Published in final edited form as:

Trends Parasitol. 2011 October ; 27(10): 434–441. doi:10.1016/j.pt.2011.05.006.

Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing

T. Nicolai Siegel^{1,2}, Kapila Gunasekera⁴, George A. M. Cross³, and Torsten Ochsenreiter⁴

¹ Biology of Host-Parasite Interactions, Parasitology, Institut Pasteur, 75724 Paris, France ² CNRS, URA2581, Institut Pasteur, 75724 Paris, France ³ Laboratory of Molecular Parasitology, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA ⁴ Institute of Cell Biology, University of Bern, Baltzerstrasse 4, 3012 Bern, Switzerland

Abstract

Trypanosoma brucei undergoes major biochemical and morphological changes during its development from the bloodstream form (BF) in the mammalian host to the procyclic form (PF) in the midgut of its insect host. The underlying regulation of gene expression, however, is poorly understood. More than 60% of the predicted genes remain annotated as hypothetical and the 5' and 3' untranslated regions (UTRs) important for regulation of gene expression are unknown for more than 90% of the genes. In this review we compare the data from four recently published high throughput RNA sequencing studies in light of the different experimental setups and discuss how these data can enhance genome annotation and give insights into the regulation of gene expression in *T. brucei*.

Polycistronic transcription in trypanosomes

Protein-coding genes in *T. brucei* are transcribed polycistronically [1,2] and are rapidly processed into monocistronic mRNAs by coupled *trans*-splicing and polyadenylation (Figure 1) [3,4]. Short polycistronic transcription units (PTUs) have also been identified in nematodes [5] and dicistronic units have been reported for *Drosophila* [6] and humans [7], but the significance of polycistronic transcription in *T. brucei* remains unknown. Genes within a PTU are transcribed from the same strand of DNA, while transcription of two neighboring PTUs located on opposite strands can be either convergent or divergent. The regions between PTUs located on opposite strands are referred to as strand-switch regions (SSRs) (Figure 1). With one exception, no other RNA polymerase II (pol II) promoter motif has been identified in *T. brucei*, and how transcription is initiated remains enigmatic [8,9].

Indications that RNA pol II transcription may initiate at divergent-SSRs came from strand-specific nuclear run-on assays performed in *Leishmania* [10], which showed that RNA pol II transcription starts at SSRs between two transcriptionally divergent PTUs (divergent-SSRs) and ends at SSRs between two transcriptionally convergent PTUs (convergent-SSRs). Given the high synteny between *T. brucei* and *L. major* [11], it was reasonable to hypothesize that divergent-SSRs in *T. brucei* are also transcription start sites.

© 2011 Elsevier Ltd. All rights reserved.

Correspondence to: Torsten Ochsenreiter.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Genome-wide analyses in *T. brucei* revealed enrichment of two histone modifications, H4K10ac and H3K4me3, and four histone variants, at convergent or divergent-SSRs — regions hypothesized to represent RNA pol II transcription start sites (TSSs) and transcription termination sites (TTSs) [12,13]. The same chromatin components were found at 60 non-SSR sites, suggesting the presence of additional RNA pol II TSSs between genes arrayed on the same DNA strand [12,13] or other functions mediated by the same modifications.

RNA maturation

To generate mature mRNAs, the primary RNA transcript is *trans*-spliced and polyadenylated. *Trans* splicing serves two functions: it dissects mRNAs from polycistronic primary transcripts and adds the 39-nt spliced leader (SL) sequence containing the cap structure required for translation and possibly other aspects of mRNA function [14]. The precursor of the SL RNA is transcribed independently from the polycistronic primary transcript, so this process of joining two independently transcribed exons is referred to as *trans* splicing, in contrast to conventional *cis* splicing where an internal RNA fragment is removed from one of the precursor transcripts [15,16]. *Trans* and *cis* splicing share remarkable similarities. Both require the same characteristic sequence motifs, a polypyrimidine tract [poly(Y) tract], a GT dinucleotide at the 5' splice site, an AG dinucleotide at the 3' splice acceptor site and, possibly, exonic enhancer motifs [17–19]. Both *trans* and *cis* splicing follow the same general mechanism, which involves two consecutive catalytic *trans*-esterification reactions. Most major components of the yeast and human spliceosomes are conserved in *T. brucei* (reviewed in [20]). With the exception of poly(A) polymerase (Tb927.03.3160) and an ATP-dependent DEAD/H RNA helicase (Tb927.03.3160), *cis*-splicing has not been observed in *T. brucei*. Because this unique feature of these two genes is conserved among kinetoplastids, one wonders whether the introns themselves might have an important but undiscovered function.

Poly(A) tails play an important role in RNA stability by providing a binding site for poly(A)-binding proteins (PABPs), which prevent decapping of the 5' end of mRNA (reviewed in [22]), one of the early steps in mRNA degradation. Experiments in *T. brucei* show that depletion of CAF1 (Ccr4-associated factor 1), an enzyme that degrades poly(A) tails, increases mRNA stability [23]. A detailed mutational analysis revealed that the same sequence motifs affect polyadenylation of one gene and *trans*-splicing of the neighboring downstream gene in the same PTU, suggesting that both reactions are functionally coupled [4,21].

Role of 3' UTRs in RNA stability

The organization of genes in large PTUs provides no obvious way of regulating the transcription of individual genes. Yet, differentiation of the parasite from one life-cycle stage to the next is accompanied by significant changes in parasite gene expression, metabolism and morphology, and numerous studies have reported life-cycle-dependent regulation of mRNA and protein abundance. It is generally assumed that such regulation occurs post-transcriptionally at the level of *trans* splicing and polyadenylation, RNA export, RNA stability, protein translation, and protein stability [24,25]. More recently, RNA stabilization and degradation have been shown to be the major post-transcriptional mechanisms responsible for the differential expression of several genes in *T. brucei* including the three isoenzymes of phosphoglycerate kinase and the plasma membrane protein ISG75 [26,27].

In yeast and other eukaryotes, mRNA degradation can be initiated by shortening of the poly(A) tail mediated by poly(A)-specific exoribonucleases. Poly(A) shortening is followed

by decapping of the 5' end of the mRNA and exonuclease-mediated RNA digestion in the 5' to 3' or the 3' to 5' direction (reviewed in [28]). The rate of mRNA degradation can be greatly influenced by sequence motifs in the 3' UTR, which serve as binding sites for components of the mRNA degradation machinery or proteins contributing to increased RNA stability [29,30].

In *T. brucei*, sequence elements of 16 nt and 26 nt conferring mRNA stability in PF were identified in the 3' UTR of different procyclin mRNAs, a group of highly abundant transcripts in PF [31–33]. Incorporation of these sequence motifs into the 3' UTR of a reporter gene led to its developmental regulation [32]. Regulatory elements were subsequently identified in the 3' UTRs of other RNA pol II transcribed genes in *T. brucei* and other Kinetoplastida [34–38]. In a more systematic approach to identify life cycle-specific stabilizing motifs, the 3' UTRs of seven components of the cytochrome oxidase complex were analyzed [37]. This complex is absent from BF, where glycolysis provides sufficient energy, but is produced upon differentiation to PF, where oxidative phosphorylation takes place [39]. Regulatory motifs were identified, including a motif resembling the 26-mer originally identified in the 3' UTR of the procyclin transcript [37].

What we have learned from RNA-seq

In order to thoroughly address questions related to polycistronic transcription, RNA maturation, or changes of transcript abundance during the life cycle of the parasite, it is essential to characterize these events on a global scale. Genome wide transcriptome analyses traditionally performed using microarrays have several disadvantages, including the high initial cost of genome-wide tiled microarrays for a novel organism, and the lack of precise delineation of UTRs. Hybridization-based approaches also suffer from cross-hybridization artifacts [40] and a limited dynamic range [41]. Four recently published genome-wide studies in *T. brucei* now used Illumina/Solexa high-throughput cDNA sequencing, also termed RNA-seq [42–45]. This high-throughput sequencing technology displays a low background, and its large dynamic range appears to be limited only by the depth of the sequencing [46]. Although all four studies are based on the same principle — high-throughput sequencing of a cDNA library and subsequent analysis of the sequence tags — the four groups used different *T. brucei* strains as well as varying approaches to avoid sequencing ribosomal RNAs (rRNAs) and to identify splice acceptor sites (SAS) and polyadenylation sites (PAS) (Table 1). In the following sections we compare the four studies and discuss how the common findings and differences relate to the biology of the organism.

5' UTRs and alternative splicing

For precise SAS mapping, sequence tags spanning the SAS junction and retaining a certain number of gene-specific nucleotides are retrieved (Figure 1), and the SL sequence is removed. The gene-specific sequence is then aligned to the genome. The combined sequencing data [42–45] enabled the mapping of >32,000 unique SAS of >8,900 genes (Table 2). The extent of alternative SAS usage was surprising and, while the actual number of alternatively spliced transcripts varies according to the criteria used in the different studies, it is agreed that the vast majority (>85%) of transcripts exhibit several different splice variants. Nilsson and coworkers reported more than 2,600 transcripts as alternatively spliced, where the major splice site contains less than 60% of the sequence tags [43], which is in line with the findings from Kolev *et al.* who reported that a mere 11% of transcripts contain only one SAS [44]. The average number of SAS per gene is 2.8. The three groups identified the same major SAS for more than 4,600 genes (Figure 2) [42–44]. The median 5' UTR length, including the 39 nt SL, ranges from 128 to 140 nt (Table 2; note that Nilsson *et al.* reported the UTR length per transcript rather than per gene, and both the Nilsson and Siegel studies [42,43] reported the length without the 39 nt SL). The high level of agreement

among the three studies underscores the reproducibility of the technique and validates the different approaches chosen by the three groups. The minor differences are most likely due to differences in the depth of sequencing, the thresholds set for SAS veracity, and the four trypanosome strains that were used (Table 1). Siegel *et al.* only considered SAS containing an AG dinucleotide, whereas Kolev *et al.* and Nilsson *et al.* did not constrain their search to this dinucleotide and found 2% and 6% of major SAS and 25% and 27% minor splice sites lacking the AG dinucleotide, although some of these non-canonical calls could be due to strain and allelic sequence variations (the current genome sequences consist of random assemblies of two allelic chromosomes). Two cases were identified where the major non-AG SAS contained *bona fide* AG SAS in the strain used for analysis [44]. A limited set of data derived from the TREU 927 genome reference strain identified 298 SAS for 288 genes that had not been identified using Lister 427 data [42].

Positive identification of SAS for so many genes permitted re-evaluation of the original ORF assignments. RNA-seq data from 488 to 542 genes indicated that the SAS was located downstream of the originally assigned initiator ATG, leading to a change in the N-terminal sequence of the corresponding protein (Figure 3). These changes in protein length could not have been predicted a priori. SAS data for another 248 to 588 transcripts indicated a possible N terminal extension of the corresponding ORF, based on the availability of an in-frame initiator ATG upstream of the originally annotated ATG. Failure to identify these ATGs was presumably a consequence of erroneous calls by gene-finding software. Recent mass spectrometry analyses had provided peptide sequence data for 12 of these predicted N terminal extensions, indicating that they are indeed translated [42,47]. The prediction of the true ATG for most genes remains difficult, mainly because of lack of experimental data on aspects such as minimum distance from the highly structured SL, or sequence context that might affect ATG choice.

None of the three studies performed an exhaustive systematic search for regulatory sequence motifs within UTRs, but all confirmed earlier findings that AG is the major splice acceptor dinucleotide ($\geq 94\%$) and that the SAS is preceded by a poly(Y) tract 14 to 43 nt upstream of the SAS. Depending on the study, 16 to 22% of UTRs contained at least one ATG (8.4% of the average 232-bp UTR) and could therefore potentially initiate translation upstream of the main open reading frames (Figure 3) [42,43]. Small upstream open reading frames (uORFs) can decrease mRNA translation of the main ORF efficiency in various organisms [48,49], and previous experiments suggested that uORFs may have strong regulatory effects on gene expression in *T. brucei*. Deletion of an ATG from the 5' UTR, for example, led to a 7-fold increase in expression of a luciferase reporter [50].

Altogether the global SAS analysis delivered a precise map of the splicing landscape across the genome and defined the major SAS for the vast majority of genes. It also provided new insights into regulatory mechanisms used by the parasite, like the inclusion or exclusion of uORFs, through alternative splicing, which could provide a novel mechanism to regulate protein translation in *T. brucei*. Most strikingly, a large number of genes contained alternative SAS that would potentially include or exclude N terminal targeting sequences of the corresponding proteins. One example is the Isoleucyl-tRNA synthetase (Tb927.10.9190), which is only encoded by one gene in *T. brucei*, where the corresponding enzyme activity is required in both the cytosolic and mitochondrial compartment. In this case, inclusion or exclusion of the mitochondrial targeting sequence by alternative splicing provides the mechanism for the dually targeted essential enzyme (Rettig and Schneider, unpublished) [43].

3' UTRs and polyadenylation

Polyadenylation sites are identified by mapping transcripts containing gene-specific nucleotides followed by long stretches of adenines (Figure 1); the poly(A) tail is removed and the gene-specific sequence is aligned to the genome. To map large numbers of PAS, 3' end-enriched libraries are the method of choice and long sequence tags (~76 bp) are preferred, owing to the generally low sequence complexity in the 3' UTR. Sequencing data enabled the mapping of >50,000 PAS for >8,000 genes [42,44]. The range of alternative PAS was far greater than the use of alternative SAS, which had been shown and predicted previously for several genes [4,21]. Alternative PAS have been described in *L. major* [3], but, prior to the availability of high-throughput sequencing data, reports were contradictory regarding the prevalence of alternative PAS in *T. brucei* [4,21,51,52]. Based on the generated PAS data, Siegel *et al.* calculated a median 3' UTR length of 400 nt, correlating well with the 388 nt reported from the more comprehensive study by Kolev *et al.* Comparison of the two studies revealed substantial differences in PAS. These differences are suspected to be a consequence of the extremely high heterogeneity in PAS, in combination with the relatively small number of unique PAS-bridging tags (Table 2), which is probably due to the low sequence complexity of 3' UTRs. A comparison of PAS for a highly abundant transcript like α -tubulin, for which many bridging tags are available, demonstrates that the overall patterns of PAS mapping are similar between the two studies (Figure 4). Thus, additional high-throughput sequencing would most likely decrease the discrepancies in PAS assignment between the two studies. Given the large amount of alternative splicing in the 5' UTR it is not inconceivable that the parasite uses alternative PAS on a larger scale in order to regulate transcript stability and translation in the different life cycle stages or during the cell-cycle, as has been suggested for individual genes previously (Figure 3, [53,54]).

Novel transcripts

1,114 novel transcripts that had not been annotated previously in the genome of *T. brucei* were identified in one study [44], and SAS mapping and RNA-seq data from other strains supports this finding for most of the novel transcripts [42,43]. Furthermore, it appears that 111 of the transcripts are developmentally regulated (≥ 2 -fold) between BF and PF parasites [43]. Using a cutoff of 25 amino acids, ~90% of the novel transcripts contain ORFs, 50 of which are also present in the closely related *Leishmania* and *T. cruzi*. Nineteen of the novel transcripts were also shown to be represented by peptide fragments reported in a proteomics study [44,55], indicating that a significant number of the transcripts are translated. Less than 10% of the novel transcripts, although with proper SAS and PAS, did not contain any ORF ≥ 25 amino acids, and some large transcripts did not have any coding potential. It remains to be investigated if these macro noncoding RNAs are similar to what has been found in other eukaryotic systems, where macro noncoding RNAs can be involved in regulating gene expression, for example by genomic imprinting [56].

Transcription initiation

While it has been generally assumed that divergent-SSRs serve as RNA pol II TSSs and convergent-SSRs serve as pol II TTSSs, this has never been experimentally demonstrated. To map TSSs, the unique 5' structure of newly transcribed RNA was utilized [44]. Unlike mature mRNA and processed non-coding RNA, which contain a 5'-terminal cap or 5' monophosphate, respectively, newly transcribed RNA contains a 5' tri-phosphate. Sequencing of a library containing transcripts enriched in tri-phosphate ends demonstrated for the first time that divergent-SSRs indeed serve as TSSs. Evidence for transcription initiation was found at all divergent-SSRs and at the 60 non SSR-associated sites for which a strong enrichment of H3K4me3, H4K10ac and H2AZ and H2BV had been observed previously [12,13,44], thus

confirming that TSSs and TTSSs are marked by a distinctive chromatin structure averaging ~ 8 kb in width. Generally, there were multiple TSSs spread across these regions [44]. However, despite the wealth of new information several questions remain unanswered: why are PTUs on the same DNA strand non-contiguous? Is there a limit to the processivity of RNA pol II? Or, are PTUs broken up to allow transcription of genes by RNA pol I or by RNA pol III? The latter appears to be the case at least for some tRNA genes found upstream of non SSR-associated RNA pol II TSSs [12].

Quantification of transcript levels

RNA-seq can be used to quantify transcript levels and identify genes that are regulated in a life-cycle-dependent manner. The technique has been validated by several studies and found to be highly reproducible, with very little technical variability [55]. The relative amounts of transcripts within one library can be determined by summing either the sequence tags mapping to a specific window or all transcripts mapping to a particular gene. The advantage of the latter approach is the greater number of tags but its disadvantage is that the number of tags needs to be normalized to the gene length, which requires an even distribution of sequence tags across the gene. When comparing different libraries, it is necessary to further normalize the results to a standard, the simplest method being to calculate the ratio of tags per gene to the total tags in the library. However, this might lead to over or under representation of the majority of transcripts if one or a few transcripts that are highly abundant change dramatically between libraries. Ideally, an internal standard would be spiked into each library, but this approach has not yet been used. The three studies that analyzed more than one life-cycle stage all validated their results differently, using a combination of biological replicates, correlation to qPCR, and comparison to previous studies and well-studied genes [42,43,45]. Nilsson and coworkers compared their data with a recently published tiling array [57] and found that, out of the 551 genes for which transcript levels changed ≥ 2 -fold, ~80% correlated in the direction of change between both studies. In total, significant changes in transcript abundance were reported for more than 2600 genes (32% of genes analyzed) when BF and PF parasites were compared [43], which is similar to the DGE study where, depending on the statistics used, up to 28% of the transcriptome was reported to be regulated between BF and PF (Table 1 and Table 2) [45]. Both studies used pleomorphic strains of *T. brucei* that are able to efficiently infect and complete the life cycle through the tsetse vector. The choice of cell line might be part of the reason why the number of genes regulated is substantially higher than seen in the study by Siegel *et al.*, who used a monomorphic cell line and only detected 5.6% of genes being regulated [42]. Other reasons are the choice of statistics as well as the library preparation itself. It remains to be seen what portion of transcripts are indeed changing in abundance between the life-cycle stages and — most importantly — what changes are biologically significant.

Conclusions

Alternative *cis* splicing greatly increase protein diversity in higher eukaryotes and plays important roles in cellular development, differentiation and diseases [58,59]. While *cis* splicing is virtually absent from *T. brucei*, the recently published RNA-seq data suggests that alternative *trans* splicing and polyadenylation may represent novel avenues of gene regulation generating a diverse pool of transcripts from a single gene, differing in stability, translation efficiency and possibility even in coding sequence (Figure 3) [42–45]. Furthermore, the life cycle-specific differences in choice of SAS [43], suggest that alternative splicing is a regulated mechanism of which the corresponding regulatory factors remain elusive. In this regard, it would be interesting to determine the SAS and PAS on a single cell level or a synchronized culture. As cost for high-throughout sequencing continues to drop, it will probably become routine to follow up a gene knock-out by RNA-seq to

assess its affect on transcript levels. In *T. brucei*, such analyses should also include a comparison of SAS and PAS to a wild type data set. Maybe this will lead to the identification of factors involved regulating alternative RNA processing. Analysis of the high-throughput sequencing data also allowed the identification of novel transcripts, and led to the re-annotation of a large number of translation initiation sites. For the first time, we now have a precisely defined transcriptome from different life-cycle stages of the parasite allowing us in the future to better study the questions of regulatory elements in 5' and 3' UTRs.

Glossary

Illumina/Solexa sequencing	high-throughput sequencing, based on a sequencing-by-synthesis chemistry developed by Solexa, Inc. Solexa was acquired by Illumina, Inc., and the technology currently yields about 600 million 100-bp sequences (60 billion base reads) per sequencing run
Long slender forms	fast-growing forms living in the bloodstream and interstitial spaces of the mammalian host. Long slender forms cannot efficiently infect the insect vector (<i>Glossina</i> , also referred to as tsetse). Often these parasites are simply referred to as bloodstream forms (BF)
Novel transcripts	previously unidentified RNA transcripts containing a spliced leader and a poly(A) tail. They comprise protein-coding and non-protein-coding RNAs
Polyadenylation sites (PAS)	mark the 3' end of the 3' UTR of each gene. A poly(A) tail is added to the 3'-most nucleotide of each gene, which is referred to as the PAS
Procyclic forms (PF)	the initial form of the parasite that proliferates in the glucose-poor midgut of the tsetse. Differentiation of short stumpy forms into procyclic forms is accompanied by activation of the mitochondrion and associated major metabolic and cellular changes
Short stumpy forms	non-proliferating forms arising from long slender forms in the mammalian host. Short-stumpy forms are pre-adapted for uptake by the tsetse, where they differentiate into procyclic forms
Splice acceptor sites (SAS)	mark the 5' end of the 5' untranslated region (UTR) of each gene. The 39 nt spliced leader is trans-spliced to the 5' nucleotide of each gene, which is referred to as the SAS. The SAS predominantly follows the dinucleotide AG
Spliced leader (SL)	a 39-nucleotide (nt) sequence that is <i>trans</i> -spliced onto the 5' end of all mRNA transcripts and onto the 5' of some non-protein-coding RNA
Strand-switch regions (SSRs)	regions between polycistronic transcription units that are located on opposite strands of the DNA. Strand-switch regions contain RNA polymerase II transcription start or transcription termination sites, depending upon the divergent or convergent orientation of the adjacent transcription units

<i>Trypanosoma brucei</i> <i>brucei</i>	a protozoan parasite and the causative agent of the wasting disease nagana in cattle in sub Saharan Africa; this organism is not infective to humans, by definition. Most laboratory strains, including the TREU 927 and Lister 427, whose genomes have been sequenced, belong to the subspecies <i>T. brucei brucei</i>
<i>T. b. gambiense</i>	a human-infective subspecies of <i>T. brucei</i> , most commonly found in central and western Africa, which causes slow-onset chronic trypanosomiasis. This subspecies accounts for more than 95% of sleeping sickness cases
<i>T. b. rhodesiense</i>	a virulent human-infective subspecies of <i>T. brucei</i> , most commonly found in southern and eastern Africa, which causes acute human African trypanosomiasis. Adding one specific gene (<i>SRA</i>) to <i>T. b. brucei</i> is sufficient to convert it to <i>T. b. rhodesiense</i>

References

1. Johnson PJ, et al. Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell*. 1987; 51:273–281. [PubMed: 3664637]
2. Berriman M, et al. The genome of the African trypanosome *Trypanosoma brucei*. *Science*. 2005; 309:416–422. [PubMed: 16020726]
3. LeBowitz JH, et al. Coupling of poly(A) site selection and *trans*-splicing in *Leishmania*. *Genes & Dev*. 1993; 7:996–1007. [PubMed: 8504937]
4. Matthews KR, et al. A common pyrimidine-rich motif governs *trans*-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes & Dev*. 1994; 8:491–501. [PubMed: 7907303]
5. Spieth J, et al. Operons in *C. elegans*: polycistronic mRNA precursors are processed by *trans*-splicing of SL2 to downstream coding regions. *Cell*. 1993; 73:521–532. [PubMed: 8098272]
6. Brogna S, Ashburner M. The Adh-related gene of *Drosophila melanogaster* is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. *EMBO J*. 1997; 16:2023–2031. [PubMed: 9155028]
7. Lee SJ. Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc Natl Acad Sci USA*. 1991; 88:4250–4254. [PubMed: 2034669]
8. Das A, et al. Trypanosomal TBP functions with the multisubunit transcription factor tSNAP to direct spliced-leader RNA gene expression. *Mol Cell Biol*. 2005; 25:7314–7322. [PubMed: 16055739]
9. Schimanski B, et al. Characterization of a multisubunit transcription factor complex essential for spliced-leader RNA gene transcription in *Trypanosoma brucei*. *Mol Cell Biol*. 2005; 25:7303–7313. [PubMed: 16055738]
10. Martinez-Calvillo S, et al. Transcription of *Leishmania* major Friedlin chromosome 1 initiates in both directions within a single region. *Mol Cell*. 2003; 11:1291–1299. [PubMed: 12769852]
11. El-Sayed NM, et al. Comparative genomics of trypanosomatid parasitic protozoa. *Science*. 2005; 309:404–449. [PubMed: 16020724]
12. Siegel TN, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes & Dev*. 2009; 23:1063–1076. [PubMed: 19369410]
13. Wright JR, et al. Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. *Mol Biochem Parasitol*. 2010; 136:434–450.
14. Perry KL, et al. Trypanosome mRNAs have unusual “cap 4” structures acquired by addition of a spliced leader. *Proc Natl Acad Sci USA*. 1987; 84:8190–8194. [PubMed: 3120186]
15. Campbell DA, et al. Apparent discontinuous transcription of *Trypanosoma brucei* variant surface antigen genes. *Nature*. 1984; 311:350–355. [PubMed: 6090933]

16. Milhausen M, et al. Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5' sequences of trypanosomatid mRNAs? *Cell*. 1984; 38:721–729. [PubMed: 6091897]
17. Huang J, Van der Ploeg LH. Requirement of a polypyrimidine tract for *trans*-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site. *EMBO J*. 1991; 10:3877–3885. [PubMed: 1935907]
18. Lopez-Estrano C, et al. Exonic sequences in the 5' untranslated region of alpha-tubulin mRNA modulate *trans* splicing in *Trypanosoma brucei*. *Mol Cell Biol*. 1998; 18:4620–4628. [PubMed: 9671472]
19. Patzelt E, et al. Mapping of branch sites in *trans*-spliced pre-mRNAs of *Trypanosoma brucei*. *Mol Cell Biol*. 1989; 9:4291–4297. [PubMed: 2479824]
20. Liang XH, et al. *trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell*. 2003; 2:830–840. [PubMed: 14555465]
21. Benz C, et al. Messenger RNA processing sites in *Trypanosoma brucei*. *Mol Biochem Parasitol*. 2005; 143:125–134. [PubMed: 15993496]
22. Wilusz CJ, et al. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol*. 2001; 2:237–246. [PubMed: 11283721]
23. Schwede A, et al. A role for Caf1 in mRNA deadenylation and decay in trypanosomes and human cells. *Nucleic Acids Res*. 2008; 36:3374–3388. [PubMed: 18442996]
24. Clayton CE. Life without transcriptional control? From fly to man and back again. *EMBO J*. 2002; 21:1881–1888. [PubMed: 11953307]
25. Haile S, Papadopoulou B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Curr Opin Microbiol*. 2007; 10:569–577. [PubMed: 18177626]
26. Haanstra JR, et al. Control and regulation of gene expression: quantitative analysis of the expression of phosphoglycerate kinase in bloodstream form *Trypanosoma brucei*. *J Biol Chem*. 2008; 283:2495–2507. [PubMed: 17991737]
27. Kramer S, et al. The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes. *J Cell Sci*. 2010; 123:699–711. [PubMed: 20124414]
28. Wilusz CJ, Wilusz J. Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet*. 2004; 20:491–497. [PubMed: 15363903]
29. Zubiaga AM, et al. The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation. *Mol Cell Biol*. 1995; 15:2219–2230. [PubMed: 7891716]
30. Yang E, et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res*. 2003; 13:1863–1872. [PubMed: 12902380]
31. Hehl A, et al. A conserved stem-loop structure in the 3' untranslated region of procyclin mRNAs regulates expression in *Trypanosoma brucei*. *Proc Natl Acad Sci USA*. 1994; 91:370–374. [PubMed: 8278396]
32. Hotz HR, et al. Mechanisms of developmental regulation in *Trypanosoma brucei*: a polypyrimidine tract in the 3'-untranslated region of a surface protein mRNA affects RNA abundance and translation. *Nucleic Acids Res*. 1997; 25:3017–3026. [PubMed: 9224601]
33. Furger A, et al. Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of *Trypanosoma brucei* by modulating RNA stability and translation. *Mol Cell Biol*. 1997; 17:4372–4380. [PubMed: 9234695]
34. Hotz HR, et al. Role of 3'-untranslated regions in the regulation of hexose transporter mRNAs in *Trypanosoma brucei*. *Mol Biochem Parasitol*. 1995; 75:1–14. [PubMed: 8720170]
35. Di Noia JM, et al. AU-rich elements in the 3'-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. *J Biol Chem*. 2000; 275:10218–10227. [PubMed: 10744707]
36. Boucher N, et al. A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3'-untranslated region element. *J Biol Chem*. 2002; 277:19511–19520. [PubMed: 11912202]

37. Mayho M, et al. Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in *Trypanosoma brucei*: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. *Nucleic Acids Res.* 2006; 34:5312–5324. [PubMed: 17012283]
38. Bringaud F, et al. Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*. *PLoS Pathog.* 2007; 3:1291–1307. [PubMed: 17907803]
39. Priest JW, Hajduk SL. Developmental regulation of mitochondrial biogenesis in *Trypanosoma brucei*. *J Bioenerg Biomembr.* 1994; 26:179–191. [PubMed: 8056785]
40. Casneuf T, et al. In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics.* 2007; 8:461. [PubMed: 18039370]
41. Nagalakshmi U, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008; 320:1344–1349. [PubMed: 18451266]
42. Siegel TN, et al. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* 2010; 38:4946–4957. [PubMed: 20385579]
43. Nilsson D, et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. *PLoS Pathog.* 2010; 6:e1001037. [PubMed: 20700444]
44. Kolev NG, et al. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog.* 2010; 6:e1001090. [PubMed: 20838601]
45. Veitch NJ, et al. Digital gene expression analysis of two life cycle stages of the human-infective parasite, *Trypanosoma brucei gambiense* reveals differentially expressed clusters of co-regulated genes. *BMC Genomics.* 2010; 11:124. [PubMed: 20175885]
46. Wilhelm BT, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature.* 2008; 453:1239–1243. [PubMed: 18488015]
47. Zhang X, et al. The *Trypanosoma brucei* MitoCarta and its regulation and splicing pattern during development. *Nucleic Acids Res.* 2010; 38:7378–7387. [PubMed: 20660476]
48. Meijer HA, Thomas AA. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem J.* 2002; 367:1–11. [PubMed: 12117416]
49. Hood HM, et al. Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. *Annu Rev Microbiol.* 2009; 63:385–409. [PubMed: 19514854]
50. Siegel TN, et al. Systematic study of sequence motifs for RNA *trans* splicing in *Trypanosoma brucei*. *Mol Cell Biol.* 2005; 25:9586–9594. [PubMed: 16227607]
51. Tschudi C, Ullu E. Polygene transcripts are precursors to calmodulin mRNAs in trypanosomes. *EMBO J.* 1988; 7:455–463. [PubMed: 3366120]
52. Hug M, et al. Hierarchies of RNA-processing signals in a trypanosome surface antigen mRNA precursor. *Mol Cell Biol.* 1994; 14:7428–7435. [PubMed: 7935457]
53. Clement SL, Koslowsky DJ. Unusual organization of a developmentally regulated mitochondrial RNA polymerase (TBMTRNAP) gene in *Trypanosoma brucei*. *Gene.* 2001; 272:209–218. [PubMed: 11470527]
54. Erondy NE, Donelson JE. Differential expression of two mRNAs from a single gene encoding an HMG1-like DNA binding protein of African trypanosomes. *Mol Biochem Parasitol.* 1992; 51:111–118. [PubMed: 1565127]
55. Marioni JC, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008; 18:1509–1517. [PubMed: 18550803]
56. Koerner MV, et al. The function of non-coding RNAs in genomic imprinting. *Development.* 2009; 136:1771–1783. [PubMed: 19429783]
57. Jensen BC, et al. Widespread variation in transcript abundance within and across developmental stages of *Trypanosoma brucei*. *BMC Genomics.* 2009; 10:482. [PubMed: 19840382]
58. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet.* 2007; 8:749–761. [PubMed: 17726481]
59. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature.* 2010; 463:457–463. [PubMed: 20110989]

60. Langmead B, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009; 10:R25. [PubMed: 19261174]
61. Li H, et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008; 18:1851–1858. [PubMed: 18714091]
62. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002; 12:656–664. [PubMed: 11932250]
63. Romaniuk E, et al. The effect of acceptor oligoribonucleotide sequence on the T4 RNA ligase reaction. *Eur J Biochem.* 1982; 125:639–643. [PubMed: 7117259]

Box 1

Sequencing strategy follows biological question

The four studies show that there are several ways to analyze the transcriptome of trypanosomes, but there seem to be distinct advantages in one particular strategy depending upon the question being asked. Transcript level changes between life cycle stages can be addressed using RNAseq, SLT or DGE. RNAseq and — to a lesser degree — DGE offer the advantage that reads cover the entire transcript, while SLT reads only cover the 5' ends of mRNAs. Sequence tag distribution across the entire transcript can be useful when transcript levels need to be compared among genes similar in sequence, e.g. canonical histones compared to histone variants. The advantage of SLT, on the other hand, is that it strongly discriminates against ribosomal RNA contamination and thus produces more mRNA sequence tags. While DGE is an established technology, it clearly suffers from the requirement of restriction sites like *Nla* III or *Dpn* II that are needed for anchoring the sequence tags: if those are absent the transcripts are missed. Furthermore the number of bases sequenced using DGE is very small (17 bp) compared to the standard capabilities of the current Illumina sequencing technology (100 bp), very often leading to sequence tags that cannot be aligned unambiguously to the genome. For the discovery and analysis of PAS, a poly (A) selection scheme followed by first strand synthesis using oligo dT primers significantly enriches 3' end reads, but ribosomal RNA contamination remains a problem, which can be overcome through an increase in sequencing depth [42,44]. The discovery and quantification of SAS is easily achieved using the SLT strategy, where complete coverage of the genome is reached with less than two million sequence tags, allowing for multiplexing of a large number of experiments in one sequencing channel [43]. To analyze RNA pol II TSSs, processed RNAs (miRNA, rRNA, etc.), containing a 5' monophosphate were removed using a 5' monophosphate-dependent exonuclease, followed by 5' RNA linker ligation, which excluded RNAs containing a cap structure [44]. Finally, RNAseq, and DGE as described in the studies here rely on linker ligation to double stranded cDNA and thus lose the strand information of the corresponding RNA. This problem can be easily overcome using RNA ligation strategies with T4 RNA ligase, which, however, has been shown to introduce a ligation bias depending on the sequence context of the acceptor molecule [63].

Figure 1.

Polycistronic transcription and processing of mRNAs in trypanosomes. *Trans* splicing of spliced leader RNA to the 5' end of the 5' UTR of protein coding genes leads to the dissection of polycistronic transcription units. Mature mRNA contains a 5' conserved spliced leader sequence and a 3' poly(A) tail. Sequence tags 'bridging' the spliced leader sequence and the 5' end of a gene or the poly(A) tail and the 3' end of a gene can be used to map SAS and PAS, respectively.

Figure 2.

Comparison of SAS. **(a)** Overview of 85 kb from chromosome I containing 42 open reading frames (ORF; yellow). SAS from the studies of Kolev (K, red) [44], Nilsson (N, black) [43], Siegel (S, green) [42] and their coworkers are depicted as bars below the ORFs. **(b)** Venn diagram indicating the overlap in the number of identical major SAS found in the *T. brucei* genome by the three studies.

Figure 3.

Biological relevance of alternative RNA splicing. **(a)** Exclusion of a small uORFs by alternative SAS usage may increase translation efficiency [48,49] **(b)** Usage of an alternative downstream SAS can shorten the coding sequence of a gene and may lead to the exclusion of a signal peptide, thus affecting the cellular targeting of the translated protein. **(c)** Usage of alternative upstream PAS can lead to the exclusion of AU-rich elements and may thus lead to increased RNA stability of the corresponding transcript [29,35].

Figure 4. Comparison of α -tubulin (Tb927.1.2360) PAS. Depicted are the PAS identified by Kolev *et al.* and Siegel *et al.* [42,44] Nucleotides labeled in red indicate PAS. Nucleotide position is shown in relation to the upstream stop codon.

Table 1Comparison of four recent *T. brucei* transcriptome studies

	Kolev <i>et al.</i> [44]	Nilsson <i>et al.</i> [43]	Siegel <i>et al.</i> [42]	Veitch <i>et al.</i> [45]
Species/strain	<i>T. b. rhodesiense</i> YTat1.1 (monomorphic ^a)	<i>T. b. brucei</i> Lister 427 clone 221 (monomorphic) and AnTat 1.1 (pleomorphic ^b)	<i>T. b. brucei</i> Lister 427 clone 221 (monomorphic) and TREU927 PF	<i>T. b. gambiense</i> STIB386 (pleomorphic)
Life stages	PF ^c	LS ^d , SS ^e , PF ^c	LS, PF ^c	LS, PF ^c
rRNA depletion	Poly(A), 5'-PPP ^f , SL ^g	Poly(A)	Poly(A)	Poly(A)
Library construction	3 library types; 5', 3' enriched and 5'-PPP	2 library types; SLT ^h (5' enriched) and RNAseq ⁱ	2 library types; RNAseq and 3' enriched	1 library type; DGE ^j
# of reads/reads aligned (%)	5' end libraries 33,338,202 (95%); 3' end libraries 30,860,548 (82%)	SLT libraries: 4,588,178 (93%) RNAseq: 24,490,701 (81%)	RNAseq libraries: 25,334,935 (72%) 3' enriched libraries: 6,402,419	DGE: 11,524,598 (28%)
Mapping algorithms	Bowtie ^k ; 2 bp mismatches in first 28 bp; whole genome used for mapping	Bowtie and MAQ ^k ; 2 bp mismatches in first 28 bp; whole genome used for mapping	BLAT ^k and Bowtie; 2 bp mismatch for 32/36 bp reads, 3 mismatches for 76 bp reads; whole genome used for mapping	MAQ; 2 bp mismatches in 21 bp read; only ORFs used for mapping

^a Monomorphic bloodstream forms are, by definition, unable to differentiate normally, via the stumpy form, and appear generally to be unable to complete the life cycle within the tsetse.

^b Pleomorphic cells exhibit a complete the life cycle in the tsetse and the mammalian host.

^c Procyclic (PF) form of the parasite

^d Long slender (LS) bloodstream form of the parasite

^e Short stumpy (SS) bloodstream form of the parasite

^f Triphosphate at the 5' end of RNA (5'-PPP), a hallmark of preprocessed mRNA when compared to rRNA and other noncoding RNAs that contain a monophosphate at the 5' end. Monophosphate containing RNAs are depleted using a monophosphate dependent Exonuclease.

^g Spliced leader (SL) sequence is used to deplete the rRNA

^h Spliced leader trapping (SLT) is a method to selectively sequence the sequence downstream of the spliced leader acceptor site (Nilsson et al.)

ⁱ Digital gene expression (DGE) is a gene expression profiling approach relying on the restriction enzyme NlaIII

^j RNAseq. is a whole transcriptome shotgun sequencing approach

^k Mapping algorithm and software packages Bowtie [60], MAQ [61], BLAT [62]

Abbreviations: ORF, open reading frame; bp, base pairs; Poly(A) polyadenylation; PAS, polyadenylation site; TSS, transcription start site.

Table 2Comparison of SAS^a and PAS^b and UTR^c length between studies

Gene features	Kolev <i>et al.</i> [44]	Nilsson <i>et al.</i> [43]	Siegel <i>et al.</i> [42]
SAS	32095 ^d	29406 ^e	10856 ^f
Genes with SAS	8935 ^d	8277 ^e	6959 ^f
PAS	52078 ^g	N/A ^h	16863 ⁱ
Genes with PAS	8038 ^g	N/A	5948 ⁱ
Median 5' UTR length/gene (based on major SAS, including 39 nt of spliced leader)	130	140 ^j	128
Median 3' UTR length/gene (based on major PAS)	388	N/A	400
SAS downstream of ATG ^k	532	542 ^j	488

^a SAS = splice acceptor site^b PAS = polyadenylation site^c UTR = untranslated region^d splice acceptor sites represented by ≥1 tag from Table S4 [44]^e splice acceptor sites represented by ≥1 tag from bloodstream form library^f splice acceptor sites represented by ≥2 tags (only unique hits)^g polyadenylation sites represented by ≥1 tag from Table S5 [44]^h N/A = not applicableⁱ polyadenylation sites represented by ≥2 tags (only unique hits)^j from the bloodstream form library^k ATG = predicted start codon as annotated in TriTrypDB 2.5