand, 67 CDSs showed potential homologues only in the closely related subspecies *T. b. gambiense* (nine were found only in *T. b. brucei*, being present in both Tb927 and Lister 427).

The CDSs restricted to the *T. brucei* subspecies are likely the youngest, arising well after the divergence of African (salivarian) trypanosomes some 300 million years ago [25]. The characteristics of these CDSs as compared to the more highly conserved CDSs are illustrated in Figs. 2B and C, and detailed in Table S3. Interest-

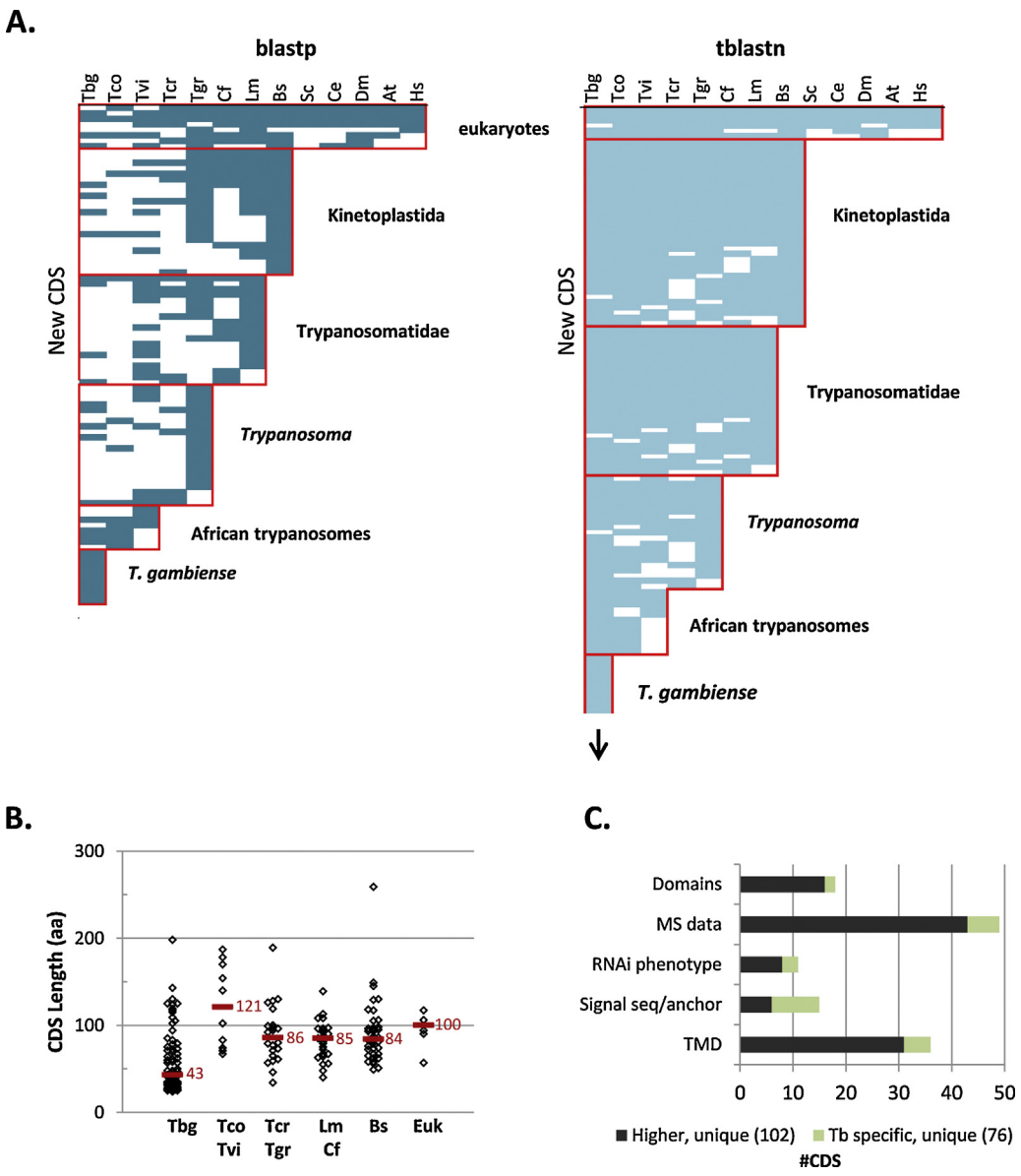

Fig. 2. Conservation and features of new CDSs.
(A) The unique new CDSs (arrayed vertically) were subjected to Blastp and tBlastn analysis against selected genomes arrayed from left to right: *T. brucei gambiense* DAL972 (Tbg), *T. congolense* (Tco), *T. vivax* (Tvi), *T. cruzi* (Tcr), *T. grayi* (Tgr), *Crithidia fasciculata* (Cf), *Leishmania major* Friedlin (Lm), *Bodo saltans* (Bs), *Saccharomyces cerevisiae* (Sc), *Caenorhabditis elegans* (Ce), *Drosophila melanogaster* (Dm), *Arabidopsis thaliana* (At), and *Homo sapiens* (Hs). Those which reached the desired cutoff (*E*-value ≥ 0.001, 30% identity, and 50% coverage), are marked by a shaded box. CDSs for which no similar proteins were detected are omitted. Arrow indicates that only some of the 67 CDSs with homology only to *T. gambiense* ORFs are shown.
(B) Unique new CDS length, stratified by highest phylogenetic conservation revealed by tBlastn. Numbers and the red bar indicate the median length for each subset. The graph does not show one 2853 aa CDSs that is present in *T. brucei* and *T. gambiense*. The higher median length of the CDSs found throughout African trypanosomes (121 aa) may result from the small sample size.
(C) Characteristics of unique new CDSs. The graph depicts the number of the 178 unique new CDSs with various attributes, parsed according to their representation on the CDSs restricted to *T. brucei* subspecies (Tb specific, green) *versus* those conserved in additional organisms ("Higher", black). Domains included Pfam and Interpro domains. MS data was as summarized [10,14]. RNAi phenotype data was from [10], which tested 42 candidate CDSs. Signal sequences and signal anchor sequences were predicted by SignalP [19] and transmembrane domains (TMD) by TMHMM [20]. See Table S3 for gene-level detail.

ingly, within the unique set of CDSs, the median length of the CDSs restricted to the *T. brucei* subspecies was considerably smaller than that of those conserved in other species (43 *vs* 93 aa) (Fig. 2B). This finding is congruent with current thinking on *de novo* gene evolution as beginning with short translated ORFs that can be extended over time and acquire function [26]. Indeed when we searched the unique CDSs for Pfam and Interpro domains of known function, only three of the 76 CDSs found only in *T. brucei* subspecies had such domains, while 16 of the remaining 102 CDSs had identifiable domains. Interestingly, unlike similar short translated ORFs conserved only in the closely related species of genus *Saccharomyces*, which showed enrichment of transmembrane domains, *T. brucei* species-specific CDSs have fewer transmembrane domains (one per 526 aa, after excluding the 2853 aa repetitive protein that contains no transmembrane domains) as compared to the CDSs conserved in additional species (one per 224 aa). However, signal or signal anchor sequences were more highly represented in *T. brucei*-specific CDSs.

Table S3 also provides edgeR-normalized read counts for both ribosome profiling and mRNA for the new CDS for animal-derived

slender BF, *in vitro* cultured BF, and PCF. The new CDSs showed a broad range of protein production (>150-fold, as measured by ribosome footprints reads per kb), with four new unique CDSs expressed at levels similar to ribosomal proteins in PCF. Interestingly, 13 of the CDSs showed differential protein production in PCF as compared to both *in vivo* derived BF and cultured BF (>2-fold change, FDR<0.01 as compared to all *T. brucei* CDSs), and 38 showed increased protein production in both *in vivo*-derived and cultured BF as compared to PCF, suggesting potential functional roles in the parasite.

3.3. CDS elimination

We consolidated our previous data on genes unlikely to encode proteins by harmonizing with the Tb927 genome v5.1 gene list, reinstating some CDSs because of incomplete data, and eliminating additional CDSs following further inspection. Together this resulted in 683CDS recommended for elimination (of which 643 were in our previous list [13]). CDSs were eliminated when the patterns of mRNA, SL, and ribosome footprint reads were incompatible with the CDS model (as is, or with an N-terminal extension or truncation). However, CDSs were not eliminated solely because of lack of ribosome or mRNA footprints, since we only examined two developmental stages of the parasite, so thus could not eliminate the possibility that these putative CDSs were translated in other stages. A modest number of CDSs were eliminated because the mRNA and/or ribosome footprint reads were on the opposite strand. An example is shown in Fig. 3A, where the region of the genome shows no experimental evidence of a CDS and the mRNA maps to the opposite strand. In Fig. 3B, ribosome profiling reveals a CDS on the opposite strand (Tb927.9.13225, encoding a hypothetical conserved protein outlined in cyan) compared to the annotated CDS. Fig. 3C shows two examples of CDSs that were removed. The 5′ putative CDS (red arrow) spans two transcripts. The ribosome footprints lie only in the 3′ half of the putative CDS, and the mRNA and spliced leader data show the footprints map to the 5′ UTR of the downstream CDS Tb927.11.10300. We and others have previously shown that ribosome footprints are common in the 5′ UTRs of translated CDSs [13,27]. The brown arrow marks an annotated CDS that clearly lies within the 3′ UTR of Tb927.11.10300, as there is no intervening SL site and the mRNA reads extend at the same level. Thus in the region depicted, of the three CDSs, two were eliminated and one retained. Of the 683 CDSs removed, 625 were annotated as "hypothetical" or "hypothetical, unlikely". In fact, 81% of the genes described as "hypothetical, unlikely" were removed in this process. Most of the other removed genes were designated hypothetical, conserved or "unspecified product" leaving only 12 with product descriptions. Table S1 provides gene-level details.

3.4. CDS changes

In many cases the patterns of mRNA, SL, and ribosome footprint reads were compatible with simple in-frame extensions or truncations of the N-terminus of the annotated CDS. In some cases, these changes yielded alterations that resulted in obvious functional changes. As shown in Fig. 4A, extension of the CDS of Tb927.10.11160 adds 117 aa and a new Pfam domain (Nfu-N). Annotated as Nfu2 (and as HIRA-interacting protein 5), the encoded protein is likely involved in biosynthesis of iron-sulfur clusters [28]. Similarly, the annotated Tb917.10.13560 (Fig. 4B) encodes a "half"-transporter. Extension of the 5′ end adds four transmembrane domains (lower panel). The revised CDS encodes a protein with homology to sphingolipid transporters and is almost identical to Tb927.10.14090, with the exception of the C-terminal tail. Truncations can alter the predicted function of a protein, as shown in Fig. 4C. Tb927.9.14000, which encodes ribosomal protein L12, uses the second methionine of the annotated CDS, as shown by the combined data. This start site keeps Pfam domains intact (blue line), but it removes a predicted signal sequence, which would sequester the protein from the pool of functional ribosomal proteins. For some genes, alternatively spliced isoforms are predicted that differ at their 5′ end. In most cases, our data can distinguish which of these is correct. For example, of the two isoforms of a putative sodium/hydrogen exchanger annotated in TriTrypDB (Tb927.11.840), our data supports isoform 1 (red arrow, Fig. 4D). Isoform 2 (brown arrow) commences at an ATG that does not lie on the same transcript and that lacks ribosome footprints. Hence it does not function as the start codon for this CDS.

3.5. Genome sequence refinement

In five cases, the patterns of ribosome profiling indicated potential issues with the underlying genome sequences. These are shown in Fig. 5 and in Fig. S2. Fig. 5A shows a case where ribosome footprints spanned two CDSs (and the "intergenic" region), suggesting that they represent a single CDS. Further analysis of this region using PCR (all primers are listed in Table S4) and sequencing showed the presence of a 5-base insertion and 1-base deletion that joins the two annotated CDSs into a single open reading frame. This fusion joins a predicted signal sequence on Tb927.10.4030 to an *s*-adenosyl methionine-dependent methyl transferase domain present on Tb927.10.4020 in a single polypeptide. Two additional examples of CDS extensions due to indels are shown in Fig. S2A and B.

Fig. 5B depicts the ribosome footprints, mRNA, and SL reads across Tb927.11.12880, annotated as a protein of the outer mitochondrial membrane [29]. The ribosome footprints show greatly increased protein production at the 3′ end of this ORF. Nonetheless, the 5′ portion appears to be translated as shown by the rescaled panel at top, which shows low abundance ribosome footprints. These two regions were separated by a large peak of multi-mapping reads. To assess whether this could represent a misassembled region of the genome, we performed PCR on genomic DNA using primers that flanked the large peak. As shown, the resulting products were larger than 5 kb, when the predicted size was ~850 bp. Interestingly, *T. brucei* strain 427 and *T. b. gambiense* both have two predicted CDSs that are related to Tb927.12880; in the first case the CDSs are adjacent, while in the second they are distant. All of these CDSs have multiple 21 aa repeats (and numerous nt sequence ambiguities). The N-terminal portion of Tb927.11.12880 is most similar to one CDS while the C-terminal portion is most similar to the other. This information, taken together with the large peak of mRNA and ribosome footprint reads, suggests that there is an intervening region of repetitive sequence separating two regions of Tb927.11.12880, and that these regions correspond to separate CDSs. However, since we were unable to obtain the intervening sequence to separate the genes, we have not modified the current gene model.

A different example in which the experimental data suggested two CDSs replacing a single annotated CDS is shown in Fig. 5C. Here, the gene Tb927.11.4650, which encodes a protein named MRPL52/Kripp13 [30,31], appears to specify two mRNAs and two protein coding regions. MRPL52 was identified in pull-downs of mitochondrial ribosomes and peptides corresponding to the downstream region were detected. The authors of one study [30] remarked that the size of the predicted *T. brucei* protein was much larger than the human orthologue (1522 *vs* 77 aa)—the redrawn boundary yields a 259 aa protein, reducing the discrepancy across species. Surprisingly, the upstream gene (renamed here to Tb927.11.4649) lacks a stop codon, even though the mRNA trace shows it lies on a separate transcript. No inconsistencies in the underlying genome sequence were found upon PCR and sequenc-

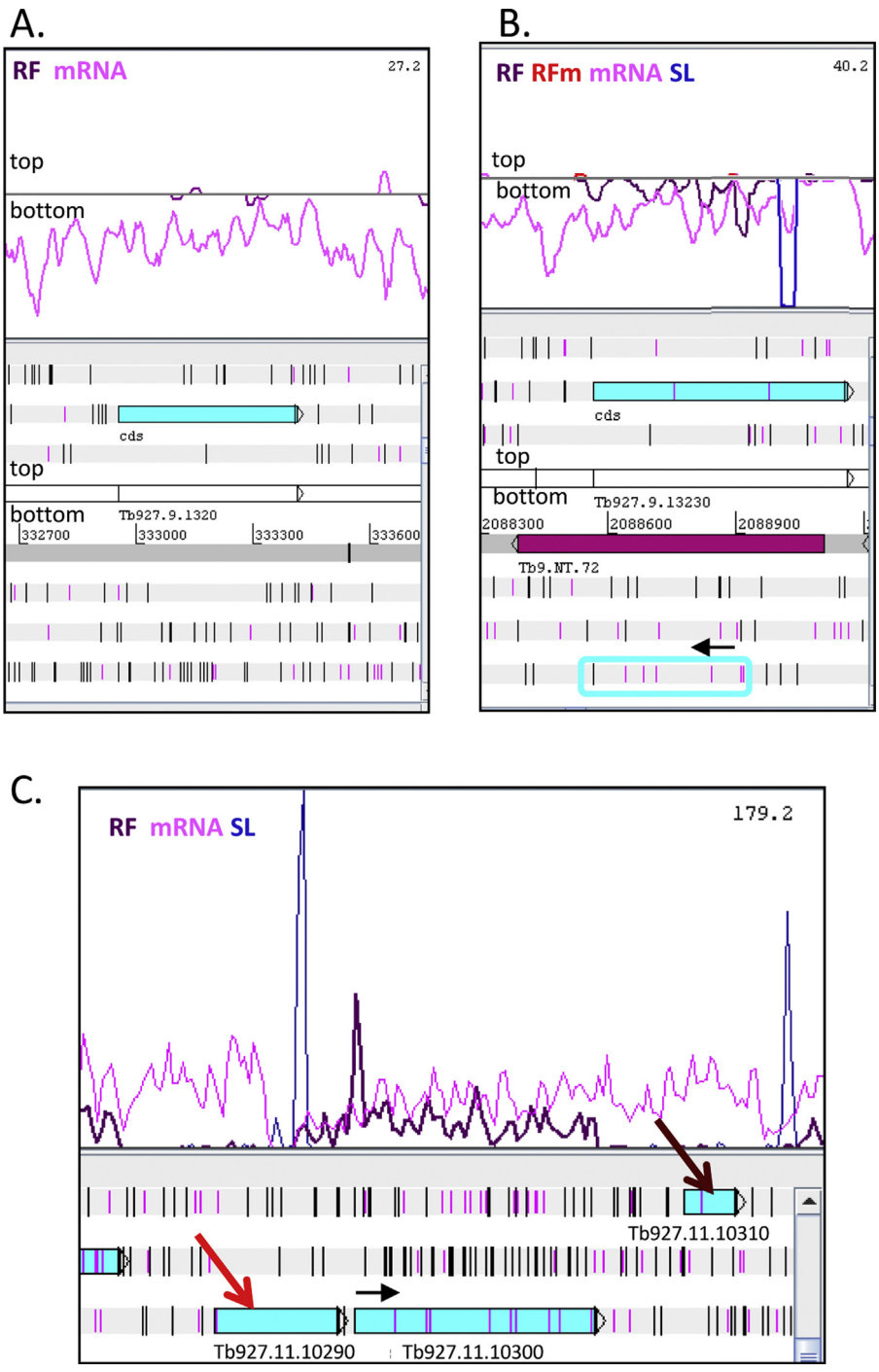

Fig. 3. Examples of CDSs that were eliminated.
(A) Example of CDS (Tb927.9.1320) removed due to mRNA reads being on the opposite strand. There are no SL reads in the vicinity. See Fig. 1B legend for a description of Artemis images.
(B) Example of CDS (Tb927.9.13230) eliminated due to opposite strand reads. A new CDS (Tb927.9.13225) is on the opposite (bottom) strand. SL reads are rescaled (peak height 201).
(C) This image shows two CDSs that were removed. The red arrow points to an example of a CDS (Tb927.11.10290) eliminated because it has both an internal SL and is not spanned by a distinct mRNA. The few ribosome profiling reads that map to the region of this "CDS" are in the 5' UTR of the downstream gene. The brown arrow points to an example of a CDS removed (Tb927.11.10310) because it lacks a dedicated SL and lacks a distinct RNA. It resides in the 3' UTR of the upstream gene.

ing, although we note that in *T. congolense* this region is annotated as two genes. Fig S2C shows an example suggesting that two protein-coding transcripts may arise from the gene Tb927.9.2130, one of which encodes the C-terminal 334 aa of the 1425 aa full-length protein. The data show the presence of an SL site preceding an increase in both mRNA and ribosome footprint reads over the C-terminal region.

4. Conclusions

The use of ribosome profiling coupled with mRNA and SL mapping allowed here the unbiased verification of 225 novel CDSs. Only 28 CDSs larger than 150 aa were discovered, and most of these were members of multi-gene families for which one or more initial members had been previously identified. If anything, the Tb927 genome

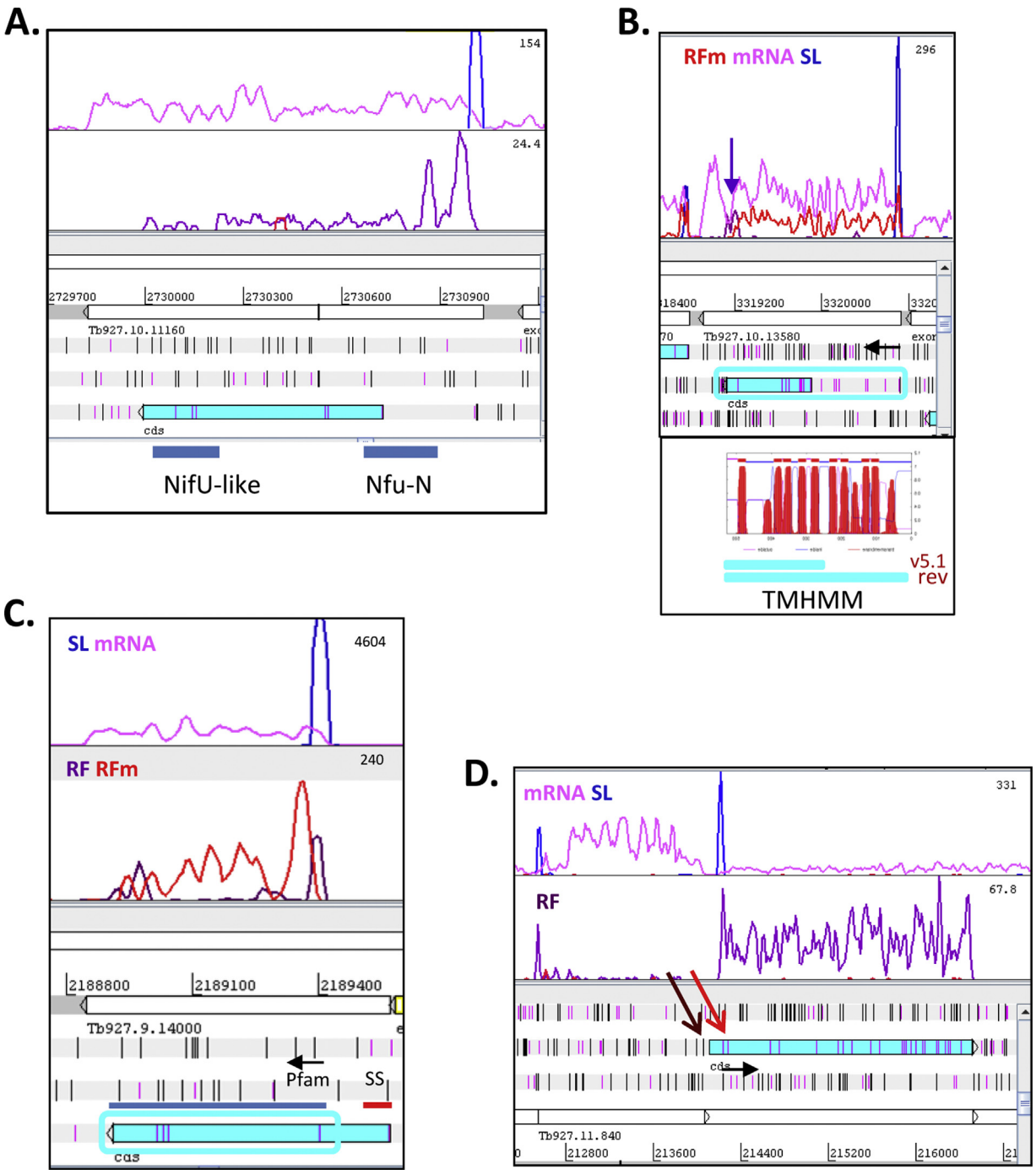

Fig. 4. Changes to coordinates of annotated genes.
(A) Extension of Tb927.10.11160 CDS 5′ end adds 117 aa and a new Pfam domain (Nfu-N). Annotated as Nfu2 (and as HIRA-interacting protein 5), the encoded protein is likely involved in biosynthesis of iron-sulfur clusters [28]. See Fig. 1B legend for a description of Artemis images.
(B) As previously annotated, Tb917.10.13560 encoded a half-transporter. Extension of the 5' end adds four transmembrane domains (lower panel) to the revised CDS (rev) as compared to the version in genome version 5.1. The encoded protein has homology to sphingolipid transporters and is almost identical to Tb927.10.14090, with the exception of the C-terminal tail (the purple arrow in the top panel marks the unique sequence peak).
(C) Tb927.9.14000 ribosomal protein L12 uses the second methionine of the annotated CDS. This start site keeps Pfam domains intact (blue bar), but eliminates the signal sequence (red bar).
(D) Two isoforms of the CDS (Tb927.11.840) are currently annotated in TriTrypDB. The red arrow marks isoform 1, where ribosome footprint reads begin. Isoform 2 (brown arrow) commences at an ATG that does not lie on the transcript as can be seen from the mRNA and SL reads.

annotation errs on the side of over-prediction, since our ribosome profiling data shows that approximately 8% (683) of annotated CDSs (version 5.1) appear not to encode proteins as determined by incompatibility of gene models with the data, combined with lack of supporting positive evidence. Additionally, 510 of currently annotated CDSs lack even minimal evidence of protein production (<180 ribosome footprints total across all libraries) in these biological samples. These include 276 hypothetical or hypothetical unlikely CDSs, which may or may not be true CDSs. The remaining CDSs lacking detectable protein production are named genes (51) and conserved hypothetical proteins (102), which may be expressed in other stages, as well as numerous VSGs, which are expressed

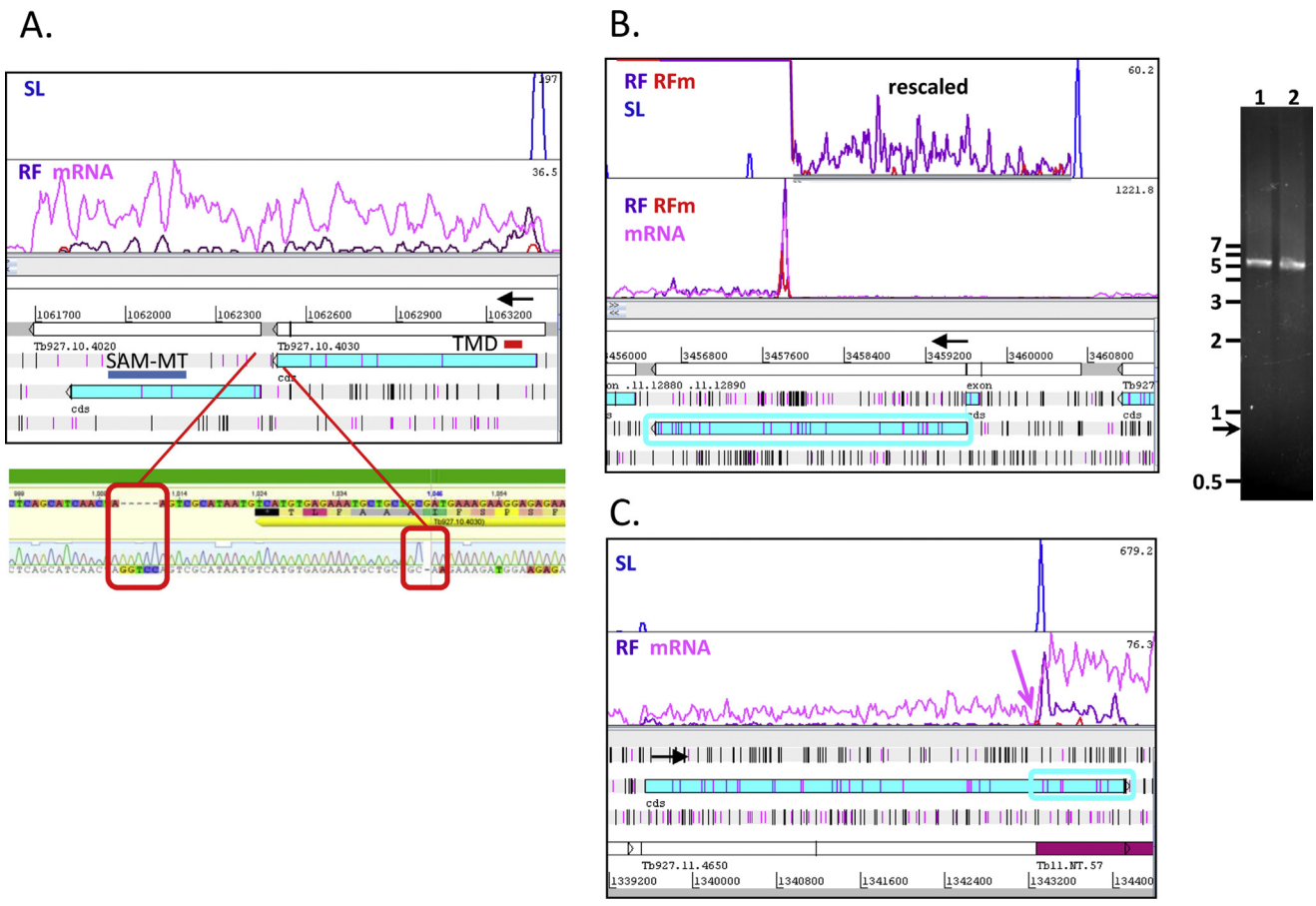

Fig. 5. Ribosome profiling suggests changes in underlying genome sequence.
(A) Fusion of two CDSs. The ribosome footprints, mRNA reads and SL mapping all suggest that the two annotated CDSs could encode one contiguous protein, although the two appear to be on different reading frames. PCR amplification and DNA sequencing revealed an additional 5 nt and a deletion of 1 nt, which results in a frameshift, joining the two CDSs in one reading frame. Thus the transmembrane domain (TMD) of Tb927.10.4030 is juxtaposed with the s-adenosylmethionine-dependent methyl transferase domain (SSF 53335) from Tb927.10.4020 to create a single CDS (which retains the name Tb927.11.4020). See Fig. 1B legend for a description of Artemis images.
(B) One annotated CDS encodes distinct proteins. The ribosome profiling data for Tb927.11.12890 shows a highly expressed region at the 3' end (middle panel, peak height 1221). The 5' portion does appear to be translated as shown by the rescaled panel at top, which shows ribosome footprints (peak height 60). PCR analysis was performed using two sets of primers on either side of the junction region. The expected sizes of the products based on the sequence were 877 bp (lane 1) and 810 bp (lane 2), which would flank the arrow marking 850 bp. As shown on right, the observed products were 5313 bp and 5193 bp respectively. No products were observed using single primers. Attempts to generate a sufficient quantity of the PCR product for sequencing were unsuccessful and we were unable to clone the product.
(C) Tb11.NT.57 (magenta) specifies a highly translated CDS within the 3′ portion of the annotated CDS Tb927.11.4650. Note the absence of mRNA reads just prior to the SL site for Tb927.NT.57 (arrow), as well as the more abundant ribosome footprints in the latter, suggesting that these two regions are separate transcripts that encode non-overlapping proteins. Surprisingly, there is no stop codon before the end of the Tb927.11.4650 mRNA. PCR across this junction did not reveal any errors in the underlying genome sequence that would add a stop codon.

clonally. Finally, the accurate evaluation of regions undergoing protein production allowed over 400 gene models in version 5.1 to be corrected, defining the proper start codon. Of the unique CDSs confirmed here, over half have candidate orthologues in other trypanosomatid genomes, most of which were not yet annotated as genes. Additionally, approximately 40% of the new unique CDSs were present only in *T. brucei* subspecies, suggesting that many are young genes or proto-genes. Indeed these tended to have far fewer functional domains and were shorter that conserved CDSs.

Acknowledgements

EJRV was supported by a fellowship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (process ID: PDE 202223/2012-4). This work was supported by grant R21 AI094129 from the National Institutes of Health. The authors are solely responsible for its content.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.molbiopara.2015.09.002.

References

[1] O. Cirovic, T. Ochsenreiter, Whole proteome analysis of the protozoan parasite *Trypanosoma brucei* using stable isotope labeling by amino acids in cell culture and mass spectrometry, Methods Mol. Biol. 1188 (2014) 47–55.
[2] F. Butter, F. Bucerius, M. Michel, Z. Cicova, M. Mann, C.J. Janzen, Comparative proteomics of two life cycle stages of stable isotope-labeled *Trypanosoma brucei* reveals novel components of the parasite's host adaptation machinery, Mol. Cell. Proteomics 12 (2013) 173–179.
[3] A.P. Jackson, H.C. Allison, J.D. Barry, M.C. Field, C. Hertz-Fowler, M. Berriman, A cell-surface phylome for African trypanosomes, PLoS Negl. Trop. Dis. 7 (2013) e2121.
[4] J.F. Ryley, Studies on the metabolism of the protozoa: 9 comparative metabolism of bloodstream and culture forms of *Trypanosoma rhodesiense*, Biochem. J. 85 (1962) 211–223.

[5] K. Vickerman, Polymorphism and mitochondrial activity in sleeping sickness trypanosomes, Nature 208 (1965) 762–766.
[6] M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D.C. Bartholomeu, N.J. Lennard, E. Caler, N.E. Hamlin, B. Haas, U. Bohme, L. Hannick, M.A. Aslett, J. Shallom, L. Marcello, L. Hou, B. Wickstead, U.C. Alsmark, C. Arrowsmith, R.J. Atkin, A.J. Barron, F. Bringaud, K. Brooks, M. Carrington, I. Cherevach, T.J. Chillingworth, C. Churcher, L.N. Clark, C.H. Corton, A. Cronin, R.M. Davies, J. Doggett, A. Djikeng, T. Feldblyum, M.C. Field, A. Fraser, I. Goodhead, Z. Hance, D. Harper, B.R. Harris, H. Hauser, J. Hostetler, A. Ivens, K. Jagels, D. Johnson, J. Johnson, K. Jones, A.X. Kerhornou, H. Koo, N. Larke, S. Landfear, C. Larkin, V. Leech, A. Line, A. Lord, A. MacLeod, P.J. Mooney, S. Moule, D.M. Martin, G.W. Morgan, K. Mungall, H. Norbertczak, D. Ormond, G. Pai, C.S. Peacock, J. Peterson, M.A. Quail, E. Rabbinowitsch, M.A. Rajandream, C. Reitter, S.L. Salzberg, M. Sanders, S. Schobel, S. Sharp, M. Simmonds, A.J. Simpson, L. Tallon, C.M. Turner, A. Tait, A.R. Tivey, S. Van Aken, D. Walker, D. Wanless, S. Wang, B. White, O. White, S. Whitehead, J. Woodward, J. Wortman, M.D. Adams, T.M. Embley, K. Gull, E. Ullu, J.D. Barry, A.H. Fairlamb, F. Opperdoes, B.G. Barrell, J.E. Donelson, N. Hall, C.M. Fraser, S.E. Melville, N.M. El Sayed, The genome of the African trypanosome *Trypanosoma brucei*, Science 309 (2005) 416–422.
[7] G.A. Cross, H.S. Kim, B. Wickstead, Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427, Mol. Biochem. Parasitol. 195 (2014) 59–73.
[8] N. Hall, M. Berriman, N.J. Lennard, B.R. Harris, C. Hertz-Fowler, E.N. Bart-Delabesse, C.S. Gerrard, R.J. Atkin, A.J. Barron, S. Bowman, S.P. Bray-Allen, F. Bringaud, L.N. Clark, C.H. Corton, A. Cronin, R. Davies, J. Doggett, A. Fraser, E. Gruter, S. Hall, A.D. Harper, M.P. Kay, V. Leech, R. Mayes, C. Price, M.A. Quail, E. Rabbinowitsch, C. Reitter, K. Rutherford, J. Sasse, S. Sharp, R. Shownkeen, A. MacLeod, S. Taylor, A. Tweedie, C.M. Turner, A. Tait, K. Gull, B. Barrell, S.E. Melville, The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism, Nucleic Acids Res. 31 (2003) 4864–4873.
[9] S. El, N.M. ayed, E. Ghedin, J. Song, A. MacLeod, F. Bringaud, C. Larkin, D. Wanless, J. Peterson, L. Hou, S. Taylor, A. Tweedie, N. Biteau, H.G. Khalak, X. Lin, T. Mason, L. Hannick, E. Caler, G. Blandin, D. Bartholomeu, A.J. Simpson, S. Kaul, H. Zhao, G. Pai, A. Van, S. ken, T. Utterback, B. Haas, H.L. Koo, L. Umayam, B. Suh, C. Gerrard, V. Leech, R. Qi, S. Zhou, D. Schwartz, T. Feldblyum, S. Salzberg, A. Tait, C.M. Turner, E. Ullu, O. White, S. Melville, M.D. Adams, C.M. Fraser, J.E. Donelson, The sequence and analysis of *Trypanosoma brucei* chromosome II, Nucleic Acids Res. 31 (2003) 4856–4863.
[10] M. Ericson, M.A. Janes, F. Butter, M. Mann, E. Ullu, C. Tschudi, On the extent and role of the small proteome in the parasitic eukaryote *Trypanosoma brucei*, BMC Biol. 12 (2014) 14.
[11] N.G. Kolev, J.B. Franklin, S. Carmi, H. Shi, S. Michaeli, C. Tschudi, The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution, PLoS Pathog. 6 (2010) e1001090.
[12] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman, Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling, Science 324 (2009) 218–223.
[13] B.C. Jensen, G. Ramasamy, E.J. Vasconcelos, N.T. Ingolia, P.J. Myler, M. Parsons, Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*, BMC Genomics 15 (2014) 911.
[14] J.J. Vasquez, C.C. Hon, J.T. Vanselow, A. Schlosser, T.N. Siegel, Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages, Nucleic Acids Res. 42 (2014) 3623–3637.
[15] P. Smircich, G. Eastman, S. Bispo, M.A. Duhagon, E.P. Guerra-Slompo, B. Garat, S. Goldenberg, D.J. Munroe, B. Dallagiovanna, F. Holetz, J.R. Sotelo-Silveira, Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in *Trypanosoma cruzi*, BMC Genomics 16 (2015) 443.
[16] T.N. Siegel, D.R. Hekstra, X. Wang, S. Dewell, G.A. Cross, Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites, Nucleic Acids Res. 38 (2010) 4946–4957.
[17] D. Nilsson, K. Gunasekera, J. Mani, M. Osteras, L. Farinelli, L. Baerlocher, I. Roditi, T. Ochsenreiter, Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*, PLoS Pathog. 6 (2010) e1001037.
[18] T. Carver, S.R. Harris, M. Berriman, J. Parkhill, J.A. McQuillan, Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data, Bioinformatics 28 (2012) 464–469.
[19] O. Emanuelsson, S. Brunak, H.G. von, H. Nielsen, Locating proteins in the cell using TargetP, SignalP and related tools, Nat. Protoc. 2 (2007) 953–971.
[20] A. Krogh, B. Larsson, H. von, G. eijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[21] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140.
[22] S. Wong, T.H. Morales, D.A. Campbell, Ubiquitin-EP52 fusion protein homologs from *Trypanosoma brucei*, Nucleic Acids Res. 18 (1990) 7181.
[23] T.N. Nguyen, B. Schimanski, A. Zahn, B. Klumpp, A. Gunzl, Purification of an eight subunit RNA polymerase I complex in *Trypanosoma brucei*, Mol. Biochem. Parasitol. 149 (2006) 27–37.
[24] H. Allison, A.J. O'Reilly, J. Sternberg, M.C. Field, An extensive endoplasmic reticulum-localised glycoprotein family in trypanosomatids, Microb. Cell 1 (2014) 325–345.
[25] J. Haag, C. O'hUigin, P. Overath, The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria, Mol. Biochem. Parasitol. 91 (1998) 37–49.
[26] A.R. Carvunis, T. Rolland, I. Wapinski, M.A. Calderwood, M.A. Yildirim, N. Simonis, B. Charloteaux, C.A. Hidalgo, J. Barbette, B. Santhanam, G.A. Brar, J.S. Weissman, A. Regev, N. Thierry-Mieg, M.E. Cusick, M. Vidal, Proto-genes and de novo gene birth, Nature 487 (2012) 370–374.
[27] N.T. Ingolia, G.A. Brar, N. Stern-Ginossar, M.S. Harris, G.J. Talhouarne, S.E. Jackson, M.R. Wills, J.S. Weissman, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes, Cell Rep. 8 (2014) 1365–1379.
[28] J. Lukes, S. Basu, Fe/S protein biogenesis in trypanosomes—a review, Biochim. Biophys. Acta 1853 (2014) 1481–1492.
[29] M. Niemann, S. Wiese, J. Mani, A. Chanfon, C. Jackson, C. Meisinger, B. Warscheid, A. Schneider, Mitochondrial outer membrane proteome of *Trypanosoma brucei* reveals novel factors required to maintain mitochondrial morphology, Mol. Cell. Proteomics 12 (2013) 515–528.
[30] A. Zikova, A.K. Panigrahi, R.A. Dalley, N. Acestor, A. Anupama, Y. Ogata, P.J. Myler, K. Stuart, *Trypanosoma brucei* mitochondrial ribosomes: affinity purification and component identification by mass spectrometry, Mol. Cell. Proteomics 7 (2008) 1286–1296.
[31] I. Aphasizheva, D. Maslov, X. Wang, L. Huang, R. Aphasizhev, Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes, Mol. Cell 42 (2011) 106–117.