Functional characterization of the *Trypanosoma brucei* polyadenylation complex

Dissertation

Vorgelegt von

**Henrik Koch**
(M.Sc. Biologie)

zur Erlangung des akademischen Grades
*doctor rerum naturalium*
(Dr. rer. nat.)

Eingereicht am Fachbereich Biologie und Chemie
der Justus-Liebig-Universität Gießen

Gießen, Juli 2016

Dekan: Prof. Dr. Volker Wissemann
Institut für Botanik (FB08)
Justus-Liebig-Universität Gießen

1. Gutachter: Prof. Dr. Albrecht Bindereif
Institut für Biochemie (FB08)
Justus-Liebig-Universität Gießen

2. Gutachter: Prof. Dr. Roland K. Hartmann
Institut für Pharmazeutische Chemie (FB16)
Philipps-Universität Marburg
# Table of contents

1 Summary .......................................................................................................................... 7

2 Zusammenfassung ............................................................................................................. 8

3 Introduction .........................................................................................................................
   3.1 Gene expression in eukaryotes ..................................................................................... 10
   3.2 Pre-mRNA splicing in eukaryotes .............................................................................. 11
   3.3 The human polyadenylation complex and its function ............................................ 14
   3.4 Alternative polyadenylation (APA) and its biological importance ......................... 19
   3.5 Regulation of polyadenylation .................................................................................... 21
   3.6 The African trypanosomes: *Trypanosoma brucei* .................................................... 22
   3.7 Expression of protein-coding genes in trypanosomes .............................................. 23
   3.8 The trypanosomatid PTB/hnRNP I homologs DRBD3 and DRBD4 ....................... 29
   3.9 Aims of this work ....................................................................................................... 30

4 Material and Methods ..................................................................................................... 31
   4.1 Material ......................................................................................................................... 31
      4.1.1 Chemicals and reagents ......................................................................................... 31
      4.1.2 Commercial Kits .................................................................................................... 34
      4.1.3 Enzymes and enzyme inhibitors .......................................................................... 34
      4.1.4 Nucleotides .......................................................................................................... 35
      4.1.5 Markers ................................................................................................................ 35
      4.1.6 Antibodies ............................................................................................................ 35
      4.1.7 Plasmids ............................................................................................................... 36
      4.1.8 Bacterial and eukaryotic cells .............................................................................. 36
      4.1.9 *T. brucei* SDM-79 medium ................................................................................ 37
      4.1.10 Laboratory equipment ......................................................................................... 37
      4.1.11 DNA and RNA oligonucleotides (in 5´ to 3´direction) ....................................... 38
   4.2 General techniques ..................................................................................................... 44
      4.2.1 Guideline ................................................................................................................. 44
      4.2.2 Working with RNA ............................................................................................... 44
      4.2.3 Isolation of nucleic acids from *T. brucei* cells .................................................. 44
      4.2.4 Purification of nucleic acids .................................................................................. 44
      4.2.5 Precipitation of nucleic acids .............................................................................. 44
4.8.3 *In vitro* transcription ................................................................. 59
4.8.4 *In vitro* polyadenylation assay .................................................. 60
4.8.5 Tandem affinity purification (TAP) and mass spectrometry ............... 60
4.8.6 *In vivo* polyadenylation assay ................................................... 61
4.8.7 Individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) ................................................................. 62
4.8.8 Systematic evolution of ligands by exponential enrichment-sequencing (SELEX-seq) ................................................................. 63
4.8.9 GST pulldown assay ..................................................................... 64
4.8.10 RNA-seq and identification of polyadenylation sites ....................... 64
4.8.11 Correlation between DRBD4 iCLIP binding and polyadenylation site choice ..................................................................................... 65
4.8.12 Flow cytometry .......................................................................... 65
4.9 Constructs ......................................................................................... 66
  4.9.1 Generation of pLew100 RNAi constructs ........................................ 66
  4.9.2 Generation of PTP-tag fusion protein constructs ........................... 66
  4.9.3 Generation of protein overexpression constructs .......................... 67
5 Results ............................................................................................. 68
  5.1 The trypanosomatid polyadenylation complex .................................... 68
    5.1.1 The trypanosomatid poly(A) polymerases .................................... 68
    5.1.2 Only the putative poly(A) polymerase Tb927.7.3780 shows *in vitro* activity .... 70
    5.1.3 *In vivo* activity of the putative poly(A) polymerases .................... 71
    5.1.4 Identification of the constituents of the polyadenylation complex .......... 73
    5.1.5 Characterization of the CPSF30 RNA-binding properties .................. 79
    5.1.6 Polyadenylation factors are required for both polyadenylation and *trans* splicing . 81
  5.2 DRBD4-mediated regulation of mRNA polyadenylation in *Trypanosoma brucei*. 88
    5.2.1 DRBD4 binds *in vivo* preferentially to gene UTRs ......................... 88
    5.2.2 DRBD4 binds purine-rich sequences containing AUGA elements *in vitro* .... 91
    5.2.3 DRBD4 depletion affects polyadenylation site usage ......................... 96
    5.2.4 DRBD4-regulated poly(A) site choice: integrating iCLIP and RNA-seq data .... 97
    5.2.5 DRBD4 is associated with polyadenylation complex constituents .......... 102
6 Discussion ......................................................................................... 104
  6.1 The poly(A) polymerase of *T. brucei* ............................................ 104
  6.2 Importance of *cis* splicing in trypanosomes .................................... 105
1. Summary

All protein-coding genes in the protozoan parasite *Trypanosoma brucei* are arranged in long clusters and are transcribed into polycistronic precursor RNAs. These pre-mRNAs require further processing by coupled *trans* splicing and polyadenylation to generate mature mRNAs. Although studies in the last decade have identified numerous spliceosomal components, we still know very little about the components, mechanisms, and dynamics of the 3´ end-processing machinery in trypanosomes. Moreover, most genes in trypanosomes contain one to three *trans* splice and even more dispersed polyadenylation sites, indicating that the regulation of both processes provides another, still not very well explored level of post-transcriptional gene regulation in trypanosomes: In terms of polyadenylation, factors regulating the polyadenylation efficiency of a primary transcript with a single poly(A) site directly affect protein expression, since unprocessed transcripts are degraded or not exported to the cytoplasm. In addition, multiple polyadenylation sites allow the generation of different transcript isoforms of a single gene by alternative polyadenylation.

To characterize the catalytic core of the polyadenylation complex in *T. brucei*, we first identified the poly(A) polymerase [Tb927.7.3780] as the major functional, nuclear-localized enzyme in trypanosomes. In contrast, another poly(A) polymerase, encoded by an intron-containing gene [Tb927.3.3160], localizes mainly in the cytoplasm and appears not to be functional in general 3´ end processing of mRNAs. Based on tandem affinity purification with tagged CPSF160 and mass spectrometry, we identified ten associated components of the trypanosome polyadenylation complex, including homologues to all four CPSF subunits, Fip1, CstF50/64, and Symplekin, as well as two hypothetical proteins. RNAi-mediated knockdown revealed that most of these factors are essential for growth and required for both *in vivo* polyadenylation and *trans* splicing, arguing for a general coupling of these two mRNA-processing reactions.

By combining genome-wide analysis of expression (RNA-seq) and *in vivo* RNA binding (iCLIP), we identified for the first time a *trans*-acting RNA-binding protein, the trypanosomatid polypyrimidine tract binding protein (PTB/hnRNP I) homolog DRBD4, as a regulator of polyadenylation. Based on SELEX-seq and iCLIP, we delineated purine-rich sequences containing AUGA elements as DRBD4 RNA-binding motif and mapped *in vivo* binding sites mainly in untranslated regions (UTRs). Integrating RNA-seq and iCLIP datasets revealed that DRBD4 binds upstream of poly(A) sites and modulates both their activation and repression, thereby affecting general transcript and isoform expression levels.
2 Zusammenfassung


Im Rahmen der Charakterisierung des *T. brucei* Polyadenylierungskomplexes haben wir zunächst die funktionelle, Zellkern-lokalisierte Poly(A) Polymerase [Tb927.7.3780] identifiziert. Die zweite Poly(A) Polymerase [Tb927.3.3160], die von einem Gen mit einem Intron kodiert wird, ist hingegen überwiegend cytoplasmatisch lokalisiert und nicht an der Polyadenylierung von mRNAs beteiligt. Mittels *tandem affinity purification* von Epitop-markiertem CPSF160 und einer anschließenden massenspektrometrischen Analyse konnten wir zehn Proteinfaktoren des Polyadenylierungskomplexes in *T. brucei* identifizieren: Homologe der vier CPSF Untereinheiten, Fip1, CstF50/64, Symplekin sowie zwei hypothetische Proteine. Des Weiteren konnten wir anhand eines RNAi-induzierten knockdowns der identifizierten Faktoren zeigen, dass die meisten sowohl für das Zellwachstum als auch für die *in vivo* Polyadenylierungs- und *trans*-Spleißreaktion essentiell sind.

Durch die Kombination von genomweiten Genexpressions- (RNA-seq) und *in vivo* RNA-Bindungsdaten (iCLIP) konnten wir das RNA-Bindeprotein DRBD4, ein Homolog des humanen *polypyrimidine tract binding proteins* (PTB/hnRNP l), als ersten *trans*-aktiven Polyadenylierungsregulator in Trypanosomen identifizieren. Mittels SELEX-seq wurden Purin-reiche Sequenzen, die AUGA Tetramere enthalten, als Bindemotiv charakterisiert, und mittels iCLIP wurde die überwiegende Zahl der DRBD4 Bindungsstellen in untranslatierten Bereichen (UTR) von mRNAs lokalisiert. Anhand der kombinatorischen Auswertung der RNA-seq und
iCLIP-Datensätze konnten wir zeigen, dass DRBD4 *upstream* von Polyadenylierungsstellen bindet und sowohl für die Aktivierung als auch die Repression essentiell ist, wodurch generelle Transkript- und Isoformexpressionslevel reguliert werden.
3  Introduction

3.1  Gene expression in eukaryotes
The genetic information of eukaryotes is mostly stored as DNA in the nucleus. For the expression of genes, the DNA is first transcribed into RNA and in case of protein-coding genes, the mRNA is exported from the nucleus and translated into protein by the ribosome. Three major DNA-dependent RNA polymerases are involved in transcription of RNA: RNA polymerase I and III are responsible for the transcription of ribosomal RNAs (rRNA) and transfer RNAs (tRNA), whereas RNA polymerase II (Pol II) transcribes messenger RNAs (mRNA) of protein-coding genes. Moreover, Pol II transcribes non-coding RNAs and most small nuclear RNAs (snRNAs). The carboxy-terminal domain (CTD) of the largest Pol II subunit is involved in regulation of transcription by differential phosphorylation and couples transcription to important RNA processing reactions.

The first transcription-coupled processing step is capping of the 5′ end of the nascent mRNA with 7-methylguanosine, forming the m^7^GpppN cap structure (Ghosh & Lima, 2010). This structure is required for mRNA stability, export, splicing and efficient translation initiation. After capping, the mRNA precursor undergoes cis splicing (see 3.2) since the genetic information of protein-coding genes is stored discontinuously in higher eukaryotes: The open reading frame (ORF) of a protein is encoded by exons (expressed regions), which are interspaced by non-coding introns (intervening regions) that are removed during co-transcriptional cis splicing. The splicing reaction is catalyzed by a macromolecular complex, the spliceosome, which assembles while transcription still takes place. During complex assembly, conserved sequences of the exon/intron borders, the splice sites, are recognized by the spliceosome and the exons are joined together, while the intron is removed. Finally, Pol II reads through the polyadenylation signal required for 3′ end maturation of the transcription (see 3.3 and 3.4): The polyadenylation complex assembles on the nascent transcript, which is then cleaved by CPSF73 and polyadenylated by the poly(A) polymerase (PAP). The added poly(A) tail has an approximate length of 250 nucleotides (nts) and is essential for RNA stability, export and translation. Afterwards transcription is terminated via the degradation of the 3′ cleavage product by exonucleases Xrn2 (Kuehner et al., 2011). During all processing steps, a multitude of proteins bind to the RNA forming an mRNA-protein complex important for further downstream processes: RNA export (Kohler & Hurt, 2007), RNA surveillance and quality control (van Hoof & Wagner, 2011, Schweingruber et al., 2013), translation into protein (Merrick, 2010), and degradation of the RNA (Houseley & Tollervey, 2009).
3.2 Pre-mRNA splicing in eukaryotes
The removal of intronic, non-coding sequences and the fusion of the coding exons is an essential step in pre-mRNA maturation (Sharp, 1994). Besides the above mentioned cis splicing (see 3.1), exons derived from two transcripts can be joined by trans splicing (Hastings, 2005). Trans splicing in trypanosomes (see 3.7) requires the spliced-leader (SL) donor RNA, from which a miniexon is transferred to the splice acceptor site (3’ splice site) of the pre-mRNA transcribed as a long polycistronic precursor.

The chemical basis of both cis and trans splicing is a two-step transesterification reaction (Figure 3.1): In the first reaction, the 2’ hydroxyl group of the branch point adenosine carries out a nucleophilic attack on the 5’ splice site phosphate, forming the lariat intron (cis splicing) or the Y structure (trans splicing) both characterized by the 2’-5’ phosphodiester bond at the branch point. In the second step, the phosphate group at the 3’ splice site is nucleophilically attacked by the 3’ hydroxyl group of the released exon/miniexon, resulting in the ligation of the exons and the release of the intron. Finally, the intron lariat or Y structure is degraded by the debranching enzyme and exonucleases.

Figure 3.1 Schematic models of cis and trans splicing.
The model splice substrates consist of two exons (cis splicing; left) or one miniexon and a second exon (trans splicing; right) depicted as boxes. The introns are shown as solid lines. Both the 5’ and 3’ splice sites (5’ ss and 3’ ss) including the phosphate group (p), the branch point (BP) adenosine (A) and the polypyrimidine tract (PPT) are indicated. The splicing reaction is based on a two-step transesterification reaction: First, the 2’ hydroxyl group of the branch point adenosine carries out a nucleophilic attack on the 5’ splice site phosphate, forming the lariat intron (cis splicing) or the Y structure (trans splicing) each characterized by the 2’-5’ bond at the branch point. In the second step, the phosphate group at the 3’ splice site is nucleophilically attacked by the 3’ hydroxyl group of the released exon/miniexon, resulting in the ligation of the exons and the release of the intron.
For spliceosome assembly and efficient splicing conserved, *cis*-acting sequence motifs in the intron are required (Figure 3.2): The first two and the last two nucleotides in the intron at the 5´ donor (GT) and 3´ acceptor splice site (AG) define the exon/intron borders. The branch point adenosine carries out the nucleophilic attack on the 5´ splice site phosphate and the polypyrimidine tract serves as another essential sequence element (Patel & Steitz, 2003).

**Figure 3.2 Consensus sequences of mammalian introns.**
Consensus sequences of the 5´ and 3´ splice sites (5´ ss and 3´ ss), the branch point (BP) adenosine (A) and the polypyrimidine tract (PPT). The letter heights at each position represent the frequency of occurrence of the corresponding nucleotides at that position. The positions that are involved in intron recognition are shown in black; other positions are shown in blue. Frequencies were derived from a set of mammalian introns. The exons 1 and 2 are shown as boxes and the intron as solid line. From Patel & Steitz, 2003; modified.

The splicing reaction is catalyzed by the macromolecular spliceosome, which is composed of five (U1, U2, U4, U5 and U6) small nuclear ribonucleoprotein particles (snRNPs) as well as a large set of proteins (Will & Lührmann, 2011, Matera & Wang, 2014). Each snRNP consists of a specific small, U-rich RNA, each containing the conserved Sm site (5´ AUUUGUG 3´), which is bound by seven common Sm-proteins (B/B′, D3, D2, D1, E, F, and G) forming a heptameric ring. In contrast, the U6 snRNA is bound by seven related LSm-proteins (Lsm2-8) that assemble to an analogous heptameric ring. In addition, each snRNP carries several specific proteins. It was recently shown that the U2, U5 and U6 snRNPs form the catalytic center making the spliceosome indeed a ribozyme (Hang *et al.*, 2015).

Spliceosome assembly is a highly ordered process involving the five snRNPs, the pre-mRNA and numerous splicing factors (Will & Lührmann, 2011). Moreover, the spliceosome undergoes major rearrangements during the splicing reaction.

Pre-mRNA splicing harbors an important regulatory potential, extending the complexity of the proteome by alternative *cis* splicing: 95% of all human genes undergo alternative *cis* splicing and approximately 100,000 of these events have been detected in major human tissues (Lee & Rio, 2015). Five variants of alternative *cis* splicing can be distinguished from constitutive splicing, when all exons of a single gene are included in the mature mRNA (Figure 3.3; page 13).
(Hui & Bindereif, 2005): The selection of alternative 5´ or 3´ splice sites results in different exon lengths. Furthermore, a whole exon can be skipped or included (cassette exon). The inclusion of a cassette exon can also lead to skipping of the adjacent exon and vice versa, making them mutually exclusive exons. In case of intron retention, an intron is not removed from the transcript. Functional consequences of alternative cis splicing are altered localization, activity or function of a protein, since single exons often encode for functional protein domains (Sharp, 1994).

Figure 3.3 Alternative cis splicing patterns.
In constitutive splicing, a single mRNA is generated from the pre-mRNA. In contrast, alternative splicing produces many splice variants from a single pre-mRNA, using the following principles: Alternative 5´ or 3´ splice sites, cassette exons, mutually exclusive exons and intron retention. Exons are depicted as boxes and introns as solid lines. Constitutive and alternative splicing is indicated by green and red lines, respectively. From Hui & Bindereif, 2005; modified.

In general, alternative splicing is tissue-, development- and differentiation-specifically regulated, but it can be also activated by external stimuli like signal transduction cascades (Lee & Rio, 2015). Regulators of alternative splicing are in most cases cis-acting proteins belonging to the SR-protein and hnRNP-protein families, which specifically bind to exonic or intronic splicing enhancers and silencers. SR-proteins (e.g. SRSF1) bind predominantly to splicing enhancer elements and activate splicing (Shepard & Hertel, 2009). They display a characteristic domain structure with a C-terminal domain containing arginine-serine dipeptides (SR-domain; protein-protein interaction) and one or two N-terminal RNA recognition motifs (RRM) for RNA binding. In contrast, most hnRNP proteins (e.g. hnRNP I) bind splicing silencer elements and are involved in repression of splicing. Common domains of these proteins are RRMs that are often present in tandem (Han et al., 2010). Both SR- and hnRNP proteins often regulate splicing by affecting
early spliceosome assembly but they can regulate splicing also by other mechanisms, e.g. looping out exons as described for hnRNP I (see 3.8).

3.3 The human polyadenylation complex and its function
Cleavage and polyadenylation of Pol II-transcribed nascent pre-mRNAs is an essential step in 3′ end maturation (Xiang et al., 2014). The only exceptions are the canonical histone (H1, H2A, H2B, H3 and H4) pre-mRNAs, which are only cleaved and end with a highly conserved stem-loop (Marzluff et al., 2008). The nascent pre-mRNA is during 3′ end maturation in most cases cut at a CA dinucleotide, and a poly(A) tail of approximately 250 nts is added, first in a distributive, and after binding of PABPN1 (nuclear poly(A)-binding protein 1) in a processive manner by the poly(A) polymerase (PAP) (Brawerman, 1981, Wahle, 1991a, Bienroth et al., 1993, Wahle, 1995a, Wahle, 1995b). For the correct procession of the 3′ end, several cis-acting elements of the RNA (Figure 3.4) are required (Proudfoot, 2011): First, the highly conserved AAUAAA hexanucleotide polyadenylation signal (PAS) typically located 15 to 30 nts upstream of the cleavage site (Sheets et al., 1990, Beaudoin et al., 2000, Hu et al., 2005). However, the PAS hexamer sequence can display microheterogeneity and an AUUAAGA hexamer is common as well (Pauws et al., 2001). The second element is the GU- (YGUGUUY; Y=pyrimidine) or U-rich (UUUU) downstream sequence element (DSE) within 20 nts downstream of the cleavage site (Hart et al., 1985, McLauchlan et al., 1985, Gil & Proudfoot, 1987, MacDonald et al., 1994, Hu et al., 2005). Third, multiple UGUA upstream sequence elements (USE) are positioned 40 to 100 nts upstream of the cleavage site (Hu et al., 2005). Mutations in these sequence elements, in particular the PAS, are associated with diseases like α- and β-thalassemia when the PAS is mutated to AAUAAG or AACAAA, respectively (Higgs et al., 1983, Orkin et al., 1985).

**Figure 3.4 Consensus sequence elements of the polyadenylation signal.**
The upstream sequence element (USE), the hexameric polyadenylation signal (PAS), the CA dinucleotide (cleavage site marked by a red arrow) and the downstream sequence element (DSE) are indicated. Moreover, the respective distances between the sequence elements are shown. From Proudfoot, 2011; modified.
Cleavage and polyadenylation is carried out by a large multiprotein complex composed of more than 85 associated proteins in humans (Shi et al., 2009). The core factors of the complex are the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), cleavage factor I (CFI), cleavage factor II (CFII), Symplekin, nuclear poly(A)-binding protein 1 (PABPN1), the poly(A) polymerase and the CTD of Pol II (Figure 3.5). The above-mentioned cis-acting elements are essential for polyadenylation complex formation and directly bound by the CPSF (PAS binding), CstF (DSE binding) and CFI (USE binding).

**Figure 3.5 Simplified model of the human core cleavage and polyadenylation complex.**
The cleavage and polyadenylation specificity factor (CPSF160/100/73/30, Fip1 and WDR33; light blue) binds the hexameric AAUAAA polyadenylation signal (PAS). However, note that the PAS is directly bound to CPSF30 and WDR33, but not to CPSF160. The red arrow indicates cleavage of the RNA substrate mediated by the endonuclease CPSF73. The cleavage stimulation factor (CstF77/64/50; yellow) binds the downstream GU- or U-rich sequence element (DSE) via CstF64, and the UGUA upstream sequence element (USE) is bound to cleavage factor I (CFIm 25/68; pink). In addition, the scaffold protein Symplekin (purple), the poly(A) polymerase (PAP; red), cleavage factor II (CFII; composed of PcfI I and Clp I; green) and the RNA polymerase II (RNAP II) including the C-terminal domain (CTD) are indicated. From Xiang et al., 2014.

The CPSF subcomplex consists of six different proteins CPSF160, 100, 73 and 30 (according to their molecular mass in kDa), Fip1 and WDR33, which are required for efficient cleavage and polyadenylation, since the PAS is directly bound to CPSF and CPSF73 is the processive endonuclease (Gilmartin & Nevins, 1989, Takagaki et al., 1989, Bienroth et al., 1991, Murthy & Manley, 1992, Benz et al., 2005, Mandel et al., 2006, Shi et al., 2009, Chan et al., 2014, Schönemann et al., 2014,). Moreover, the complex is recruited to the initiation complex and accompanies Pol II throughout transcription, coupling both transcription and 3´ end processing (Dantonel et al., 1997).
The largest protein, CPSF160, is composed of tandem WD40 repeats that cluster into three major β-propellers, which generally serve as protein scaffold but can also bind nucleic acids (Neuwald & Poleksic, 2000, Scrima et al., 2008, Stirnimann et al., 2010). This is consistent with CPSF160 RNA and protein interactions (CPSF100, Fip1, CstF77 and PAP) (Murthy & Manley, 1992, Murthy & Manley, 1995).

Both CPSF100 and 73 belong to the metallo-β-lactamase superfamily, whose members are mostly hydrolases depending on metal ions (Jenny et al., 1996, Callebaut et al., 2002). CPSF100 and CPSF73 have a β-lactamase domain with a β-caspase domain (β-casp) inserted like a cassette with the active site of the protein at the interface of these two domains. However, only CPSF73 harbors the conserved residues (mainly histidine) that coordinate two zinc atoms involved in RNA cleavage making it the processive endonuclease (Mandel et al., 2006). Both proteins are tightly associated via their C-terminal domains, arguing for heterodimerization being required for cleavage, which is in line with homodimerization of other β-casp proteins (Dominski, 2007, Michalski & Steiniger, 2015).

CPSF30 consists of five C3H1 zinc fingers and a C-terminal C2HC zinc finger and binds in cooperation with WDR33 via zinc fingers two and three (minor contribution of the C-terminal zinc finger) directly to the PAS (Barabino et al., 1997, Chan et al., 2014, Schönenmann et al., 2014, Shimberg et al., 2016). In addition, CPSF30 interacts with the Pol II body and likely participates in the association of CPSF and Pol II during transcription (Nag et al., 2007).

Fip1 was discovered more than a decade later than the above mentioned proteins and interacts with the poly(A) polymerase, CPSF30, CPSF160 and CstF77 via the N-terminal region (Kaufmann et al., 2004). Moreover, the C-terminal arginine-rich domain can bind RNA, in particular U-rich sequences.

The last CPSF subunit is WDR33 consisting of an N-terminal WD40 repeats, a middle collagen-like domain and a C-terminal glycine-proline-arginine domain (Ito et al., 2001, Shi et al., 2009). WDR33 is essential for RNA binding of the polyadenylation complex since it recognizes in cooperation with CPSF30 the hexameric PAS (Chan et al., 2014, Schönenmann et al., 2014, Shimberg et al., 2016).

The second subcomplex of the polyadenylation machinery is termed CstF and composed of CstF77, 64 and 55 with two copies of each subunit (Takagaki et al., 1990, Gilmartin & Nevins, 1991). The complex couples transcription and polyadenylation via the association with Pol II and directly binds the cis-acting DSE on the RNA transcript (MacDonald et al., 1994, McCracken et al., 1997).

The largest factor, CstF77, harbors an N-terminal, half a TPR domain that is most likely involved in self-dimerization followed by a proline-rich segment, which binds the hinge domain of CstF64.
and the WD40 repeats of CstF50 bridging them, since they do not make direct contact (Takagaki & Manley, 1994, Preker & Keller, 1998, Takagaki & Manley, 2000, Bai et al., 2007, Hockert et al., 2010). Additionally, the Pol II CTD is bound to CstF77 as well (McCracken et al., 1997).

The second factor, CstF64, consists of an N-terminal RRM for DSE binding, a central hinge domain for CstF77 or Symplekin association, and a proline-glycine-rich region (Takagaki et al., 1992, MacDonald et al., 1994, Takagaki & Manley, 1997, Takagaki & Manley, 2000, Perez Canadillas & Varani, 2003, Ruepp et al., 2011b). CstF64 as well as the isoform tauCstF64 are involved in the regulation of alternative polyadenylation (APA; see 3.4 and 3.5) and may have redundant functions (Wallace et al., 1999, Li et al., 2012, Shankarling & MacDonald, 2013, Yao et al., 2013).

The last CstF subunit is CstF50 harboring an N-terminal self-dimerization domain and a C-terminal WD40 repeat domain which functions as a binding platform for CstF77 (Takagaki & Manley, 1992, Takagaki & Manley, 2000). Moreover, CstF50 binds the Pol II CTD with high affinity and couples in cooperation with CstF77 (see above) transcription and 3’ end processing (McCracken et al., 1997).

CFI is composed of two CFIm25 proteins and two proteins of the paralogous genes CFIm59, CFIm68 or CFIm72, of which the latter one is a larger isoform of CFIm68 (Rüegsegger et al., 1996, Rüegsegger et al., 1998, Ruepp et al., 2011a). CFI associates early with the transcription elongation complex, recognizes the USE to stabilize the CPSF complex and is involved in mRNA export and regulation of APA (Rüegsegger et al., 1996, Venkataraman et al., 2005, Ruepp et al., 2009, Kim et al., 2010, Ruepp et al., 2011a, Gruber et al., 2012, Martin et al., 2012). Knockdown of both CFIm25 and CFIm68 leads to an increased usage of proximal poly(A) sites, resulting in shorter 3’ untranslated regions (UTR), whereas in the wildtype situation the use of distal poly(A) sites (long 3’ UTRs) is preferred.

The smallest protein of this complex, CFIm25, encompasses a central NUDIX domain without hydrolase activity serving as a binding platform for the USE (the CFIm25 dimer binds two USEs in an antiparallel fashion via CFIm68-mediated RNA looping), CFIm68, PABPN1 and the poly(A) polymerase (Dettwiler et al., 2004, Coseno et al., 2008, Tresaugues et al., 2008, Yang et al., 2010, Li et al., 2011, Yang et al., 2011).

In contrast to CFIm59 and 72, is CFIm68 the most common second subunit of the CFI complex. It is composed of an N-terminal RRM, a middle proline-rich region and a C-terminal RS domain resembling the domain organization of the previously mentioned SR proteins (see 3.2) (Rüegsegger et al., 1998). CFIm68 is involved in USE binding (see above) and interacts with the nuclear export machinery through the Thoc5 protein (TREX complex) and the export receptor NXF1/TAP (Gooding et al., 1998, Ruepp et al., 2009, Li et al., 2011, Yang et al., 2011).
CFII is the least-well characterized subcomplex of the polyadenylation machinery, since the factors with the exception of Clp I and PcfI I remain poorly defined and were mostly studied in yeast (de Vries et al., 2000). Knockdown of PcfI I impairs RNA cleavage and transcription termination, and Clp I was shown to interact with CPSF as well as CFI and likely tethers them to CFIIIm (de Vries et al., 2000, West & Proudfoot, 2008).

Symplekin is an essential scaffold protein of the polyadenylation complex and contains seven pairs of antiparallel α-helices (ARM repeat) at the N-terminus that are often involved in protein-protein interactions (Takagaki & Manley, 2000, Andrade et al., 2001, Kennedy et al., 2009, Xiang et al., 2010). The central domain binds the hinge domain of CstF64 and, in addition, Symplekin associates with both CPSF73 and CPSF100 forming a stable complex required for RNA cleavage (Takagaki & Manley, 2000, Hofmann et al., 2002, Sullivan et al., 2009, Ruepp et al., 2011b).

The canonical, nuclear localized poly(A) polymerase of the polyadenylation core complex, which adds the poly(A) tail to the pre-mRNA, is encoded by the human PAPOLA gene and belongs to the nucleotidyltransferase superfamily of DNA polymerase β (Edmonds, 1990, Raabe et al., 1991, Wahle, 1991a, Wahle, 1991b, Colgan & Manley, 1997). At the N-terminus of the protein, the catalytic nucleotidyltransferase domain is inserted like a cassette in the PAP central domain (Raabe et al., 1994, Zhelkovsky et al., 1995, Martin & Keller, 1996, Martin et al., 1999, Martin et al., 2000). The catalytic domain contains the conserved aspartate triad that coordinates the magnesium ions that are essential for catalysis and ATP binding. Following the PAP central domain, a RNA-binding region and a C-terminal nuclear localization signal (NLS) are located. The general processivity of the poly(A) polymerase is regulated by the PABPN1 protein and the poly(A) tail is first synthesized in a distributive, and after binding of PABPN1 to the poly(A) tail, in a processive manner (Wahle, 1991a, Bienroth et al., 1993, Wahle, 1995b).

As discussed above, PABPN1 stimulates the catalytic activity of the poly(A) polymerase to ensure a proper length of the poly(A) tail (see above) and may also regulate APA (see 3.4 and 3.5) (Jenal et al., 2012). The middle coiled-coiled domain directly stimulates the poly(A) polymerase and the following RRM and C-terminal region are required for RNA binding, allowing PABPN1 to coat the newly synthesized poly(A) tail (Nemeth et al., 1995, Keller et al., 2000, Kerwitz et al., 2003). It was proposed that this structure restricts the CPSF complex to the PAS, to facilitate the interaction between CPSF and the poly(A) polymerase until the correct length of the poly(A) tail is reached (Kuhn et al., 2009).

The last core subunit of the polyadenylation complex is the Pol II CTD that links both transcription and polyadenylation and is directly required for the latter one (McCracken et al.,
1997, Hirose & Manley, 1998). However, the precise mechanism is not yet understood but a platform role was proposed, since several proteins (e.g. CstF77 and CstF50) bind to the CTD. Recent data from two studies suggested a new model for the previously introduced CPSF complex, dividing it into a RNA-binding module composed of CPSF160, CPSF30, Fip1 as well as WDR33 and a RNA-cleavage module built by CPSF100, CPSF73 and Symplekin (Chan et al., 2014, Schönemann et al., 2014, Shimberg et al., 2016). The hexameric PAS is directly bound to CPSF30 (zinc fingers two and three) and WDR33, whereas Fip1 binds U-rich sequences upstream and downstream of it. Moreover, CPSF160 binds RNA upstream of the cleavage site over a brought region (Figure 3.6). RNA cleavage is catalyzed by the endonuclease CPSF73 (see above) of the cleavage module, which is only essential for cleavage but not for PAS binding and polyadenylation.

*Figure 3.6 A new model for CPSF-RNA interaction.*

Multiple proteins mediate CPSF-RNA interactions: CPSF30 and Wdr33 directly interact with the hexameric AAUAAA polyadenylation signal (PAS), and CPSF30 binds to the AAUAAA motif via its zinc fingers two and three (shown as two fingers). Fip1 binds the U-rich auxiliary sequence upstream and downstream of the AAUAAA motif (shown only upstream of the AAUAAA). CPSF160 binds to RNA over a broad region upstream of the cleavage site. CPSF73, CPSF100 and Symplekin do not participate in AAUAAA binding, but function as the RNA cleavage module mediated by the endonuclease CPSF73. From Chan et al., 2014.

### 3.4 Alternative polyadenylation (APA) and its biological importance

Recent discoveries showed that most human genes contain more than one polyadenylation site, suggesting that APA is a widespread phenomenon generating RNA transcripts with different 3´ ends (Millevoi & Vagner, 2010, Shi, 2012, Elkon et al., 2013). Four classes of APA events can be distinguished (Figure 3.7; page 20): Tandem 3´ UTR APA and alternative-terminal-exon APA are the most frequent, the less frequent intonic APA and the last frequent internal-exon APA type.

In tandem 3´ UTR APA, alternative poly(A) sites are located in the last exon of the transcript, allowing the generation of mRNAs with different 3´ UTR lengths via the usage of proximal or distal poly(A) sites without any effect on the encoded protein. In contrast, the other three APA
types can directly affect the coding potential of the mRNA. Alternative splicing (see 3.2) can generate transcripts with alternative-terminal-exons causing the second APA type. An intronic, cryptic poly(A) site is activated during intronic APA resulting in the extension of an internal-exon making it the terminal one. Finally, in internal-exon APA, a poly(A) site in an internal exon is used for premature cleavage and polyadenylation in the coding region of the mRNA.

Figure 3.7 The four different APA types.
Tandem 3´ UTR APA involves alternative poly(A) sites within the same terminal-exon and hence generates multiple isoforms that differ in their 3´ UTR length, without affecting the protein encoded. The three other APA types potentially affect the coding sequence in addition to the 3´ UTR. In alternative-terminal-exon APA, alternative splicing generates isoforms that differ in their last exon. Intronic APA involves cleavage at a cryptic, intronic poly(A) site extending an internal-exon making it the terminal one. Internal-exon APA involves premature cleavage and polyadenylation within an internal-exon of the coding region. Internal-exons are depicted as large boxes, 3´ UTRs as small red boxes and introns as solid black lines. Constitutive spliced (green lines) exons are shown in blue and alternatively spliced (orange lines) exons in green. From Elkon et al., 2013; modified.

APA increases like alternative splicing (see 3.2) the complexity of the transcriptome by generating mRNA isoforms that differ either in their coding sequence or in their 3´ UTR, affecting protein function, mRNA stability or translation efficiency (Andreassi & Riccio, 2009, Fabian et al., 2010). In particular, the 3´ UTR is very important since it functions as a binding platform for miRNAs that mainly regulate translation efficiency and for trans-acting proteins that influence mRNA stability.
Diverse biological processes are linked to APA: Many tissue-specific polyadenylation events and global trends in poly(A) site selection (brain and nervous system: distal poly(A) site; placenta, ovaries and blood: proximal poly(A) site) argue for a general role in development and cellular differentiation (Zhang et al., 2005). Perhaps the most profound association of APA was found with cellular proliferation (Sandberg et al., 2008, Elkon et al., 2012): The activation of T cell proliferation is linked to a usage of proximal poly(A) sites, and a general shortening of 3´ UTRs was also observed in cancer cells (Mayr & Bartel, 2009, Lin et al., 2012, Morris et al., 2012). This can most likely be explained by binding sites for miRNAs or trans-acting proteins in the 3´ UTR that can negatively affect mRNA stability or translation efficiency. In addition, APA is involved in neuronal activation upon external stimuli (Timmusk et al., 1993, An et al., 2008, Lau et al., 2010). The brain-derived neurotrophic factor gene contains two poly(A) sites, allowing the generation of two isoforms with either a short 3´ UTR, which is translated into protein at basal levels, or a long 3´ UTR only translated after external stimulation. Such a regulation is most likely mediated by a trans-acting factor like an RNA-binding protein or a miRNA, but remains elusive.

3.5 Regulation of polyadenylation

The above described APA (see 3.4) requires precise regulation, since it is involved in diverse biological processes, and dysregulation can result in non-functional proteins and is generally associated with cancer (Millevoi & Vagner, 2010, Shi, 2012, Elkon et al., 2013). Differential regulation of the polyadenylation efficiency of a transcript containing a single poly(A) site directly affects protein expression, since unprocessed transcripts are degraded or not exported to the cytoplasm (Millevoi & Vagner, 2010). Three simplified models for the regulation of polyadenylation have been proposed: “First come, first serve”, “Survival of the fittest” and “Agonist/Antagonist” (Davis & Shi, 2014).

The “First come, first serve” model describes the predominant usage of the proximal poly(A) site since it is transcribed earlier, reflecting a direct linkage to transcription. This has been observed for both slow Pol II mutants in Drosophila melanogaster and drugs that reduce the transcription elongation rate (Pinto et al., 2011, Yu & Volkert, 2013). Further studies also suggested a direct influence of the chromatin structure because it often influences the elongation rate under physiological conditions (Brown et al., 2012). The promoter of the retrogene Mcts2 is located within an intron of the mouse H13 gene, which is due to CpG methylation imprinted and only expressed from the paternal allele (Wood et al., 2008, Cowley et al., 2012). Depending on the Mcts2 expression status, either the proximal (Mcts2 expressed) or the distal (Mcts2 not expressed) H13 poly(A) site is used, indicating that the expression of the retrogene reduces the transcription rate favoring the usage of the proximal H13 poly(A) site.
The “Survival of the fittest model” is based on the observation that distal poly(A) sites are often stronger than proximal ones because their PAS and DSE better resemble the conserved canonical sequences (Martin et al., 2012, Smibert et al., 2012, Tian & Graber, 2012, Yao et al., 2012, Lackford et al., 2014). However, poly(A) sites are regulated by additional factors including the concentration of various core polyadenylation complex constituents (Takagaki et al., 1996, Takagaki & Manley, 1998). For example, Immunoglobulin M (IgM) is subjected to intronic APA that produces an mRNA encoding a secreted protein, whereas polyadenylation at the distal site generates a membrane-bound IgM. This process is directly controlled by the up-regulation of CstF64 levels during B cell activation: In contrast to low CstF64 levels, where the strong, distal poly(A) site is used, high CstF64 levels favor the weak proximal poly(A) site. This is in line with an increased usage of distal poly(A) sites upon co-depletion of CstF64 and tauCstF64, indicating that both factors are essential for proximal poly(A) site activation (Yao et al., 2012). Depletion of CFIm25 or CFIm68 has the opposite effect, the activation of proximal poly(A) sites, suggesting the requirement of both factors for the maintenance of long 3´ UTRs (Gruber et al., 2012, Martin et al., 2012).

The last model “Agonist/Antagonist” is based on individual positive or negative effects on poly(A) site choice by trans-acting factors such as RNA-binding proteins (Shi, 2012). The polypyrimidine tract binding protein (PTB/hnRNP I) represses poly(A) sites via competing with CstF64 for the downstream GU/U-rich sequence element, resulting in an inhibition of mRNA 3´ end cleavage and an accumulation of unprocessed pre-mRNAs (Castelo-Branco et al., 2004). In addition, PTB can activate polyadenylation by, first, promoting RNA binding of hnRNP H close to the polyadenylation site, which in turn facilitates the recruitment of CstF or the poly(A) polymerase, and, second, by increasing the processivity of the poly(A) polymerase (Millevoi et al. 2009). Finally, general splicing factors can also participate in the regulation of polyadenylation. The U1 snRNP is essential for cis splicing and binds to nascent pre-mRNAs repressing cryptic poly(A) sites, often located in introns preventing premature cleavage and polyadenylation (Kaida et al., 2010, Berg et al., 2012).

### 3.6 The African trypanosomes: Trypanosoma brucei

*Trypanosoma brucei* is a protozoan parasite belonging to the order of Kinetoplastida and the name of the order is derived from the kinetoplast, the DNA-containing compartment in their single mitochondrion (Achcar et al., 2014). The parasite undergoes a complex life-cycle in an insect vector, the tsetse fly (*Glossina spp.*) and a vertebrate host. *Trypanosoma brucei* can be divided in three subspecies causing different forms of trypanosomiasis, a lethal tropical disease common in Africa (Malvy & Chappuis, 2011, Migchelsen et al., 2011, Lejon et al., 2013): *T.*
brucei gambiense is most prevalent in central and western Africa and causes chronic human trypanosomiasis, leading to death of the untreated patient within three years. In contrast, T. brucei rhodesiense causes acute human trypanosomiasis, which is deadly within six months and is most common in southern and eastern Africa. The last subspecies is T. brucei brucei that is non-infective for humans but causes the Nagana cattle disease (also caused by several other Trypanosoma species). About 10,000 people get infected every year by human trypanosomiasis, which in over 96% is caused by T. brucei gambiense, and 70 million people in Africa live in the risk of being infected. Moreover, Trypanosoma brucei became due to some special biochemical and cellular characteristics an important model organism: The expression of variant surface glycoproteins allowing the evasion of the hosts immune response, and RNA editing of mitochondrial mRNAs (deletion or insertion of uridine nucleotides) by guide RNAs and the editosome. The special characteristics in gene expression, namely spliced-leader trans splicing and the polycistronic organization of protein-coding genes, which will be discussed in the following (see 3.7).

### 3.7 Expression of protein-coding genes in trypanosomes

In trypanosomes all protein-coding genes are arranged in long, polycistronic transcription units (PTU) with up to 100 functionally unrelated genes, which are transcribed by Pol II (Figure 3.8; page 24) (Johnson et al., 1987, Berriman et al., 2005). Genes within a PTU are transcribed from the same strand, whereas neighboring PTUs are transcribed from the opposite strand separated by a strand switch region (SSR). Transcription of neighboring PTUs is either divergent or convergent and the SSR functions as transcription start site (TSS) or transcription termination site (TTS), respectively (Martinez-Calvillo et al., 2003, El-Sayed et al., 2005). Surprisingly, no Pol II promoter sequence has been identified in trypanosomes with the exception of the SL RNA gene (see below and 3.2), suggesting that transcription of protein-coding genes is not regulated and proceeds roughly at the same rate (Luo et al., 1999, Ruan et al., 2004, Das et al., 2005, Schimanski et al., 2005a, Palenchar & Bellofatto, 2006, Schimanski et al., 2006, Lee et al., 2007, Lee et al., 2009). An enrichment of specific histone modifications and variants at both TSS and TTS argues for an epigenetic mechanism of transcription initiation and termination: The histone modifications H4K10ac and H3K4me3, histone variants H2A.Z and H2B.Z and the bromodomain factor 3 protein (Bdf3), which binds acetylated lysine, are enriched at convergent SSRs functioning as transcription start sites (Hassan et al., 2002, Siegel et al., 2009, Wright et al., 2010). Nucleosomes containing H2A.Z are less stable, suggesting less condensed chromatin-favoring transcription (Jin & Felsenfeld, 2007, Siegel et al., 2009). In contrast, transcription terminates at convergent SSRs, which display an enrichment of H3K76m1/2, H3.V,
H4.V and base J, suggesting that they serve as marks for transcription termination, and two recent studies showed indeed that H3.V and base J are required for transcription termination (Siegel et al., 2009, Gassen et al., 2012, van Luenen et al., 2012, Reynolds et al., 2016, Schulz et al., 2016). Moreover, Pol I transcribed tRNA genes are often located at TTS and may contribute to termination of Pol II transcription (Hull et al., 1994).

Figure 3.8 Organization of polycistronic transcription units in trypanosomes.
Transcription of polycistronic transcription units (PTU; indicated as colored bold arrow) starts at (transcription start site, TSS) divergent (div.) strand switch regions (SSR) and terminates (transcription termination site, TTS) at convergent (conv.) SSRs. Genes within a PTU are transcribed from the same strand, whereas neighboring PTUs are transcribed from opposing strands separated by a SSR. At TSS H4K10ac and H3K4me3, H2A.Z, H2B.Z and bromodomain factor 3 protein (Bdf3) are enriched, whereas at TTS H3K76m1/2, H3.V, H4.V, base J and tRNA genes are enriched. From Siegel et al., 2011; modified.

After transcription of polycistronic transcription units, single mRNAs need to be processed by two coupled mechanisms, namely trans splicing and polyadenylation (Figure 3.9; page 25) (Boothroyd & Cross, 1982, Agabian, 1990, Huang & van der Ploeg, 1991a, Huang & Van der Ploeg, 1991b, LeBowitz et al., 1993, Ullu et al., 1993, Matthews et al., 1994, Benz et al., 2005, Günzl, 2010, Preußer et al., 2012, Preußer et al., 2012). Both processes are regulated by the polypyrimidine tract upstream of a trans splice site, which affects polyadenylation of the upstream gene and trans splicing of the downstream gene. During trans splicing a 39 nts long miniexon derived from the SL RNA, which already contains a highly modified m7G cap structure termed cap4, is added to the 5´ end of all mRNAs (Bangs et al., 1992). The trans splicing reaction is catalyzed by the spliceosome (see 3.2) and it is assumed that the SL RNP replaces the U1 snRNP. However, two genes, the putative poly(A) polymerase Tb927.3.3160 and an ATP-dependent DEAD box helicase Tb927.8.1510, contain an intron with a conserved sequence and require additional U1 snRNP-dependent cis splicing (see 3.2) (Tschudi & Ullu, 1990, Mair et al., 2000, Liang et al., 2003, Berriman et al., 2005, Kolev et al., 2010, Siegel et al., 2010, Tkacz et al., 2010).
Figure 3.9 mRNA processing in trypanosomatids.
Schematic representation of the major mRNA processing steps in trypanosomatids. After transcription of polycistronic transcription units (colored boxes), which are separated by strand switch regions (SSR), the pre-mRNAs undergo trans splicing and polyadenylation [(A)n]. Each of the protein-coding genes (shown in different colors) is processed through trans splicing, whereby the SL RNA with its cap 4 structure provides the 5′ exon (miniexon, in magenta); in addition, very few genes require cis splicing of an internal intron (see green boxes). Both cis and trans splicing proceed through two transesterification steps and require the U2, U4/U6, and U5 small nuclear ribonucleoproteins (snRNPs) as well as many additional protein factors. Moreover, cis splicing requires the U1 snRNP, trans splicing the SL RNP as specific factors. From Preußer et al., 2012.

The 3´ end of each mRNA is processed by cleavage and polyadenylation (see 3.3 and 3.4), which is catalyzed in other eukaryotes by the polyadenylation complex. However, in trypanosomes most constituents of the polyadenylation machinery remain still elusive. So far only three subunits of the CPSF complex -CPSF30, CPSF73, and Fip1- have been identified as orthologues of their mammalian counterparts (see 3.3) and were biochemically characterized to some extent. T. brucei CPSF30 has a conserved domain structure similar to its human counterpart, harboring five C3H1 zinc fingers and a C-terminal C2HC zinc finger. CPSF30 is essential for cell viability and processing of polycistronic pre-mRNAs, since its depletion results in an accumulation of dicistronic α/β-tubulin pre-mRNA (Hendriks et al., 2003). Fip1 is mainly localized in the nucleus, essential for cell viability and contains the conserved Fip1 domain, which is required for interaction with CPSF30 (zinc fingers four and five) in trypanosomes (Bercovich et al., 2009). Moreover, the protein harbors an C3H1 zinc finger and a histidine-
rich) and a proline-rich (P-rich) region. CPSF73 was recently co-purified with the spliceosomal U1 snRNP protein U1A, arguing for a coupling of both pre-mRNA processing reactions (Tkacz et al., 2010). However, even for the poly(A) polymerase, in other eukaryotes probably the best delineated component, a definitive identification and biochemical characterization have not been accomplished in trypanosomes. At least two genes code for putative poly(A) polymerases [Tb927.3.3160 and Tb927.7.3780], the first of which is peculiar in that it represents one of the two genes in trypanosomes that require cis splicing (see above). In addition, the polyadenylation signal itself is poorly defined, since the two canonical cis-acting elements, the AAUAAA motif (PAS) and the GU/U-rich sequence (DSE), are not conserved in trypanosomes. The only known cis-acting element required for polyadenylation is the above mentioned polypyrimidine tract.

Genome-wide data showed that most genes contain one to three trans splice- and even more dispersed polyadenylation sites, allowing both alternative trans splicing and polyadenylation (Kolev et al., 2010, Nilsson et al., 2010, Siegel et al., 2010, Veitch et al., 2010, Siegel et al., 2011). For alternative trans splicing four biological roles have been discussed (Figure 3.10 A; page 28) (Preußer et al., 2012): (1) Skipping of the canonical AUG start codon; (2) in- or exclusion of targeting signals; (3) in- or exclusion of regulatory elements (e.g. upstream open reading frames); (4) usage of an alternative ORF. However, only for the in- or exclusion of targeting signals a specific example has been described so far (Rettig et al., 2012). Alternative trans splicing creates a long and a short isoform of the isoleucyl-tRNA synthetase mRNA. The long isoform encodes for a protein, which is due to an additional N-terminal signal sequence imported into the mitochondrion, whereas the second protein isoform, encoded by the short mRNA, remains in the cytoplasm. This indicates an important role of alternative trans splicing in post-transcriptional gene regulation in trypanosomes. However, trans-acting factors, which contribute to the regulation of alternative splicing still remain elusive. So far only a few splicing repressors, such as the trypanosomatid orthologue of hnRNP F/H, and splicing activators, like the T. brucei PTB homologs DRBD3/4 (see 3.8), have been identified (Stern et al., 2009, Gupta et al., 2013). With to alternative polyadenylation, which may affect mature transcript levels by altering the length of the 3´ UTR and also 5´ UTR of the downstream gene, so far neither any regulatory elements nor trans-acting factors have been identified (Figure 3.10 B; page 28). Variations in the length of the 3´ UTR and 5´ UTR possibly alter the inclusion of important regulatory elements in the mature mRNA. These RNA elements, located mainly in the 3´ UTR, are of special interest in trypanosomes, since gene expression is regulated mostly on the post-transcriptional level with effects on either mRNA stability or translation efficiency (Clayton & Shapira, 2007, Haile & Papadopoulou, 2007, Kramer & Carrington, 2011, Kramer, 2012, Clayton, 2013, Clayton, 2014, Kolev et al., 2014). Key players in the regulation of mRNA stability
are trans-acting RNA-binding proteins: As an example, the RRM-containing proteins DRBD3 and DRBD4 are involved in positive and negative regulation of mRNA stability, and in case of DRBD3 it was shown that the regulatory sequence elements are located in the mRNA 3´ UTR (Estevez, 2008, Stern et al., 2009). The best understood example is the stabilization of chaperone mRNAs during heat shock of bloodstream form trypanosomes by the ZC3H11 zinc finger protein (Delhi et al., 2011, Droll et al., 2013, Singh et al., 2014). The N-terminal zinc finger binds to UAU repeats in the mRNA 3´ UTR, while the C-terminal domain is required for stabilization. ZC3H11 interacts with MKT1 and PBP1, which in turn interacts with PABP that is known to increase mRNA stability when recruited to the 3´ UTR. RNA degradation in trypanosomes is in general similar to other eukaryotes. First, the RNA is deadenylated by the CAF1-NOT complex (CAF1 is the processive deadenylase) and afterwards decapped at the 5´ end by an enzyme still not identified (Milone et al., 2002, Schwede et al., 2008, Schwede et al., 2009, Farber et al., 2013, Erben et al., 2014). Finally, the mRNA is either degraded from the 5´ end by the exonuclease Xrna or from the 3´ end by the exosome (Clayton & Estevez, 2011, Manful et al., 2011).
Figure 3.10 Biological functions of alternative trans splicing and polyadenylation.

Alternative trans splicing (A) and polyadenylation (B) patterns of pre-mRNAs and the different resulting mature mRNAs are schematically shown. (A) Protein-coding exons (green or yellow) and the SL RNA-derived miniexon (SL) are represented as large boxes, targeting signals and regulatory elements as small boxes (red or blue) and intergenic regions as black lines. Four potential biological roles of alternative trans splicing have been reported in trypanosomes, including skipping of the AUG start codon, in- or exclusion of targeting signals, in- or exclusion of regulatory elements, and the usage of an alternative open reading frame (orange box). (B) Protein-coding genes (green or gray) and the SL RNA-derived miniexon (SL) are represented as large boxes. The intergenic region forming alternative 5´ and 3´ UTRs between two genes is indicated by small colored boxes (blue, purple and yellow). Alternative polyadenylation affects the 3´ UTR (tandem 3´ UTR APA) and can, due to the polycistronic gene organization, also affect the 5´ UTR of the downstream gene by altering trans splice site choice (APA affecting the 5´ UTR). From Preußer et al., (2012); modified.
3.8 The trypanosomatid PTB/hnRNP I homologs DRBD3 and DRBD4
The *Trypanosoma brucei* homologs of vertebrate PTB are DRBD3 and DRBD4 (De Gaudenzi *et al.*, 2005, Stern *et al.*, 2009). Vertebrate PTB belongs to the hnRNP protein family (see 3.2) and is a classical repressor of alternative splicing (Romanelli *et al.*, 2013). PTB is a shuttling protein mainly localized in the nucleus that binds via its four RRMs 15-25 nts long polypyrimidine tracts containing UCUUC and CUCUCU elements (Perez *et al.*, 1997, Oberstrass *et al.*, 2005, Ray *et al.*, 2009, Reid *et al.*, 2009, Lamichhane *et al.*, 2010, Maynard & Hall, 2010). The RRMs one and two are connected by a flexible linker allowing independent RNA binding, whereas the RRMs three and four interact, resulting in an antiparallel orientation of their bound RNA. Moreover, the polypyrimidine motifs bound to the RRMs three and four are separated by a minimum of 15 nts, allowing RNA looping. The expression of PTB and the neuronal-specific paralog nPTB is auto- and cross-regulated by alternative splicing (Markovtsov *et al.*, 2000, Rahman *et al.*, 2002, Wollerton *et al.*, 2004). PTB represses splicing of its own exon 11, causing nonsense-mediated decay (NMD) and promotes skipping of nPTB exon 10, causing NMD as well. However, in neuronal cells nPTB autoregulates exon 10 inclusion, resulting in higher nPTB protein levels (Ni *et al.*, 2007).

Several PTB-mediated mechanisms for repression of exon inclusion during cis splicing have been characterized: PTB binding sites within the polypyrimidine tract overlap with U2AF and U2 snRNP binding sites, arguing for a competition of these factors for their respective binding sites resulting in regulation of splicing. Based on the RNA-binding properties of PTB (see above) it can also loop out the branch point adenosine or a whole exon leading to exon skipping (Liu *et al.*, 2002, Oberstrass *et al.*, 2005, Spellman *et al.*, 2005, Auweter & Allain, 2008, Lamichhane *et al.*, 2010, Maynard & Hall, 2010).

Additionally, PTB can also promote exon inclusion and two different models have been proposed. First, via PTB binding to a polypyrimidine stretch positioned exclusively downstream of the respective regulated exon (Izquierdo *et al.*, 2005). Second, by binding near the splice sites of the exon antagonizing the action of different splicing repressors (Xue *et al.*, 2009).

In addition to splicing, PTB also increases the stability of various mRNAs e.g. pre-proinsulin, prohormone convertase 1/3 and 2 (Knoch *et al.*, 2004, Fred & Welsh, 2009).

Moreover, PTB can both stimulate and repress polyadenylation: PTB activates polyadenylation by, first, promoting RNA binding of hnRNP H close to the polyadenylation site, which in turn facilitates the recruitment of CstF or the poly(A) polymerase, and, second, by increasing the processivity of the poly(A) polymerase (Millevoi *et al.* 2009). In contrast, PTB represses polyadenylation by competing with CstF64 for the downstream GU/U-rich sequence element,
resulting in an inhibition of 3’ end cleavage and an accumulation of unprocessed pre-mRNAs (Castelo-Branco et al., 2004).

The trypanosomatid homologs DRBD3 and DRBD4 are required for cell viability and contain two and four RRM domains, respectively (De Gaudenzi et al., 2005, Stern et al., 2009). In contrast to DRDB4, which is predominantly nuclear, DRBD3 can be detected in both cytoplasm and nucleus (Stern et al., 2009, Fernandez-Moya et al., 2012). Both proteins are no general splicing factors but are required for efficient trans splicing of genes having a C-rich polypyrimidine tract and moreover, DRBD3 is needed for cis splicing. In addition, DRBD3 and 4 regulate the stability of different subsets of mRNAs either in a positive or negative manner (Estevez, 2008, Stern et al., 2009). DRBD3 was shown to have a polypyrimidine-rich binding motif and to bind mRNAs preferentially in their 3’ UTR (Das et al., 2015). Moreover, DRBD3 is involved in the intracellular trafficking of bound mRNAs in form of ribonucleoprotein complexes, in response to stress conditions such as heat shock and starvation (Fernandez-Moya et al., 2012).

3.9 Aims of this work

The expression of polycistronic protein-coding genes in the protozoan parasite Trypanosoma brucei requires coupling of the two major mRNA processing steps, trans splicing and polyadenylation, to produce mature mRNAs (see 3.7). The protein components, the mechanisms and dynamics of both processes and in particular their mechanistic linkage are still not well characterized. Although studies in the last decade have identified numerous spliceosomal components, we still know very little about the composition and functioning of the polyadenylation machinery in trypanosomes: Only CPSF73, CPSF30 and Fip1 had previously been identified and to some extent functionally described.

The first aim of this work was the initial characterization of the polyadenylation complex in T. brucei, including the identification of both the functional poly(A) polymerase and the other constituents of the polyadenylation machinery. Afterwards the proteins should be further investigated with regard to their role in trans splicing and polyadenylation, since these processes are coupled in vivo.

The second aim was to analyze the regulation of polyadenylation in trypanosomes. This was based on the exemplary DRBD4 protein, which is a homolog of human PTB that is known to regulate both splicing and polyadenylation in vivo. As a starting point for our functional analysis, we planned to determine both the DRBD4 RNA-binding motif and genome-wide binding sites. Afterwards DRBD4-dependent poly(A) sites should be identified and analyzed for DRBD4 binding, with the aim to characterize poly(A) sites directly regulated by DRBD4 in vivo.
4. Material and Methods

4.1 Material

4.1.1 Chemicals and reagents

2 (β)-mercaptoethanol Roth
4´,6-diamidino-2-phenylindole (DAPI) Roth
Acetic acid Roth
Agarose ultra pure Roth
Amberlite® MB-1 hydrogen and hydroxide form Sigma-Aldrich
Ammonium persulfate (APS) Bio-Rad
Ampicillin Roth
Antibiotic-antimycotic (100×) ThermoFischer Scientific
Anti-protein C affinity matrix Roche Applied Science
Blasticidin Sigma-Aldrich
Boric acid Roth
Bovine serum albumin (20 mg/ml) (BSA) Roche Applied Science
Bromphenol blue Merck
Calcium chloride (CaCl$_2$) Merck
Cellfectin® II reagent ThermoFischer Scientific
Chloroform Roth
Ciprofloxacin AppliChem
Cold-water fish gelatin Sigma-Aldrich
Dimethyl pyrocarbonate (DMPC) Sigma-Aldrich
Dimethyl sulfoxide (DMSO) Sigma-Aldrich
Dipotassium phosphate (K$_2$HPO$_4$) Roth
Disodium hydrogen phosphate (Na$_2$HPO$_4$·2H$_2$O) Roth
Dithiothreitol (DTT) Roth
DNAzol® ThermoFischer Scientific
Doxycyclin (Dox) Sigma-Aldrich
Ethanol (≥ 99,8%) Roth
Ethidium bromide Roth
Ethylene-diaminetetraacetic acid (EDTA) Roth
Ethylene glycol tetraacetic acid (EGTA) Sigma-Aldrich
<table>
<thead>
<tr>
<th>Material and Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetal bovine serum (FBS)</td>
</tr>
<tr>
<td>Ficoll® PM 400</td>
</tr>
<tr>
<td>Formamide</td>
</tr>
<tr>
<td>Genitalicin</td>
</tr>
<tr>
<td>Glucose</td>
</tr>
<tr>
<td>Glutamin</td>
</tr>
<tr>
<td>Glutathione sepharose 4B beads</td>
</tr>
<tr>
<td>Glycerol</td>
</tr>
<tr>
<td>Glycine</td>
</tr>
<tr>
<td>GlycoBlue (15 mg/ml)</td>
</tr>
<tr>
<td>Glycogen (20 mg/ml)</td>
</tr>
<tr>
<td>Glyoxal solution ~40% in H₂O (~8.8 M)</td>
</tr>
<tr>
<td>Hemin</td>
</tr>
<tr>
<td>Heparin sodium salt</td>
</tr>
<tr>
<td>Hygromycin</td>
</tr>
<tr>
<td>Hypoxanthine</td>
</tr>
<tr>
<td>IgG sepharose™ 6 fast flow</td>
</tr>
<tr>
<td>Immu-Mount™</td>
</tr>
<tr>
<td>InstantBlue-protein stain</td>
</tr>
<tr>
<td>Isopropanol</td>
</tr>
<tr>
<td>Isopropyl-β-D-thiogalactopyranosid (IPTG)</td>
</tr>
<tr>
<td>LB-agar (Luria/Miller)</td>
</tr>
<tr>
<td>LDS loading-buffer (4×)</td>
</tr>
<tr>
<td>L-proline</td>
</tr>
<tr>
<td>L-threonine</td>
</tr>
<tr>
<td>Magnesium chloride (MgCl₂)</td>
</tr>
<tr>
<td>Mangan chloride (MnCl₂)</td>
</tr>
<tr>
<td>Methanol</td>
</tr>
<tr>
<td>Milk-powder (fat-free)</td>
</tr>
<tr>
<td>Monopotassium phosphate (KH₂PO₄)</td>
</tr>
<tr>
<td>N-2-hydroxyethylpiperazine (HEPES)</td>
</tr>
<tr>
<td>N,N,N’,N’-tetramethylenediamine (TEMED)</td>
</tr>
<tr>
<td>Nonidet P-40 (NP-40)</td>
</tr>
<tr>
<td>PageBlue™ protein staining solution</td>
</tr>
<tr>
<td>Penicillin G</td>
</tr>
</tbody>
</table>
4. Material and Methods

Phleomycin: Sigma-Aldrich
Phosphate-buffered saline (PBS; 10×): ThermoFischer Scientific
Polyoxyethylene (20) sorbitan monolaurate (Tween 20): Sigma-Aldrich
Polyvinyl alcohol (PVA): Sigma-Aldrich
Polyvinylpyrrolidone: Sigma-Aldrich
Potassium chloride (KCl): Roth
Potassium phosphate (KH₂PO₄): Roth
Puromycin: Sigma-Aldrich
Reducing agent (10×): ThermoFischer Scientific
Rotiphorese gel 30: Roth
(30% Acrylamide/bisacrylamide 37.5:1): Roth
Rotiphorese gel 40: Roth
(40% Acrylamide/bisacrylamide 19:1): Roth
SDM-79 CGGGPPTA: PAA
Sf-900™ II SFM: ThermoFischer Scientific
Sodium acetate: Merck
Sodium azide (NaN₃): Merck
Sodium chloride (NaCl): Roth
Sodium citrate: Roth
Sodium dodecyl sulfate (SDS): Roth
Sodium pyrophosphate tetrabasic (Na₄P₂O₇): Sigma-Aldrich
Sodium pyruvate: Sigma-Aldrich
StrataClean resin: Agilent Technologies
Tris-hydroxymethylaminomethane (Tris): Roth
Triton X-100: Merck
TRIzol® reagent: ThermoFischer Scientific
tRNA from yeast: Roche Applied Science
Trypton: Roth
Urea: Roth
Xylene cyanole: Fluka
Yeast extract: Roth
4. Material and Methods

4.1.2 Commercial Kits

- Bac-to-Bac® baculovirus expression system: ThermoFischer Scientific
- Bio-Rad silver stain: Bio-Rad
- HiScribe™ T7 high yield RNA synthesis kit: Agilent Technologies
- LabChip XT DNA 750 assay kit: PerkinElmer
- Lumi-Light western blotting substrate: Roche Applied Science
- PerfeCTa® SYBR® green FastMix®: Quanta Biosciences
- Prime-It® RmT random primer labeling kit: Agilent Technologies
- Qiaquick gel extraction kit: QIAGEN
- Qiaquick PCR purification kit: QIAGEN
- qScript™ cDNA synthesis kit: Quanta Biosciences
- qScript™ flex cDNA kit: Quanta Biosciences
- Qubit® dsDNA HS assay kit: ThermoFischer Scientific
- RNeasy mini kit: QIAGEN

4.1.3 Enzymes and enzyme inhibitors

- Accuprime supermix 1: ThermoFischer Scientific
- AcTEV protease (10 U/µl): ThermoFischer Scientific
- Circligase II (100 U/µl): Epicentre
- cOmplete™ EDTA-free protease inhibitor: Roche Applied Science
- Fast digest BamHI: ThermoFischer Scientific
- Phusion® high-fidelity DNA polymerase: New England Biolabs
- Proteinase K (PK; 20 mg/ml): Roth
- Restriction endonucleases: New England Biolabs
- RNase I (100 U/µl): ThermoFischer Scientific
- RNase A/T1 Mix (2 mg/ml; 5000 U/ml): ThermoFischer Scientific
- RNaseOUT RNase inhibitor (40 U/µl): ThermoFischer Scientific
- RQ1-DNase (1 U/µl): Promega
- Shrimp alkaline phosphatase (SAP; 1 U/µl): ThermoFischer Scientific
- Superscript III reverse transcriptase (200 U/µl): ThermoFischer Scientific
- T4 DNA ligase (400 U/µl): New England Biolabs
- T4 polynucleotide kinase (PNK; 10 U/µl): New England Biolabs
4. Material and Methods

T4 RNA ligase (10 U/μl)  ThermoFischer Scientific
T7 RNA polymerase (20 U/μl)  ThermoFischer Scientific
Taq DNA polymerase  expressed and purified by Silke Schreiner
TURBO DNase (2 U/μl)  ThermoFischer Scientific
USB® sequenase™ version 2.0 DNA polymerase  Affymetrix

4.1.4 Nucleotides

Deoxynucleoside triphosphate mix  Peqlab
(dNTPs, 100 mM)
Ribonucleoside triphosphates (100 mM)  Roche Applied Science
\([\alpha^{32}P]\)-ATP (3000 Ci/mmol; SCP-207)  Hartmann Analytic
\([\alpha^{32}P]\)-CTP (3000 Ci/mmol; SCP-209)  Hartmann Analytic
\([\alpha^{32}P]\)-dCTP (3000 Ci/mmol; SCP-205)  Hartmann Analytic
\([\gamma^{32}P]\)-ATP (800 Ci/mmol; SCP-801)  Hartmann Analytic

4.1.5 Markers

DIG-labeled DNA molecular weight marker VII  Roche Applied Science
Gene ruler DNA ladder mix  ThermoFischer Scientific
Gene ruler low range DNA ladder mix  ThermoFischer Scientific
pBR322 (Bolivar et al., 1977); *Hpall* digested  kindly provided by Michael R. Green (University of Massachusetts Medical School, Worcester, USA)
pcDNA3.1(+)  ThermoFischer Scientific
peqGOLD protein-marker IV  Peqlab

4.1.6 Antibodies

Anti-GFP antibody  Roche Applied Science
Anti-protein A antibody produced in rabbit  Sigma-Aldrich
Anti-protein C antibody ab18591 (rabbit)  Abcam
Anti-rabbit IgG; peroxidase antibody (goat)  Sigma-Aldrich
Goat anti-rabbit IgG; secondary antibody; Alexa fluor® 488 conjugate  ThermoFischer Scientific
Goat anti-rabbit IgG; secondary antibody; Alexa fluor® 594 conjugate  ThermoFischer Scientific
4.1.7 Plasmids

pBR322 (Bolivar et al., 1977) kindly provided by Michael R. Green (University of Massachusetts Medical School, Worcester, USA)

pcDNA3.1(+) ThermoFischer Scientific

pc-PTP-NEO (Schimanski et al., 2005b) kindly provided by Arthur Günzl (University of Connecticut Health Center, Connecticut, USA)

pFastBac™ HT B ThermoFischer Scientific

pGEX-6P-2 GE Healthcare

pLEW100 (Shi et al., 2000) kindly provided by Arthur Günzl (University of Connecticut Health Center, Connecticut, USA)

pJM326 (Schrader et al., 1998) kindly provided by Arthur Günzl (University of Connecticut Health Center, Connecticut, USA)

4.1.8 Bacterial and eukaryotic cells

DRBD4-PTP/YFP all DRBD4 cell lines were kindly provided by Mark Carrington (University of Cambridge, Department of Biochemistry, Cambridge, UK)

DRBD4-RNAi (Tb427/pSPR2/p3217/p4379)

DRBD4-RNAi (Tb427/p4106/p3217/p4379)

MAX efficiency® DH10Bac™ competent cells ThermoFischer Scientific

One shot® BL21 Star™ (DE3) chemically competent E. coli ThermoFischer Scientific

Sf9 insect cells in Sf-900™ II SFM ThermoFischer Scientific

Single step (KRX) competent cells Promega

Tb427 29.13 (Wirtz et al., 1999) kindly provided by Ulrich Göringer (TU Darmstadt, Molecular Genetics, Darmstadt, GER)
4. Material and Methods

4.1.9  *T. brucei* SDM-79 medium

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ciprofloxacin</td>
<td>1.5 mg/l</td>
</tr>
<tr>
<td>Fetal bovine serum (FBS)</td>
<td>10%</td>
</tr>
<tr>
<td>Glucose</td>
<td>1 g/l</td>
</tr>
<tr>
<td>Hemin (2.5 mg/ml in 0.05 M NaOH)</td>
<td>5 mg/l</td>
</tr>
<tr>
<td>L-proline</td>
<td>600 mg/l</td>
</tr>
<tr>
<td>L-threonine</td>
<td>350 mg/l</td>
</tr>
<tr>
<td>Penicillin G</td>
<td>60 mg/l</td>
</tr>
<tr>
<td>SDM-79 CCGGPPTA</td>
<td>One pack for 5 l medium</td>
</tr>
<tr>
<td>Sodium pyruvate</td>
<td>100 mg/l</td>
</tr>
</tbody>
</table>

The pH is adjusted with NaOH to 7.4 and the medium is filter sterilized with a Sartolab-P20 plus 0.20 µm filter (Sartorius).

4.1.10 Laboratory equipment

<table>
<thead>
<tr>
<th>Equipment</th>
<th>Manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>ÄKTA purifier 10</td>
<td>GE Healthcare</td>
</tr>
<tr>
<td>Agilent 2100 Bioanalyzer</td>
<td>Agilent Technologies</td>
</tr>
<tr>
<td>Axioskop 20 fluorescence microscope</td>
<td>Zeiss</td>
</tr>
<tr>
<td>BioPhotometer</td>
<td>Eppendorf</td>
</tr>
<tr>
<td>BLX-254 UV-Crosslinker</td>
<td>Bio-Link</td>
</tr>
<tr>
<td>Cooling-Thermomixer MKR 13</td>
<td>HLC</td>
</tr>
<tr>
<td>Curix60</td>
<td>AGFA</td>
</tr>
<tr>
<td>Electro cell manipulator 600</td>
<td>BTX</td>
</tr>
<tr>
<td>Foto/PrepI</td>
<td>Fotodyne</td>
</tr>
<tr>
<td>G:BOX gel documentation system</td>
<td>Syngene</td>
</tr>
<tr>
<td>GeneAmp PCR system 9700</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>LabChip XT/XTe</td>
<td>PerkinElmer</td>
</tr>
<tr>
<td>Mastercycler ep realplex² S</td>
<td>Eppendorf</td>
</tr>
<tr>
<td>Mini-PROTEAN tetra electrophoresis system</td>
<td>Bio-Rad</td>
</tr>
<tr>
<td>NanoDrop 1000 spectrophotometer</td>
<td>ThermoFischer Scientific</td>
</tr>
<tr>
<td>NuPAGE electrophoresis and blotting system</td>
<td>ThermoFischer Scientific</td>
</tr>
<tr>
<td>PerfectBlue gelsystem mini M</td>
<td>Peqlab</td>
</tr>
<tr>
<td>Qubit® 2.0 fluorometer</td>
<td>ThermoFischer Scientific</td>
</tr>
<tr>
<td>Safety stand 630 A</td>
<td>BTX</td>
</tr>
<tr>
<td>Slab gel dryer SE1160</td>
<td>Hoefer Scientific Instruments</td>
</tr>
<tr>
<td>SpeedVac® plus SC100A</td>
<td>Savant</td>
</tr>
</tbody>
</table>
4. Material and Methods

Sub-Cell GT agarose gel system  Bio-Rad
SureCycler 8000  Agilent Technologies
Trans-Blot® turbo™ transfer system  Bio-Rad
Tri-Carb 1600TR  PACKARD
Typhoon™ FLA 9500  GE Healthcare
Veriti thermal cycler  Applied Biosystems
Wide mini sub cell 41E/1261  Bio-Rad

4.1.11 DNA and RNA oligonucleotides (in 5´ to 3´ direction)

Cloning of *T. brucei* pC-PTP-NEO constructs (Sigma-Aldrich)

PAP-PTP Tb927.3.3160 fw.  ATAAAGGGCGCCGTTTTGTGAGGTCTTCCAT
PAP-PTP Tb927.3.3160 rv.  ATAAAGGGCGCCGCGCTCTCAGGTCTTGGAT
PAP-PTP Tb927.7.3780 fw.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT
PAP-PTP Tb927.7.3780 rv.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT
CPSF160-PTP fw.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT
CPSF160-PTP rv.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT
CPSF30-PTP fw.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT
CPSF30-PTP rv.  GATTAGGGCCGCGCCGAGTCCCAGCGCTT

Cloning of *T. brucei* pLEW100 RNAi constructs (Sigma-Aldrich)

PAP Tb927.3.3160 RNAi fw.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
PAP Tb927.3.3160 RNAi rv.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
PAP Tb927.7.3780 RNAi fw.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
PAP Tb927.7.3780 RNAi rv.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF160 RNAi fw.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF160 RNAi rv.  GCATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF100 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF100 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF73 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF73 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF30 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CPSF30 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
Fip1 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
Fip1 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CstF50 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CstF50 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CstF64 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CstF64 RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
Symplekin RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
Symplekin RNAi rv.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
CFIm25 RNAi fw.  ATATAAAGCTTACGGCGTGAGTCTTCCAT
4. Material and Methods

CFIm25 RNAi rv.
ATATTCTAGAGAGATCGAACAATGGAGCGG
CFIm RNAi fw.
ATATAAGCTTACCGGCTGTTGAACATGAGGGGCTTG
CFIm RNAi rv.
ATATTCTAGACCGGGGAGAGAAACGAAAAG
Tb927.11.13860 RNAi fw.
ATATAAGCTTACGCGTGGTAACTGACATGGGGCTTG
Tb927.11.13860 RNAi rv.
ATATTCTAGACCGGGGAGAGAAACGAAAAG
Tb927.8.4480 RNAi fw.
ATATAAGCTTACCGGCTGTTGAACATGAGGGGCTTG
Tb927.8.4480 RNAi rv.
ATATTCTAGACCGGGGAGAGAAACGAAAAG

Cloning of baculovirus pFastBac™HT B expression constructs (Sigma-Aldrich)

PAP Tb927.3.3160 fw.
CATAGGATCCAGTGGATGTATGGCCCAAC
PAP Tb927.3.3160 rv.
CATATCTAGATTACTTCAGGTCTTGTAACAGGT
PAP Tb927.7.3780 fw.
CATAGGATCCATCAGCAACAGACAAGTCGTTC
PAP Tb927.7.3780 rv.
CATATCTAGATAAAAAACCGAGTCCAGCAGC

Cloning of bacterial pGEX-6P-2 expression constructs (Sigma-Aldrich)

GST-CPSF30 fw.
CCTGGGGATCCCTTTTACTGACAACGCTGCCCAAC
GST-CPSF30 rv.
CCGGGAATTCTTACTCGCTTCTCCGGTACATC
GST-DRBD4 fw.
ATATGTCGACTTACACCTGCTCTTTTGAATACATT
GST-DRBD4 rv.
ATATGCGGCGCTCACTCATCTCTGTTCGTCTTG

Sequencing primer (Sigma-Aldrich)

pc-PTP-NEO seq rv.
AATTCTCGCTAGCAGTAGTTGG
pGEX-6P-2 seq fw.
AAGCCACGTTGGTGTTTG
pGEX-6P-2 seq rv.
CCGGGAGCTGATGTCAGAGG
pFastBac™HT B seq.
AAATGATACCATCTCGC

mRNA detection primer (Sigma-Aldrich)

mRNA CPSF160 fw.
AAGTCCTCGAGTGGTCTCTC
mRNA CPSF160 rv.
AAACATGCCTGCTGTCAGTC
mRNA CPSF100 fw.
GGGGATCTGAGTGGGAGG
mRNA CPSF100 rv.
CCGCAACTATATGGAGGTCG
mRNA CPSF73 fw.
GAGACACCAAATCCAGGAG
mRNA CPSF73 rv.
ATTCCTAACTACCCAGGTG
mRNA CPSF30 fw.
CCCGCGTGTGTGTTTTTACC
mRNA CPSF30 rv.
CAGTGGCATACCCGGACAG
mRNA Fip1 fw.
CTGCACGCTTTTCTGAG
mRNA Fip1 rv.
GGTGTGTTGTTGTGTTGTG
mRNA CstF50 fw.
GGACTGGCACCCTTATCACT
mRNA CstF50 rv.
CTTTACACTCTCAGACTG
mRNA CstF64 fw.
AACACCCAGCCGATGTAT
mRNA CstF64 rv.
TTTCAAGGAAGGGCGCAATCT
mRNA Symplekin fw.
CCGGATATGAGGATGGTCCGCT
4. Material and Methods

mRNA Symplekin rv. TCGAGAAGCAACACTGGGCAT
mRNA CFIm25 fw. GAAGTGGGTGTTGATGTCGG
mRNA CFIm25 rv. GAGATCGAACAATGGAGCGG
mRNA CFIm fw. CGTTCACTTCATCCACGCTT
mRNA CFIm rv. AAGGTGGCAGTTTTATCCCGT
mRNA Tb927.11.13860 fw. ACAACTTAGTGGGCTCGTCA
mRNA Tb927.11.13860 rv. CAACGATGACCACGACACAA
mRNA Tb927.8.4480 fw. CTGCAGATGTCGCCAGTTTT
mRNA Tb927.8.4480 rv. TAATCCATGCCTCTTCGCCA

Polyadenylation assay (Sigma-Aldrich)
Poly(T) with G CGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG

Trans splicing assay (Sigma-Aldrich)
PPIase exon rv. CGTTGCGACCACCTTCTGCA
PRP8 exon rv. GATGACCACCTTCTCAGTAC
SL RNA 6-28 fw. ACGCTATTATTAGAAACAGTTTCT
SL RNA rv. CAGGAACCAACACGACAAATG

PCR primer for loading controls (Sigma-Aldrich)
7SL RNA fw. GTTGCGTTGACTTGTTGTTC
7SL RNA rv. TC GGTCGGTCTCTGGCAAC
U3 RNA fw. CCGTACTCTGAACAGAATCGTTTT
U3 RNA rv. CCGTTCATCGAACAGCTCTC'

SELEX annealing oligonucleotides (Sigma-Aldrich); N: random nucleotide
SLX-N20 GCGTCTCGAGCGTAGTTANNNNNNNNNNNNNNNNNNAG
TCGGCATCTTGGTACCCTATAGTGAGTCGTATTA
T7 fw. TAATACGACTCACTATAGGG

SELEX RT-PCR primer (Sigma-Aldrich)
SLX-RT GCGTCTCGAGCGTAGTTA
SLX-T7 fw. TAATACGACTCACTATAGGGTACCAAGATGCCGACT

SELEX RT barcoding primer (Sigma-Aldrich); N: random nucleotide; experimental barcode
SLX_R1 CAAGCAGAAGACGGCATACGAGATCGGTCACAGCTGCTCCGATCCCTGN
NCCAAANNNGCGTCTCGAGCGTAGTTA
SLX_R2 CAAGCAGAAGACGGCATACGAGATCGGTCACAGCTGCTCCGATCCCTGN
NCCAAANNNGCGTCTCGAGCGTAGTTA
SLX_R3 CAAGCAGAAGACGGCATACGAGATCGGTCACAGCTGCTCCGATCCCTGN
NGTTCNNGCGTCTCGAGCGTAGTTA
### 4. Material and Methods

| SLX_R4       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NGGCTNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R5       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTTGCAANNNGCGTCTCGAGCGGTAGTTA |
| SLX_R6       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NAGCTNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R7       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTTCTNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R8       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTACNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R9       | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTCTTNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R10      | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTAACNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R11      | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NTCTTNNNGCGTCTCGAGCGGTAGTTA |
| SLX_R12      | CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN NACTTNNNGCGTCTCGAGCGGTAGTTA |

SELEX library amplification primer (Sigma-Aldrich); N: random nucleotide
SLX-Sol-5×N fw. AATGATACGGCGACGTCCGAGATCTACACTCTTTCCCTACAC GACGCTCTTCAGTGCTNWNNGGTACCAAGATGCCGACT
SLX_Sol rv. CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTCTGN

Oligonucleotides for T7 *in vitro* transcription (Sigma-Aldrich)
- DRBD4 motif_1_s (T7 fw.) TAATACGACTCACTATAGGG
- DRBD4 motif_1_as TTGGGTCTATTGGGGTTGCTTATTGGTCTTCTATAGTGAG TCGTATTA
- DRBD4 motif_2_s TAATACGACTCACTATAGGGAAATGAAATGACGTCGAG
- DRBD4 motif_2_as CTCGACGCTTCATTTCCCTCTATAGTGAGTCGTATTA
- DRBD4 motif_3_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_3_as CTCGACGTCATTCTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_4_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_4_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_5_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_5_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_6_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_6_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_7_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_7_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_8_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_8_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
- DRBD4 motif_9_s TAATACGACTCACTATAGGAAGAAATGAAAGACGTCGAG
- DRBD4 motif_9_as CTCGACGCTTCATTTCCCTATAGTGAGTCGTATTA
4. Material and Methods

**DRBD4 motif_9_s**
TAATACGACTCACTATAGGGAAAAAAAAAAACGTCGAG

**DRBD4 motif_9_as**
CTCGACGTTTTTTTTTTTTTTTCCCTATAGTGAGTCGATTA

**DRBD4 motif_10_s**
TAATACGACTCACTATAGGGAAATTAATTA

**DRBD4 motif_10_as**
CTCGACGTTTTTTTTTTTTTTCTATAGTGAGTCGATTA

**Tb927.1.2340_PPT_s**
GAGGGGGTGAGGAGAGGAGGGAACCTATAGTGAGTCGATTA

**Tb927.1.2340_PPT_as**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.11.3630_PPT_s_1**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.11.3630_PPT_as_1**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.11.3630_PPT_s_2**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.11.3630_PPT_as_2**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.6.510_PPT_s (T7 fw.)**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**Tb927.6.510_PPT_as**
AAAGAAGAGAAGGAAGGGAAGAGGGAACCTATAGTGAGTCGATTA

**iCLIP RNA oligonucleotide** (Dharmacon); P: phosphate

**3' linker**
P-UGAGAUCGGAAGAGCCGUUCAG-Puromycin

**iCLIP RT-primer** (Sigma-Aldrich); P: phosphate; N: random nucleotide; experimental barcode

**iCLIP-RT4**
P-NNAGGTNNNAGATCGGAAGAGGATCGTGGATCCTGACCC

**iCLIP-RT8**
P-NNCATTNNNAGATCGGAAGAGGATCGTGGATCCTGACCC

**iCLIP-RT9**
P-NNGCCANNNAGATCGGAAGAGGATCGTGGATCCTGACCC

**iCLIP-RT10**
P-NNGACCNNNAGATCGGAAGAGGATCGTGGATCCTGACCC

**iCLIP oligonucleotide for linearization of circular ssDNA** (MWG Eurofins)

**Cut_oligo**
GTTCAGGATCCACGACGCTCTTCAAAA

**iCLIP PCR-primer compatible with high-throughput sequencing** (MWG Eurofins)

**Solexa P5**
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACAC

**Solexa P3**
CAAGCAGAAGACGGCATACGAGATCGGTCTCCGACCTTCTGC

**PCR-primer for Northern blot probes** (Sigma-Aldrich)

**Probe Tb927.8.760 fw.**
CCGAAGATACCCGAAGACCA

**Probe Tb927.8.760 rv.**
AACATCCGAACATGAGGGGA
4. Material and Methods

<table>
<thead>
<tr>
<th>Probe</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.8.770 fw.</td>
<td>GATAACGTGTCGCTGACCAC</td>
</tr>
<tr>
<td>Tb927.8.770 rv.</td>
<td>TGCCTCCTCGTCAAAGTGAT</td>
</tr>
<tr>
<td>Tb927.11.5440 fw.</td>
<td>TGGTCTGAAAGTGCTGACC</td>
</tr>
<tr>
<td>Tb927.11.5440 rv.</td>
<td>CGTGAAGTTGCTACC</td>
</tr>
<tr>
<td>Tb927.6.1470 fw.</td>
<td>CAAGGTGAGTGAGGAAGGA</td>
</tr>
<tr>
<td>Tb927.6.1470 rv.</td>
<td>TGCCTCTTTTGTCTACCATG</td>
</tr>
<tr>
<td>Tb927.2.6000 fw.</td>
<td>CCAATCAGGACCCGCTTACA</td>
</tr>
<tr>
<td>Tb927.2.6000 rv.</td>
<td>CTGTCGCTATTACATGCA</td>
</tr>
<tr>
<td>18S rRNA fw.</td>
<td>CAGCGGATCACCGCTTACA</td>
</tr>
<tr>
<td>18S rRNA rv.</td>
<td>CTGTCGCTATTACATGCA</td>
</tr>
</tbody>
</table>
4.2 General techniques

4.2.1 Guideline
All methods were adapted from laboratory protocols or the cited literature. Enzymes, reagents and kits were, if not otherwise stated, used according to the manufactures instructions. If not otherwise mentioned, all reactions and incubation steps were performed at room temperature.

4.2.2 Working with RNA
All steps were performed with certified RNase-free material and all buffers were prepared with RNase-free water. This was prepared by addition of 1:1,000 volumes of dimethyl pyrocarbonate (DMPC) to deionized water, stirring for 1 h and two times autoclaving (120°C, 20 min, 2 bar). In the following, RNase-free water is termed “DMPC-water”. If not otherwise mentioned, samples were always kept on ice.

4.2.3 Isolation of nucleic acids from *T. brucei* cells
For the isolation of genomic DNA or total RNA from procyclic *T. brucei* cells, the DNAzol® or TRIzol® reagent (both ThermoFischer Scientific) was used, respectively. If high-quality RNA was required, e.g. for RT-PCR, the RNA precipitation step was skipped, the RNA solution was mixed 1:1 with 70% ethanol and further purified using RNeasy columns (QIAGEN).

4.2.4 Purification of nucleic acids
Phenol/chloroform extraction was used to purify nucleic acids form enzymatic reactions. The nucleic acid/protein solution was dissolved with DMPC-water >300 µl and an equal volume of RNA-grade phenol/chloroform/isoamylacohol (25:24:1) was added. The mixture was vigorously mixed for 30 sec and centrifuged (4°C, 5 min, 25,000 × g) for phase separation. The upper, aqueous phase was transferred in a new reaction tube and nucleic acids were precipitated (see 4.2.5). If the yield was critical, PhaseLock heavy gel tubes (5 Prime) were used.

4.2.5 Precipitation of nucleic acids
To concentrate or desalt nucleic acids, ethanol or isopropanol precipitation was used. The nucleic acid solution was mixed with 1:10 volumes of 3M sodium-acetate (pH 5.2) and 10-20 µg glycogen (PeqLab) or 7.5-15 µg GlycoBlue (ThermoFischer Scientific). To precipitate nucleic acids, 2.5 volumes of 100% ethanol or 0.7-0.8 volumes of isopropanol were added, followed by incubation (4°C, -20°C or -80°C for 30 min to 16h) and centrifugation (4°C, 30 min, 25,000 × g). The pellet was washed with 70% ethanol and centrifuged for 5 min using the same conditions.
Precipitated nucleic acids were dissolved in deionized water or in case of RNA in DMPC-water. RNA-pellets were for subsequent denaturing polyacrylamide/urea gel electrophoresis directly dissolved in 1× formamide loading buffer (0.5× TBE, 45% formamide, 0.025% bromphenol blue, 0.025% xylene cyanol).

4.2.6 Photometric determination of nucleic acid concentration
To determine the nucleic acid concentration, the NanoDrop 1000 spectrophotometer (ThermoFischer Scientific) or BioPhotometer (Eppendorf) was used. In case of iCLIP and SELEX-seq libraries, the nucleic acid concentration was measured using the Qubit® 2.0 fluorometer (ThermoFischer Scientific) and the Qubit® dsDNA HS assay kit (ThermoFischer Scientific).

4.2.7 Databases, web-based bioinformatics and programs
As a general database for T. brucei specific genome-data “TriTrypDB” was used (Aslett et al., 2010). Pairwise nucleotide and protein sequence alignments were performed by the “EMBOSS Needle” and multiple sequence alignments by the “Clustal Omega” program. Pattern and profile search were done by “InterProScan” (Jones et al., 2014). Protein structure prediction was carried out via the “GeneSilico metaserver” (Kurowski & Bujnicki, 2003). PCR primer were designed using “Primer3” (Rozen & Skaletsky, 2000) and melting temperatures were calculated by “Oligo Calc” (Kibbe, 2007). In addition, useful tools like “Reverse complement” from “The sequence manipulation suite” (Stothard, 2000), “NEBcutter V2.0” as well as “Double digest finder” provided by New England Biolabs, "LIGATION CALCULATOR" and “BLAST®” (Altschul et al., 1990) were used. For planning of molecular cloning strategies “Vector NTI 10.3.1” (ThermoFischer Scientific) and for analysis of genome-wide data the “Integrative genomics viewer” were utilized (Robinson et al., 2011, Thorvaldsdottir et al., 2013). Mass spectrometry data were evaluated with “Scaffold version 4.0.4” (Proteome Software) and relative band intensities were measured using “ImageJ” (Schneider et al., 2012). Literature search was performed using “PubMed”.

Clustal Omega http://www.ebi.ac.uk/Tools/msa/clustalo/
EMBOSS Needle http://www.ebi.ac.uk/Tools/psa/emboss_needle/
GeneSilico metaserver https://www.genesilico.pl/meta2/
4. Material and Methods

InterProScan  http://www.ebi.ac.uk/interpro/search/sequence-search
LIGATION CALCULATOR  http://www.insilico.uni-duesseldorf.de/Lig_Input.html
NEBcutter V2.0  http://nc2.neb.com/NEBcutter2/
Oligo Calc  http://biotools.nubic.northwestern.edu/OligoCalc.html
Primer3  http://primer3.ut.ee/
The sequence manipulation suite  http://www.bioinformatics.org/sms/index.html
TriTrypDB  http://tritrypdb.org/tritrypdb/
4.3 Cell culture of procyclic *T. brucei*

4.3.1 Cultivation of procyclic *T. brucei* cells

Cell culture of eukaryotic cells was carried out under sterile conditions, using a clean bench in a designated cell culture laboratory. Commercially available flasks and pipettes were used. Procyclic *T. brucei* cells were grown for two to three days at 28°C in SDM-79 medium (see 4.1.9) after dilution to $2 \times 10^6$ cells/ml. Glycerol stocks were prepared in SDM-79 (containing 10% glycerol) and stored at -80°C or in liquid nitrogen. According to the type of cell line antibiotics were added to the media (Table 4.1).

Table 4.1 Antibiotics for *T. brucei* cell culture.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Antibiotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTP/YFP-tag</td>
<td>Geniticin [40 µg/ml]</td>
</tr>
<tr>
<td>Tb427 29.13</td>
<td>Geniticin [15 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Hygromycin [50 µg/ml]</td>
</tr>
<tr>
<td>Tb427 29.13 + pLEW100</td>
<td>Geniticin [15 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Hygromycin [50 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Phleomycin [2.5 µg/ml]</td>
</tr>
<tr>
<td>DRBD4-RNAi (pSPR2/p3217/p4379)</td>
<td>Blasticidin [10 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Geniticin [15 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Puromycin [1 µg/ml]</td>
</tr>
<tr>
<td>DRBD4-RNAi (p4106/p3217/p4379)</td>
<td>Blasticidin [10 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Geniticin [15 µg/ml]</td>
</tr>
<tr>
<td></td>
<td>Puromycin [1 µg/ml]</td>
</tr>
</tbody>
</table>

4.3.2 Transfection of procyclic *T. brucei* cells

Procyclic *T. brucei* cells (see 4.3.1) were transfected with 10 µg linearized plasmid (in 20 µl sterile water) by electroporation. The cells ($8 \times 10^7$ to $1 \times 10^8$) were suspended in 380 µl Cytomix pH 7.6 (120 mM KCl, 0.15 mM CaCl$_2$, 10 mM K$_2$HPO$_4$/KH$_2$PO$_4$ pH 7.6, 25 mM HEPES pH 7.6, 5 mM MgCl$_2$, 2 mM EGTA pH 7.6, 0.5% glucose, 100 µg/ml BSA, 1 mM hypoxanthine), mixed with the plasmid, transferred in a Gene pulser®/MicroPulser™ electroporation cuvette 0.2 cm (Bio-Rad) and electroporated (1500 V, 25 µF, 24 Ω) using a Electro cell manipulator 600 (BTX). Afterwards 10 ml fresh SDM-79 medium was added and the cells were grown for 8 h before the selective antibiotic was added (see 4.3.1). Finally, the cells were spread (200 µl each) in 96-well plates for growing to obtain clonal cell lines.
4.3.3 Generation of clonal cell lines

To generate a stable RNAi cell line, the linearized pLEW100 construct (see 4.9.1) was transfected into Tb427 29.13 cells and integrated into an rDNA spacer. The plasmid (10 µg) was linearized using SacI (see 4.6.2), isopropanol precipitated (see 4.2.5, no addition of glycogen or sodium-acetate) and transfected (see 4.3.2). To generate a C-terminal PTP-tagged cell line, the pC-PTP-NEO construct (see 4.9.2) was transfected (see 4.3.2) in T. brucei 427 (WT) cells and integrated in the endogenous locus. The constructs were linearized in the insert using XhoI (Tb927.7.3780, Tb927.3.3160, CPSF160) or BsmI (CPSF30), precipitated and transfected as described above.

4.3.4 Reverse knockdown using procyclic T. brucei RNAi cell lines

For a specific, RNAi-mediated knockdown of a single gene (see 4.3.3), 2×10^6 cells/ml were treated for 24 h with 1 mg/ml doxycyclin (Sigma-Aldrich, in 70% ethanol). Afterwards the cells were counted, diluted to 2×10^6 cells/ml and treated again with doxycyclin. As control 70% ethanol was used. The cells were harvested for a large, preparative knockdown after day one, two and three after RNAi induction. The knockdown was extended for up to five days for growth curve experiments. RNAi for RNA-seq was induced in the cell line Tb427/pSPR2/p3217/p4379 by the addition of 0.2 mg/ml doxycycline and cells were diluted every day to 2×10^4 cell/ml. For Northern blotting (cell line: Tb427/p4106/p3217/p4379) RNAi was induced by the addition of 0.2 mg/ml doxycycline at a cell density of 2×10^4/ml or 1 mg/ml doxycycline at a cell density of 2×10^5/ml.
4. Material and Methods

4.4 Reverse transcription (RT) and polymerase chain reaction (PCR)

4.4.1 Reverse transcription (RT)

RNA was reverse transcribed for PCR-based analysis using the qScript™ cDNA synthesis kit (Quanta Biosciences) containing oligo(dT) and random hexamer primer. For specific RT the Superscript III reverse transcriptase (ThermoFischer Scientific) or the qScript™ flex cDNA kit (Quanta Biosciences) was used. Total RNA was isolated as described in 4.2.3.

4.4.2 Polymerase chain reaction (PCR)

Taq polymerase and a 10× Taq PCR buffer (100 mM Tris-Cl pH 9.0, 500 mM NaCl, 15 mM MgCl₂, 1% Triton X-100) were used for standard PCR. Preparative PCR for cloning of protein-coding genes was performed using Phusion® high-fidelity DNA polymerase (New England Biolabs). As template either cDNA (see 4.4.1) or genomic DNA (see 4.2.3) was used.

Standard 25 µl PCR and temperature profile

<table>
<thead>
<tr>
<th>Volume</th>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5 µl</td>
<td>10× Taq PCR buffer</td>
<td>1× (1.5 mM MgCl₂)</td>
</tr>
<tr>
<td>0.5 µl</td>
<td>Taq DNA polymerase</td>
<td>-</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>MgCl₂ (25 mM)</td>
<td>1.0 mM</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>dNTPs (10 mM each)</td>
<td>0.4 mM</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>forward primer (10 µM)</td>
<td>0.4 µM</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>reverse primer (10 µM)</td>
<td>0.4 µM</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>DNA template</td>
<td>-</td>
</tr>
<tr>
<td>17.0 µl</td>
<td>deionized water</td>
<td></td>
</tr>
<tr>
<td>25.0 µl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- initial denaturation: 94°C, 2 min
- denaturation: 94°C, 10 sec
- annealing: 52-62°C, 20 sec
- elongation: 72°C, 30-120 sec
- final elongation: 72°C, 7 min
- storage: 4°C, hold

If required, the PCR reaction was scaled up to 2× (50 µl) or 4× (100 µl) and the magnesium chloride concentration was altered for optimization. The annealing temperature of each primer pair was empirically determined or by using “Oligo Calc” (see 4.2.7). For primer design “Primer3” (see 4.2.7) was used (with the exception of open reading frame primer for protein expression constructs). The elongation time was set depending on the length of the PCR product by
assuming an average elongation of 1 kb per min. The cycle number was determined empirically. PCR reactions were analyzed by gel electrophoresis and ethidium bromide staining (see 4.5.1). Preparative PCR products were gel purified (see 4.5.2). Primer sequences are summarized in 4.1.11.

### 4.4.3 Quantitative RT-PCR (RT-qPCR)

Since PCR described in 4.4.2 allows only semiquantitative measurements, quantitative RT-PCR (RT-qPCR) was used to determine RNA levels precisely. Total RNA was isolated (see 4.2.3) and converted into cDNA according to 4.4.1. All experiments were performed as biological replicates and technical triplicates. The Quanta PerfeCTa® SYBR® green FastMix® (Quanta Biosciences) and Mastercycler ep realplex² S (Eppendorf) were used. qPCR data was analyzed using the provided software and quantified according to an established strategy (Pfaffl, 2001). Primer amplification efficiencies were determined by a dilution series of cDNA (see 4.4.1) from *T. brucei* 427 (WT) cells. C_T values were plotted against the logarithm of the dilution factor to obtain the slope. Finally, the amplification efficiency was calculated according to the formula: 

\[ E = 10^{\frac{1}{\text{slope}}} \]

Relative expression levels were calculated using the following formula:

\[
\text{relative expression} = \frac{(E_{\text{target RNA}})^{\Delta C_T \text{target RNA}(\text{control-knockdown})}}{(E_{\text{reference RNA}})^{\Delta C_T \text{reference RNA}(\text{control-knockdown})}}
\]

E is the amplification efficiency and ΔC_T is the difference between the control and knockdown sample.
4. Material and Methods

4.5 Gel electrophoresis-based analysis and detection methods

For gel electrophoresis of nucleic acids the Tris-Borate-EDTA (TBE) buffer system was used. The 10× TBE buffer stock solution (890 mM Tris-Cl pH 8.0, 890 mM boric acid, 20 mM EDTA) was diluted as indicated.

4.5.1 Gel electrophoresis of DNA

Agarose gel electrophoresis was used as standard method for the analysis of DNA. The agarose concentration was chosen depending on the fragment length to obtain optimal resolution (Table 4.2).

Table 4.2 Agarose concentration in DNA gels.
DNA fragment-dependent agarose concentration used for gel electrophoresis including running behavior of tracking dyes

<table>
<thead>
<tr>
<th>Fragment size</th>
<th>Agarose concentration (w/v)</th>
<th>Bromphenol blue</th>
<th>Xylen cyanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5-7 kb</td>
<td>1.0</td>
<td>300 bp</td>
<td>3.0 kb</td>
</tr>
<tr>
<td>0.4-6 kb</td>
<td>1.2</td>
<td>200 bp</td>
<td>1.5 kb</td>
</tr>
<tr>
<td>0.2-3 kb</td>
<td>1.5</td>
<td>120 bp</td>
<td>1.0 kb</td>
</tr>
<tr>
<td>0.1-2 kb</td>
<td>2.0</td>
<td>&lt;100 bp</td>
<td>0.8 kb</td>
</tr>
</tbody>
</table>

Agarose was dissolved in 0.5× TBE by boiling in a microwave, supplemented with ethidium bromide (final concentration 0.5 µg/ml) and stored at 50°C. Gels were poured and run in horizontal gel systems (PerfectBlue gelsystem mini M [Peqlab] or Sub-Cell GT [Bio-Rad]). DNA samples were diluted 1:6 with a 6× DNA loading dye (1× TBE, 20% glycerol, 0,025% xylen cyanol, 0,025% bromphenol blue) and run for 30-60 min at 120 V in 0.5× TBE. Gels were analyzed and documented under UV-irradiation using the G:BOX gel documentation system (Syngene). Gene ruler DNA ladder mix or Gene ruler low range DNA ladder mix (both ThermoFischer Scientific) was used as marker.

Polyacrylamide gel electrophoresis (Mini-PROTEAN tetra electrophoresis system [Bio-Rad]) was used for high resolution analysis of very short DNA fragments (<200 bp, e.g. iCLIP library preparation PCR). A 6% solution of acrylamide/bisacrylamide (19:1) was prepared in 1× TBE, and polymerization was initiated by adding 1:125 volumes of 10% APS and 1:1,250 volumes of TEMED. Gels were run for 25 min at 200 V (1× TBE), stained for 10 min with ethidium bromide (5 µg/ml in 1× TBE) and destained for 2 min in deionized water. Analysis was carried out as
described above. As marker, Gene ruler low range DNA ladder (ThermoFischer Scientific) was used.

4.5.2 Extraction of DNA from agarose gels

Gel purification was used for the isolation of a single DNA species from a complex mixture. Agarose gel electrophoresis was performed with an agarose concentration <2% according to 4.5.1, followed by purification using the QIAGen QIAquick gel extraction kit (QIAGEN).

4.5.3 Denaturing polyacrylamide/urea gel electrophoresis of RNA

For the analysis of RNA, the samples were diluted 1:2 in 2× formamide loading buffer (1× TBE, 90% formamide, 0.05% bromphenol blue, 0.05% xylene cyanol) or as pellets directly dissolved in 1× formamide loading buffer after precipitation (see 4.2.5). RNA was denatured for 5 min at 95°C and put on ice for additional 3 min. The samples were loaded onto denaturing TBE polyacrylamide/urea gels for electrophoresis, using either the Mini-PROTEAN tetra electrophoresis system (Bio-Rad) or a custom gel system (21 cm × 18 cm × 0.5 mm). The gel was prepared using a 10 to 15% solution of acrylamide/bisacrylamide (19:1) in 1× TBE containing 50% urea. Polymerization was initiated by the addition of 1:125 volumes of 10% APS and 1:1,250 volumes of TEMED. The gels were run for 20 min at 23 W prior loading of the sample. Samples were loaded and separated for 40 to 110 min at 23W (0.5× TBE). For unspecific detection of RNA Bio-Rad silver stain (Bio-Rad) was used. As a marker, DIG-labeled DNA molecular weight marker VII (Roche Applied Science) was used. Radioactive labeled samples were detected by autoradiography (see 4.5.4). As marker, pcDNA3.1(+) or pBR322 were HpaII digested, fragments were SAP (ThermoFischer Scientific) treated and radioactive labeled at their 5´ end using T4 PNK (New England Biolabs) and [γ-32P]-ATP (800 Ci/mmol; SCP-801 [Hartmann Analytic]).

4.5.4 Autoradiography

After separation of radioactively labeled RNA samples by gel electrophoresis (see 4.5.3), autoradiography was performed for detection. One glass plate was removed and the gel was covered with plastic wrap. An X-ray film (-80°C) or phosphorimaging screen (room temperature) was exposed in a film cassette for up to several days depending on the signal strength. Phosphorimaging screens were scanned using a Typhoon FLA 9500 (GE Healthcare) and quantified using the provided “ImageQuant TL” program.
4. Material and Methods

4.5.5 Glyoxal RNA gel and Northern blotting

Five µg of total RNA form *T. brucei* (see 4.2.3) was denatured for 40 min at 50°C (in 50% DMSO, 4% glyoxal, 10 mM sodium phosphate, pH 6.9) and separated on a 1.2 % agarose gel (10 mM sodium phosphate, pH 6.9). In the following, RNA was capillary blotted overnight on an Amersham Hybond-N membrane (GE Healthcare) using 20× SSC (150 mM NaCl, 15 mM sodium citrate), crosslinked at 254 nM (120 mJ/cm²), incubated at 80°C for 60 min and pre-hybridized (5× SSC, 5× Denhardt’s solution, 10% SDS, 0.1% Na₄P₂O₇, 0.1 % heparin sodium salt) for 1 h at 65°C. Following overnight (65°C) hybridization with ³²P-dCTP-labeled PCR probes (3000 Ci/mmol; SCP-205 [Hartmann Analytic]) complementary to the ORF of the target gene (Prime-It® RmT random primer labeling kit [Agilent Technologies]), the membrane was wash twice (0.1× SSC, 0.1% SDS) for 30 min at 60°C and quantified by autoradiography using the Typhoon FLA 9500 (GE Healthcare) imaging system and X-ray films (see 4.5.4).

4.5.6 Denaturing polyacrylamide/SDS gel electrophoresis of proteins

For gel electrophoresis of proteins, the discontinuous Laemmli gel/buffer system and Mini-PROTEAN tetra electrophoresis system (Bio-Rad) were used (Laemmli, 1970). A 10-15% acrylamide/bisacrylamide (37.5:1) solution was prepared in separating gel buffer (375 mM Tris-Cl pH 8.8, 0.1% SDS) depending on the protein to be analyzed. Polymerization was initiated by adding 1:125 volumes of 10% APS and 1:1,250 volumes of TEMED. Stacking gels were poured after polymerization of the separating gel, using a 5% acrylamide/bisacrylamide (37.5:1) solution in stacking gel buffer (62.5 mM Tris-Cl pH 6.8, 0.05% SDS). Polymerization was initiated by adding the same ratios of TEMED and APS. Protein samples were mixed 1:1 with 2× SDS sample buffer (125 mM Tris-Cl pH 6.8, 20% glycerol, 4% SDS, 5% β-mercaptoethanol, 0.025 mg/ml bromphenol blue) and boiled for 10 min in a water bath. Electrophoresis was performed for 45-90 min in SDS running buffer (25 mM Tris-Cl, 192 mM glycine, 0.1% SDS) with a current limit of 15 mA in stacking- and 35 mA in separating gels. For unspecific detection of proteins, the PageBlue™ protein staining solution (ThermoFischer Scientific) or InstantBlue-protein stain (Expedeon) was used. After destaining with water, the gels were dried on a Slab gel dryer SE1160 (Hoefer Scientific Instruments). As marker, peqGOLD protein marker IV (Peqlab) was used.

4.5.7 Western blotting

Western blotting was performed for specific detection of proteins in cellular lysates. After gel electrophoresis, (see 4.5.6) the gel, the PVDF western blotting membrane (Roche Applied Science) and the blotting paper (Bio-Rad) were soaked in blotting buffer (25 mM Tris, 190 mM
glycine, 0.025% SDS). The blotting sandwich was assembled from the bottom as following: blotting paper, PVDF membrane, gel, blotting paper. Transfer was performed using the Trans-Blot® turbo™ transfer system (Bio-Rad) and the standard SD (semi-dry) program. The membrane was blocked (5% milk powder in 1× PBS, 0.1% Tween 20) for 1 h or overnight at 4°C. The primary antibody (Table 4.3) was added for 1 h in blocking buffer and after binding, the membrane was washed (1× PBS, 0.1% Tween 20) three times for 10 min. The secondary antibody (Table 4.3) was bound for 1 h in blocking buffer and the membrane was washed as described above. Finally, the membrane was dried and Lumi-Light western blotting substrate (Roche Applied Science) was added for detection using X-ray films (GE Healthcare).

**Table 4.3 Primary and secondary antibodies used for Western blotting.**

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Dilution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-protein A antibody produced in rabbit</td>
<td>1:5,000</td>
</tr>
<tr>
<td>Anti-protein C antibody ab18591 (rabbit)</td>
<td>1:400</td>
</tr>
<tr>
<td>Anti-rabbit IgG; peroxidase antibody (goat)</td>
<td>1:10,000</td>
</tr>
</tbody>
</table>
4.6 Cloning of DNA fragments in *E. coli*

4.6.1 Cultivation of *E. coli*

Bacterial cells were cultured under sterile conditions. All used consumables were certified sterile, autoclaved (120°C, 20 min, 2 bar) or heated (180°C, 2 h). LB-medium (10 g/l NaCl, 10 g/ml trypton, 5 g/ml yeast extract) or LB-agar (Roth) for cultivation was dissolved in deionized water, autoclaved and handled using a Bunsen burner. For selective growing according to the resistance of the transformed plasmid, the following antibiotics were added to the media: Ampicillin [100 µg/ml], Gentamicin [7 µg/ml], Kanamycin [50 µg/ml] and Tetracyclin [10 µg/ml].

4.6.2 Digestion of DNA fragments

PCR products were amplified (see 4.4.2) using primer with 5´ overhangs containing restriction sites. Plasmid DNA and PCR products were digested with New England Biolabs restriction enzymes. For simultaneous double digestion, the optimal buffer conditions were selected using the “Double digest finder” on the New England Biolabs webpage (see 4.2.7).

4.6.3 Ligation of DNA fragments

For cloning of DNA fragments, both the plasmid and insert were digested according to 4.6.2. Linearized plasmid and insert were gel purified (see 4.5.2) and molar ratios were determined by agarose gel electrophoresis (see 4.5.1). A five to seven fold molar excess of the insert was mixed with the plasmid (calculated using the “LIGATION CALCULATOR”; see 4.2.7) and incubated for 1 h or overnight at 16°C for ligation, using the T4 DNA ligase (New England Biolabs).

4.6.4 Transformation of *E. coli* cells

One vial of Single step (KRX) competent cells (Promega) was thawed on ice, mixed with 5 µl ligation reaction or 20 ng plasmid and incubated for 5 min on ice. For One shot® BL21 (DE3) chemically competent *E. coli* cells (ThermoFischer Scientific), the incubation time was prolonged to 20-30 min. Cells were incubated for 2 min on ice after head shock (42°C, 30 sec), mixed with 450 µl LB medium and shacked for 1 h at 37°C. Cells were plated on LB-agar plates, supplemented with antibiotics according to 4.6.1 and grown overnight at 37°C.

4.6.5 Identification and analysis of bacterial clones

Colony PCR was used for screening of bacterial clones. A single colony was picked using a pipet tip and stirred in a 1× PCR reaction (see 4.4.2) with insert specific primer. To ensure
efficient cell lysis, the initial denaturation step in the temperature profile was extended to 5 min. LB-medium (5 ml) containing antibiotics (see 4.6.1) was inoculated with selected clones and incubated overnight at 37°C while shaking. Plasmid DNA was isolated using the QIAprep spin miniprep kit (QIAGEN). Plasmids were analyzed by restriction digestion (see 4.6.2) and gel electrophoresis (see 4.5.1). Plasmids containing an insert with the expected length were sequenced by SeqLab (Sequence Laboratories Göttingen). For large scale-preparation of plasmid DNA, the Qiagen® plasmid plus maxi kit (QIAGEN) was used.
4. Material and Methods

4.7 Protein expression and purification

4.7.1 Expression of GST-fusion proteins in *E. coli*
For the expression of N-terminal GST-fusion proteins in *E. coli*, the recombinant (see 4.9.3) pGEX-6-P2 vector (GE Healthcare) was transformed in One shot® BL21 (DE3) chemically competent *E. coli* cells (ThermoFischer Scientific) according to 4.6.4. The expression culture was inoculated 1:1000 and overexpression (16 °C, shaking, overnight) was induced with 0.1 mM IPTG when OD$_{600}$ 0.4-0.6 was reached. The cells were harvested by centrifugation (4°C, 5 min, 5,000 × g) and stored at -80°C after washing with 1× PBS.

4.7.2 Purification of GST-fusion proteins from *E. coli*
The cell pellet (see 4.7.1) was thawed on ice and resuspended in GST-lysis buffer (10 mM Tris-Cl pH 7.5, 100 mM KCl, 2.5 mM MgCl$_2$, 0.1 mM DTT, cOmplete™ EDTA-free protease inhibitor [Roche Applied Science]) and incubated for 30 min at 4°C on wheel. Afterwards, the cells were sonicated (3×10) on ice and the lysate was supplemented with 0.01% NP-40. Cell debris was removed by two times centrifugation (4°C, 20 min, 20,000 × g) and the cell lysate was finally cleared using a syringe with a 0.45 µm filter (Rotilabo® syringe filter CME [Roth]). Prior addition of cell lysate and protein binding for 30 min on wheel, 80 µl (packed bead volume) pre-blocked (1× PBS, 0.2 mg/ml glycogen, 0.1 mg/ml BSA, 10 mM DTT, 0.05% Tween 20) glutathione sepharose 4B beads (GE Healthcare) were washed 3× with GST-binding buffer (10 mM Tris-Cl pH 7.5, 100 mM KCl, 2.5 mM MgCl$_2$, 0.1 mM DTT, 0.01% NP-40). Beads were washed with 5 ml GST-binding buffer, twice with 5 ml GST-binding buffer 600 (10 mM Tris-Cl pH 7.5, 600 mM KCl, 2.5 mM MgCl$_2$, 0.1 mM DTT, 0.01% NP-40), transferred in a 1.5 ml tube and washed again (3× GST-binding 600 and twice GST-binding buffer). Proteins bound to beads were stored in GST-lysis buffer at 4°C for up to 4 weeks. An aliquot of the slurry was analyzed by SDS-PAGE (see 4.5.6), including a BSA standard to determine the protein concentration and purity of the sample.

4.7.3 Cultivation of Sf9 insect cells
Cell culture of Sf9 insect cells (ThermoFischer Scientific) was carried out under sterile conditions using a clean bench in a designated cell culture laboratory. Commercially available flasks and pipettes were used. The adherent growing Sf9 insect cells were cultured at 28°C in Sf-900™ II SFM medium (supplemented with 10% FBS and 0.5× Antibiotic-antimycotic, all ThermoFischer Scientific). The cells were resuspended in fresh medium and split in a 1:5 or 1:10 dilution after three to four days, when a high confluency was reached.
4.7.4 Expression of recombinant proteins in Sf9 insect cells
Recombinant pFastBac™HT B (ThermoFischer Scientific) vector was cloned according to 4.9.3 and baculovirus for protein expression in Sf9 insect cells was obtained, using the Bac-to-Bac® baculovirus expression system (ThermoFischer Scientific).

4.7.5 Purification of baculovirus-expressed proteins
Sf9 insect cells (see 4.7.4) were harvested by centrifugation (4°C, 5 min, 1,000 × g), washed with 1× PBS and resuspended in a fourfold packed cell volume of lysis buffer (50 mM NaH₂PO₄, 300 mM NaCl, 10 mM imidazole, 1% NP-40, Complete™ EDTA-free protease inhibitor [Roche Applied Science], pH 8.0 using NaOH). The cells were lysed for 10 min on ice and cell debris was removed by centrifugation (4°C, 20 min, 20,000 × g). Recombinant proteins were affinity-purified, using an ÄKTA purifier 10 and 1ml HisTrap FF Crude columns (both GE Healthcare). The column was washed (50 mM NaH₂PO₄, 300 mM NaCl, 10 mM imidazole, pH 8.0 using NaOH) with a tenfold column volume, using a flow rate of 1 ml/min and a maximal pressure of 0.5 MPa. The protein was eluted (50 mM NaH₂PO₄, 300 mM NaCl, 100 mM imidazole, pH 8.0 using NaOH) and further gel-purified on an ÄKTA purifier 10 and a Superdex 200 16/60 column (both GE Healthcare) using storage buffer (20 mM Tris-Cl, pH 8.0, 200 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol) and a flow rate of 1 ml/min and a maximal pressure of 0.3 MPa. For storage of purified protein, the samples were frozen in liquid nitrogen.
4.8 Special techniques

4.8.1 Indirect Immunofluorescence

*Trypanosoma brucei* cells (2×10⁷ to 4×10⁷) expressing PTP-tagged proteins were harvested, washed with 1× PBS and resuspended in 1× PBS. 20 µl of the cell suspension were fixed for 30 min at 4°C with 4% paraformaldehyde in 1× PBS on a coverslip. The cells were washed (1× PBS), permeabilized for 5 min (1% TritonX-100 in 1× PBS), washed again (1× PBS) and blocked (1× PBS, 0.05% Tween 20, 0.02% NaN₃, 1% cold-water fish gelatin [Sigma Aldrich]). Afterwards the cells were incubated for 1 h with a 1:40,000 dilution of anti-protein A antibody (Sigma-Aldrich) in blocking solution. After washing (1× PBS, 0.05% Tween 20), goat anti-rabbit Alexa fluor® 488/594 secondary antibodies (1:400 dilution [ThermoFischer Scientific]) were added for 1 h in blocking solution (including 1 μg/ml DAPI). Finally, cells were washed (1× PBS, 0.05% Tween 20) and coverslips were mounted in Immu-Mount™ (ThermoFischer Scientific). For imaging, an Axioskop 20 fluorescence microscope (Zeiss) and the provided Axio Vision software were used.

4.8.2 DNA oligonucleotide annealing

400 pmol of complementary DNA oligonucleotides (see 4.1.11) were annealed in 25 µl annealing buffer (10 mM Tris-Cl pH 7.5, 50 mM NaCl, 1 mM EDTA) as follows:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>94°C</td>
<td>1°C in 45 sec</td>
</tr>
<tr>
<td>4°C</td>
<td>hold</td>
</tr>
</tbody>
</table>

4.8.3 *In vitro* transcription

*In vitro* transcription was performed from annealed DNA oligonucleotides (see 4.8.1) or a linearized plasmid that contained the 3’ UTR of the *T. brucei* α-Tubulin gene (provided by Christian Preußer), using T7 RNA polymerase (ThermoFischer Scientific). The reaction was incubated for 3 h at 37°C and set up as follows:

<table>
<thead>
<tr>
<th>Volume (µl)</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.00</td>
<td>5× transcription buffer</td>
</tr>
<tr>
<td>2.50</td>
<td>DTT (100 mM)</td>
</tr>
<tr>
<td>6.75</td>
<td>DMPC-H₂O</td>
</tr>
<tr>
<td>2.50</td>
<td>ATP or CTP (2mM; depending on labeling)</td>
</tr>
<tr>
<td>1.25</td>
<td>UTP (10 mM)</td>
</tr>
<tr>
<td>1.25</td>
<td>GTP (10 mM)</td>
</tr>
<tr>
<td>1.25</td>
<td>ATP or CTP (10 mM; depending on labeling)</td>
</tr>
</tbody>
</table>
The DNA template was digested by the addition of 3 μl RQ1 DNase (1 U/μl) and incubation for 30 min at 37°C. Prior RNA purification using Mini quick spin RNA columns (Roche Applied Science), a 1 μl aliquot was taken for scintillation (Tri-Carb 1600TR [PACKARD]). Afterwards the transcripts were ethanol precipitated (see 4.2.5), separated on a 15% denaturing polyacrylamide/urea gel (see 4.5.3), detected by autoradiography (see 4.5.4) and excised from the gel. To isolate RNA, the gel piece was crushed and rotated overnight in 400 μl 1× PK buffer (100 mM Tris-Cl pH 8.0, 12.5 mM EDTA, 150 mM NaCl, 1% SDS). The liquid portion was transferred, the RNA was ethanol precipitated (see 4.2.5), subjected to scintillation (Tri-Carb 1600TR [PACKARD]) and resuspended in DMPC-H$_2$O after calculating the yield according to the following formula:

\[
RNA [ng] = NTP_{total unlabeled} [\mu M] \times \text{reaction volume} [\mu l] \times \frac{RNA activity}{activity reaction} \times 1.32
\]

4.8.4 In vitro polyadenylation assay

For standard in vitro polyadenylation assays, 4 nM in vitro transcribed $^{32}$P-labeled RNA (see 4.8.3), derived from the 3’ end of T. brucei $\alpha$-tubulin mRNA, was assembled with poly(A) buffer (25 mM Tris-Cl pH 7.4, 50 mM KCl, 1.5 mM MgCl$_2$, 1.5 mM MnCl$_2$, 0.05 mM EDTA, 0.5 mM DTT, 10% glycerol, 2.6% polyvinyl alcohol [PVA]). 32 nM recombinant purified poly(A) polymerase (see 4.7.5) was added on ice. The reaction was started by the addition of 1 mM ATP and incubation at 37°C. RNAs were ethanol precipitated (see 4.2.5) and analyzed by denaturing polyacrylamide/urea gel electrophoresis (see 4.5.3) in combination with autoradiography (see 4.5.4).

4.8.5 Tandem affinity purification (TAP) and mass spectrometry

Tandem affinity purification of PTP-tagged proteins was done as described by Schimanski et al., 2005, with minor modifications: Briefly, T. brucei cells were collected from 3-liter cultures (corresponding to ~4 ml packed cell volume) and lysed in 7 ml extraction buffer (150 mM KCl, 20 mM Tris-Cl pH 7.7, 3 mM MgCl$_2$, 0.5 mM DTT, cOmplete™ EDTA-free protease inhibitor [Roche Applied Science]). For IgG affinity chromatography, 300 μl of IgG sepharose 6 fast flow beads (packed bead volume [GE Healthcare]) were incubated with extract for two hours at 4°C. Beads
were washed extensively with PA-150 buffer (150 mM KCl, 20 mM Tris-Cl pH 7.7, 3 mM MgCl₂, 0.5 mM DTT, 0.1% Tween 20), followed by TEV protease buffer (150 mM KCl, 20 mM Tris-Cl pH 7.7, 3 mM MgCl₂, 0.5 mM EDTA, 0.5 mM DTT, 0.1% Tween 20). Tagged proteins were eluted in 2 ml of TEV protease buffer containing 250 units of AcTEV protease (ThermoFischer Scientific). For anti-protein C affinity purification, CaCl₂ was added to the eluate to a final concentration of 2 mM. The eluate was incubated for 2 hours at 4°C with 300 µl of anti-protein C affinity matrix (packed bead volume [Roche Applied Science]). The beads were washed with PC-150 buffer (150 mM KCl, 20 mM Tris-Cl pH 7.7, 3 mM MgCl₂, 1 mM CaCl₂, 0.1% Tween 20), and the protein C-tagged proteins were eluted with 0.5 ml EGTA elution buffer (5 mM Tris-Cl pH 7.7, 10 mM EGTA, 5 mM EDTA). Eluted proteins were concentrated, using a SpeedVac® plus SC100A (Savant) vacuum concentrator and StrataClean resins (Stratagene), analyzed on 4-12% NuPAGE gels (ThermoFischer Scientific), and stained with InstantBlue-protein stain (Expedeon) (see 4.5.6) Proteins were in-gel digested with endoproteinase trypsin and analyzed by LC-MS/MS under standard conditions. Data were searched against a T. brucei protein sequence database (NCBInr v.11.09.2013, 38,057 entries). Data were evaluated with “Scaffold version 4.0.4” (Proteome Software), and proteins were ranked by their unique peptide counts in LC-MS/MS analysis, using a peptide and protein threshold of 95.0% and a minimum of three peptides.

4.8.6 In vivo polyadenylation assay
To determine the general polyadenylation status of mRNAs, we measured the length of the poly(A) tails by splint labeling according to Tkacz et al., 2010. Briefly: Two to five µg of total RNA (see 4.2.3) was mixed with 200 pmol oligo(dG)-oligo(dT) primer in 1× sequinase buffer, heated to 85°C and quenched on ice for 30 min. Prior heating, 1 µl of the reaction was analyzed by RT-PCR (see 4.4.1 and 4.4.2) using U3 or 7SL RNA specific primer, serving as loading control for normalization. Afterwards 0.1 nmol DTT, 4.8 U USB® sequenase™ version 2.0 DNA polymerase (Affymetrix) and 10 pmol dCTP including 2.5 µCi [α-³²P]-dCTP (3000 Ci/mmol; SCP-205 [Hartmann Analytic]) were added. The reaction was incubated for 30 min, ethanol precipitated (see 4.2.5) and resuspended in 50 µl RNase A/T1 mix (2 µg RNase A / 5U RNase T1 mix [ThermoFischer Scientific], 20 µg yeast tRNA [Roche Applied Science], 300 mM NaCl, 10 mM Tris-Cl pH 7.5, 5 mM EDTA) to digest (37°C, 30 min) the RNA leaving intact only the radioactively labeled poly(A) tails. The poly(A) tails were purified by phenol/chloroform extraction (see 4.2.4) and ethanol precipitation (see 4.2.5), separated on a denaturing 15% polyacrylamide/urea gel (see 4.5.3) and detected by autoradiography (see 4.5.4). Quantitation was performed by measuring the relative intensities of ³²P-labeled product bands, using
4. Material and Methods

“ImageJ”. Fold changes were obtained by calculating the middle “Riemann sum”, which can be understood as a weighted average of the polyadenylation efficiency:

\[ S = \sum_{i=1}^{n} f \left( 1 + \frac{i}{2} \right) \times \frac{1}{2} \]

in which \( S \) is the “Riemann sum” of the function \( f(x) \) over the total measured intensities \( n \) divided into separate sections \( i \).

4.8.7 Individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP)

Three biological replicates of iCLIP experiments using a stable PTP-tagged DRBD4 cell line and wildtype \( T. brucei \) 427 (WT) cells as negative control were performed according to an established strategy with minimal changes (Preußer et al., 2014). In brief: After UV-C cross-linking, cell lysates were prepared in 1.2 ml extraction buffer (150 mM KCl, 20 mM Tris-Cl pH 7.7, 3 mM MgCl\(_2\), 0.5 mM DTT, cOmplete™ EDTA-free protease inhibitor [Roche Applied Science]) using a Dounce homogenizer (25 strokes with a type B pestle) followed by sonication. The RNP-complexes were purified after DNase (Promega) and limited RNase I (final concentration 1 U/ml [ThermosFischer Scientific]) treatment by applying the first step of tandem affinity purification (IgG sepharose 6 fast flow beads [GE Healthcare]). After SAP (ThermoFischer Scientific) treatment, 3’ RNA-linker ligation (T4 RNA ligase [ThermoFischer Scientific], AcTEV protease (ThermoFischer Scientific) cleavage, the second affinity step (anti-protein C affinity matrix [Roche Applied Science]) and 5’ terminal \(^{32}\)P-labeling (T4 PNK [New England Biolabs]) the RNP-complexes were gel separated under denaturing conditions (NuPAGE electrophoresis and blotting system; NuPAGE™ Novex™ 4-12% Bis-Tris protein gels 1.0 mm 12-wells [both ThermoFischer Scientific]. Afterwards the complexes were electroblotted onto a nitrocellulose membrane (0.45 \( \mu \)m [Bio-Rad]) and bound RNA was recovered by PK (Roth) treatment and precipitation (see 4.2.5). RNA was reverse transcribed (Superscript III reverse transcriptase [ThermoFischer Scientific]), using oligonucleotides which introduce a 5’ barcode as well as a BamHI restriction site. The cDNAs were size-fractionated by denaturing polyacrylamide gel electrophoresis (Novex® TBE-Urea gels 6% 10-wells [ThermoFischer Scientific]), circularized (Circligase II [Epicentre]) and after annealing of a complementary oligonucleotide, linearized between the two adapter regions by BamHI (ThermoFischer Scientific) digestion. The cDNA was PCR amplified (30-32 cycles; Accuprime supermix 1 [ThermoFischer Scientific]) and the library was purified using QIAquick PCR purification kit (QIAGEN) and Caliper (XT DNA 750 assay kit [Perkin-Elmer]). After quality control using an Agilent 2100 Bioanalyzer and High sensitivity DNA analysis kit (both Agilent Technologies), the
library was subjected to high-throughput sequencing on an Illumina MiSeq instrument (single-read; 100 bp for DRBD4_1/2 and WT; 75 bp for DRBD4_3).

The YFP-based iCLIP protocol was adapted from the protocol described above with minimal changes, including skipping of ActEV protease cleavage and the second affinity step: 2 µg anti-GFP antibody (Roche Applied Science) in TBS-T (50 mM Tris-Cl pH 7.5, 150 mM NaCl, 0.05% Tween20) was bound for 2 h (4°C) to Dynabeads protein G (ThermoFischer Scientific). The beads were washed with TBS-T and cell lysate was applied for 2 h at 4°C after DNase and RNase (final concentration 10 U/ml) treatment. Afterwards the beads were extensively washed using TBS-T1000 (50 mM Tris-Cl pH 7.4, 1 M NaCl, 0.05% Tween 20), followed by SAP treatment. The library was PCR amplified (27 cycles), purified and subjected to high-throughput sequencing on an Illumina MiSeq instrument (single-read; 75 bp).

Sequence reads were aligned to \textit{T. brucei} 927 genome assembly (TriTrypDB; http://tritrypdb.org; release 6: TriTrypDB-6.0_TbruceiTREU927) to determine the crosslink sites and their corresponding crosslink-tag counts. For details of sequencing data analysis, see Preußer \textit{et al.}, 2014 and references therein. The \textit{T. brucei} 927 genome was used for mapping because of its more comprehensive genome assembly and gene annotation in comparison with \textit{T. brucei} 427.

4.8.8 Systematic evolution of ligands by exponential enrichment-sequencing (SELEX-seq)

An RNA pool with a degenerated sequence of 20 nts was prepared by T7 \textit{in vitro} transcription (HiScribe™ T7 high yield RNA synthesis kit [New England Biolabs]), using a DNA oligonucleotide containing the random sequence annealed to a T7 promoter complementary DNA oligonucleotide. Selection was performed in a total volume of 200 µl (10 mM Tris-HCl pH 7.5, 100 mM KCl, 2.5 mM MgCl₂, 0.1% Triton X-100) with 40 pmol GST-DRBD4/GST-CPSF30 or GST (as negative control) bound (see 4.7.2) to pre-blocked glutathione sepharose 4B beads (GE Healthcare) and 4 nmol transcript. After a 20 min incubation at room temperature, the samples were washed three times with 1 ml of washing buffer (10 mM Tris-HCl pH 7.5, 100/300/600 mM KCl, 2.5 mM MgCl₂, 0.1% Triton X-100), treated with PK (Roth), phenol/chloroform extracted (see 4.2.4), and ethanol precipitated (see 4.2.5). The stringency of washing steps was increased for each round of selection. Selected RNAs were reverse-transcribed (qScript™ flex cDNA kit [Quanta Biosciences]) and PCR amplified (see 4.4.1 and 4.4.2). Transcripts for the next round of selection were generated by \textit{in vitro} transcription as described above. After five rounds of selection, RNA aliquots from each round and from the fifth round of GST selection were used for barcoding (barcodes 1-5: GST-DRBD4 and 6: GST or barcodes 7-11: GST-CPSF30 and 12: GST) by reverse transcription (qScript™ flex cDNA kit
4. Material and Methods

[Quanta Biosciences]). cDNA libraries were amplified by PCR (see 4.4.2), pooled in equal amounts and purified by QIAquick PCR purification kit (Qiagen) and Caliper (XT DNA 750 assay kit [Perkin-Elmer]). After quality control using an Agilent 2100 Bioanalyzer and High sensitivity DNA analysis kit (both Agilent Technologies), the library was subjected to high-throughput sequencing on a MiSeq instrument (single-read 100 bp [Illumina]). Sequence reads were first sample-barcode sorted, trimmed by PCR primer sequences on both ends, and further random-barcode filtered to obtain 20 nts sequence tags of the enriched RNA pools. The numbers of filtered sequence tags (from each SELEX round) containing either one of the 256 or 4,096 possible tetramer or hexamer motifs, respectively, was summarized, and the z-score values were calculated for enrichment of each motif. Moreover, the top eight tetramers from the final selection round were combined, allowing up to two nucleotides spacing between them, and ranked according to their frequency in the SELEX-sequence tags.

4.8.9 GST pulldown assay

For assaying protein-RNA interactions, 10 pmol of GST-DRBD4 (or GST as negative control) was immobilized (see 4.7.2) on glutathione sepharose 4B beads (GE Healthcare) and incubated with 500 fmol denatured, in vitro transcribed, $^{32}$P-labeled RNAs (see 4.8.3) in a final volume of 200 µl binding buffer (10 mM Tris-Cl pH 7.5, 100 mM KCl, 2.5 mM MgCl$_2$, 0.1 mM DTT, 0.01% NP-40). After binding at room temperature for 30 min, beads were washed four times with 1 ml binding buffer. Bound RNA was recovered by the addition of 400 µl PK buffer (100 mM Tris-Cl pH 7.5, 12.5 mM EDTA, 150 mM NaCl, 1% SDS) and 10 min incubation at 80ºC. Afterwards the RNA was extracted with phenol/chloroform (see 4.2.4), ethanol precipitated (see 4.2.5), analyzed on a 15% denaturing polyacrylamide/urea gel (see 4.5.3) and detected by autoradiography (see 4.5.4). For quantification of pulldown efficiencies the Typhoon FLA 9500 (Ge Healthcare) imaging system was used (see 4.5.4).

4.8.10 RNA-seq and identification of polyadenylation sites

RNA-seq libraries for the determination of polyadenylation sites were made and sequenced at the Beijing Genomics Institute (Shenzhen, China). Poly(A)$^+$ RNA was purified from total RNA, converted to cDNA using a 5’-T15VN-3’ oligonucleotide (V=A, G or C; N=T, A, G or C), sheared, and size-selected for fragments ~200 bp in length, using the Illumina TruSeq RNA sample preparation kit v2 (Illumnia). Sequencing was performed on an Illumina Hiseq 2000 (Illumina) platform yielding 90 bp paired-end reads. Polyadenylation sites were identified by the SLaP mapper tool (Fiebig et al., 2014), using the T. brucei 927 genome assembly.
4.8.11 Correlation between DRBD4 iCLIP binding and polyadenylation site choice

The poly(A) sites identified in uninduced control cells by RNA-Seq were clustered with a maximal spacing of 20 nucleotides. The poly(A) site reads from the control and the DRBD4-knockdown cells in each of the 8,962 clusters were summarized. The Bioconductor package DESeq (Anders & Huber, 2010) was applied to determine the differential expression of poly(A) site clusters. As a result, 234 down- and 248 up-regulated clusters with a minimum of 10 poly(A) site reads in control cells (14 in knockdown cells) were selected for further analysis, using as a criterion an absolute fold change greater than 0.75 in log₂ value. Poly(A) site clusters with p-values greater than 0.75 (expression not affected by DRBD4 knockdown) were selected for background group selection. For each regulated cluster, four clusters with similar expression levels (in the range of 0.9 to 1.1) were randomly picked as background clusters. The position 0 of the poly(A) site cluster was defined as the nucleotide with the highest poly(A) read count within the cluster. The iCLIP crosslink-tag counts in the region between -200 nts and +200 nts of the regulated and background clusters were summarized. For the distribution of iCLIP crosslink tags, fitted values with 21-nts windows were used for smooth-line presentation, with iCLIP tag counts plotted on the Y-axis and positions relative to the defined position 0 on the X-axis.

4.8.12 Flow cytometry

Mid-log phase density cells (5×10⁶ cells/ml) were analyzed with and without doxycycline induction, using a BD FACScan (BD Biosciences).
4.9 Constructs

4.9.1 Generation of pLew100 RNAi constructs

For a specific RNAi-mediated knockdown of a single gene (Table 4.4), 300-500 bp of the 3' UTR were cloned (see 4.6) as a stem loop construct in the pLEW100 vector, allowing the expression of a double stranded RNA (Wang et al., 2000). The target gene insert was amplified by PCR (see 4.4.2) using a forward primer containing a HindIII and MluI restriction site and a reverse primer containing a Xbal restriction site. The “sense insert” was first cloned in pJM325, using HindIII and Xbal restriction sites. The “antisense insert” was directly cloned into pLew100, using MluI and Xbal restriction sites. Afterwards the “sense insert” and a ~530 bp long stuffer derived from pJM325 were cloned into pLEW100, using HindIII and Xbal restriction sites.

Table 4.4 Generation of pLEW100 RNAi constructs.

<table>
<thead>
<tr>
<th>Protein</th>
<th>TriTrypDB annotation</th>
<th>cloned nucleotides of the gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAP</td>
<td>Tb927.7.3780</td>
<td>1527-1997</td>
</tr>
<tr>
<td>PAP</td>
<td>Tb927. 3.3160</td>
<td>944-1440</td>
</tr>
<tr>
<td>CPSF160</td>
<td>Tb927.11.14650</td>
<td>3243-3720</td>
</tr>
<tr>
<td>CPSF100</td>
<td>Tb927.11.230</td>
<td>584-997</td>
</tr>
<tr>
<td>CPSF73</td>
<td>Tb927.4.1340</td>
<td>1683-2106</td>
</tr>
<tr>
<td>CPSF30</td>
<td>Tb927.11.12750</td>
<td>271-759</td>
</tr>
<tr>
<td>Fip1</td>
<td>Tb927.5.4320</td>
<td>285-765</td>
</tr>
<tr>
<td>Symplekin</td>
<td>Tb927.8.7490</td>
<td>3754-4235</td>
</tr>
<tr>
<td>CstF64</td>
<td>Tb927.8.8210</td>
<td>4-396</td>
</tr>
<tr>
<td>CstF50</td>
<td>Tb927.6.1830</td>
<td>744-1191</td>
</tr>
<tr>
<td>CFIm25</td>
<td>Tb927.7.1620</td>
<td>392-819</td>
</tr>
<tr>
<td>CFIm</td>
<td>Tb927.6.3690</td>
<td>729-1183</td>
</tr>
<tr>
<td>hyp. Protein</td>
<td>Tb927.11.13860</td>
<td>234-710</td>
</tr>
<tr>
<td>hyp. Protein</td>
<td>Tb927.8.4480</td>
<td>264-724</td>
</tr>
</tbody>
</table>

4.9.2 Generation of PTP-tag fusion protein constructs

For the expression of C-terminal PTP-tagged proteins from the endogenous locus, the pC-PTP-NEO vector was used (Schimanski et al., 2005b).

The ORFs of Tb927.3.3160 (nts 672-1461), Tb927.7.3780 (nts 1172-2061), CPSF160 (nts 3601-4356) and CPSF30 (nts 74-831), were PCR-amplified (see 4.4.2) and inserted in-frame (see 4.6)
into the pC-PTP-NEO vector upstream of the PTP tag sequence, using the Apal and NotI restriction sites.

4.9.3 Generation of protein overexpression constructs

For overexpression of recombinant proteins in Sf9 insect or One shot® BL21 (DE3) chemically competent E. coli cells, the pFastBac™HT B (ThermoFischer Scientific) and pGEX-6P-2 (GE Healthcare) vectors were used, respectively. The ORFs of PAP Tb927.3.3160 (nts 3-1464) and PAP Tb927.7.3780 (nts 3-2064) were PCR amplified (see 4.4.2) and cloned (see 4.6) into the pFastBac™HT B vector, using the BamHI and EcoRI restriction sites. The ORFs of Tb927.11.12750 (nts 3-834; CPSF30) and Tb927.11.14100 (nts 3-1518; DRBD4) were PCR amplified (see 4.4.2) and cloned (see 4.6) into the pGEX-6P-2 vector, using the BamHI and EcoRI or SalI and NotI restriction sites, respectively.
5. Results

5.1 The trypanosomatid polyadenylation complex

5.1.1 The trypanosomatid poly(A) polymerases

To identify the major functional poly(A) polymerase in *T. brucei*, we focus here on two putative poly(A) polymerases [Tb927.3.3160 and Tb927.7.3780] and initially compared the sequences with their human homolog, PAPOLA (Figure 5.1 and Supplementary Figure 8.1; page 128-129). Based on the Needleman-Wunsch alignment algorithm, the human PAPOLA protein shows 17.4% and 24% identity with the trypanosomatid Tb927.3.3160 and Tb927.7.3780 proteins from *T. brucei*, respectively. Although there are regions of low sequence similarity, structure predictions of these two *T. brucei* poly(A) polymerases suggest a strong structure conservation of the two proteins. Moreover, by domain structure prediction of these two *T. brucei* poly(A) polymerases, the highly conserved domains of the poly(A) polymerase described in higher eukaryotes, such as the central domain harboring the nucleotidyltransferase activity including the three catalytic aspartic acid residues, are present in both trypanosomatid proteins (Zhelkovsky *et al*., 1995, Martin & Keller, 1996, Martin *et al*., 1999). Following the central domain, an RNA-binding domain was predicted for all three proteins. However, both trypanosomatid proteins lack a recognizable, classical nuclear localization signal at their C-terminus.

![Figure 5.1 Domain structure of trypanosomatid poly(A) polymerases.](image)

Domain organization of the PAPs, comparing the human poly(A) polymerase alpha (hPAPOLA) and two putative trypanosomatid PAPs [Tb927.7.3780 and Tb927.3.3160]. The PAP central domain, harboring the nucleotidyltransferase domain and the conserved catalytic aspartic acid residues (in blue), the RNA-binding domain (in green), and the known nuclear localization signals (NLS; in magenta) are illustrated.
To compare the two putative enzymes, we next generated two clonal procyclic cell lines, which stably express either Tb927.3.3160 or Tb927.7.3780, each with a C-terminal PTP-tag (protein C epitope / TEV cleavage site / 2× protein A epitope). Expression of both proteins was monitored by Western blot analysis (Figure 5.2 A) and in addition, the subcellular distribution of both proteins was visualized by indirect immunofluorescence (Figure 5.2 C and D). Interestingly, the PTP-tagged version of the Tb927.7.3780 predominantly localizes to the nucleus, whereas the putative poly(A) polymerase Tb927.3.3160 clearly localizes mainly in distinct cytoplasmic spots, suggesting differential functions. However, one has to consider that the large C-terminal tag may influence the localization within the cell.

Figure 5.2 Expression and cellular localization of two poly(A) polymerases in T. brucei.
(A) The expression of PTP-tagged putative poly(A) polymerase [Tb927.7.3780] and [Tb927.3.3160] in T. brucei cells was detected by Western blotting, using polyclonal anti-protein A antibodies (lane PTP). As control, T. brucei 427 wildtype (lane WT) cells were included. Protein size markers in kDa.
(B and C) Cell lines stably expressing PTP-tagged versions of the two poly(A) polymerases were fixed and stained with DAPI. The localization of tagged proteins was assessed by indirect immunofluorescence, using anti-protein A antibodies followed by Alexa-594 and Alexa-488-coupled secondary antibodies for Tb927.7.3780 and Tb927.3.3160, respectively. Superimpositions of DAPI and PTP-tagged proteins (merge) as well as with brightfield are shown (merge with brightfield).
5. Results

5.1.2 Only the putative poly(A) polymerase Tb927.7.3780 shows *in vitro* activity

To assess and compare the enzymatic activity of both proteins, we expressed His-tagged versions of the two poly(A) polymerases in Sf9 insect cells, using the Bac-to-Bac® baculovirus expression system followed by HPLC purification (Figure 5.3).

![Figure 5.3 Recombinant expression of trypanosomatid poly(A) polymerases in Sf9 cells.](image)

Recombinant proteins were expressed in baculovirus-infected Sf9 cells, HPLC-purified, and analyzed by SDS-PAGE and Coomassie staining. BSA standards (100-500 ng) were used for quantification. Protein size markers (M) in kDa.

The purified recombinant proteins were assayed for unspecific polyadenylation activity, using an RNA substrate derived from the 3′ end of the *T. brucei* α-tubulin mRNA. Interestingly, only the Tb927.7.3780 protein showed *in vitro* activity under these conditions, in contrast to the Tb927.3.3160 protein (Figure 5.4; page 71). The RNA substrate was efficiently elongated by Tb927.7.3780 (approximately 150 nts within 10 min), suggesting a high processivity of the enzyme. As previously shown for the human poly(A) polymerase, non-specific activity is very inefficient in the absence of specificity factors, such as CPSF or CFI; however, if Mg$^{2+}$ is replaced by Mn$^{2+}$, the unspecific activity is strongly enhanced (Christofori & Keller, 1988, Wahle *et al.*, 1991). This is also the case for the poly(A) polymerase Tb927.7.3780 in trypanosomes (Figure 5.4; page 71).
5. Results

Figure 5.4 Recombinant putative poly(A) polymerase Tb927.7.3780 polyadenylates in vitro.
Only the putative poly(A) polymerase Tb927.7.3780 shows in vitro activity and requires manganese as a cofactor for unspecific polyadenylation. In vitro transcribed $^{32}$P-labeled RNA (derived from the *T. brucei* α-tubulin 3’ UTR) was incubated at 37°C with baculovirus-expressed and purified putative poly(A) polymerase [Tb927.3.3160 or Tb927.7.3780]. Reactions were performed in the presence of MnCl$_2$ or MgCl$_2$ or a mixture of both (1.5 mM), and either in presence (+) or absence (-) of ATP. Reactions were analyzed after 30 minutes on a denaturing 12.5% polyacrylamide/urea gel. Marker sizes in nucleotides.

5.1.3 *In vivo* activity of the putative poly(A) polymerases
To compare the activities of the two trypanosomatid poly(A) polymerases *in vivo*, we silenced them individually in procyclic *T. brucei* by doxycycline-inducible RNAi. The knockdown efficiency was confirmed by RT-qPCR, using primer detecting the respective Tb927.3.3160 and Tb927.7.3780 mRNAs (Figure 5.5 A; left side; page 72). After three days of RNAi induction, we observed a knockdown efficiency of approximately 80% and 70%, respectively. In addition, we performed semiquantitative RT-PCR using the same primer pairs, confirming the knockdown (Figure 5.5 A; right side; page 72). When we monitored the cell viability upon knockdown, only
after silencing Tb927.7.3780 expression, cell growth was rapidly affected; in contrast, knockdown of Tb927.3.3160 did not result in a phenotypic growth defect (**Figure 5.5 B**).

**Figure 5.5 Only the putative poly(A) polymerase Tb927.7.3780 is required for cell viability.**

(A) The mRNA levels were analyzed by real-time RT-PCR (left) or semiquantitative RT-PCR (right) from uninduced cells (t0) and after 1, 2, and 3 days of RNAi induction (t24-t72, as indicated). U3 RNA served for normalization and as control. M, markers (100, 200, 300, 400, and 500 bp). (B) Growth curves of representative cell lines, in which RNAi was induced by doxycycline, (+ Dox, lines with asterisks) and respective uninduced controls (-Dox, lines with triangles).

To investigate whether individual knockdown of the two poly(A) polymerases affects the global mRNA polyadenylation status *in vivo*, total RNA was analyzed after three days, and the poly(A) tail length was examined according to Tkcaz *et al.*, 2010. In brief, poly(A) tails were labeled at their 3’ end by splint ligation, using an oligo(dG)-oligo(dT) primer to allow subsequent fill-in with \(^{32}\)P-dCTP. Total RNA was then treated with RNase A and T1, leaving intact poly(A) tails, which were analyzed by gel electrophoresis. The relative band intensities of the poly(A) tails were measured by ImageJ and normalized against signals of U3 RNA (**Figure 5.6**; page 73). The results clearly indicate that Tb927.7.3780 depletion severely reduced relative amounts and lengths of the poly(A) tails, unlike Tb927.3.3160 depletion, where an opposite effect of a slight increase in activity was observed. Taken together, our results indicate that the poly(A)
polymerase Tb927.7.3780 (annotated as putative) represents the major functional nuclear poly(A) polymerase in trypanosomes. In contrast, the poly(A) polymerase from the intron-containing gene Tb927.3.3160 localizes mainly in the cytoplasm and appears not to be functional in general 3’ polyadenylation.

![Image](https://example.com/image.png)

**Figure 5.6 Only the putative poly(A) polymerase Tb927.7.3780 is active in polyadenylation in vivo.** Depletion of poly(A) polymerase Tb927.7.3780, but not of Tb927.3.3160, impairs polyadenylation in vivo. RNA from uninduced (- Dox) and induced (+ Dox) cells was splint-labeled, digested with RNases A/T1, followed by separation of the labeled tails on a denaturing 15% polyacrylamide/urea gel. U3 RNA detected by RT-PCR served as loading control. Marker sizes are indicated in nucleotides. The relative intensities of both signals were measured with ImageJ and plotted. The black and grey lines represent the tail length distributions in uninduced and induced cells, respectively.

### 5.1.4 Identification of the constituents of the polyadenylation complex

To systematically investigate the composition of the trypanosomatid polyadenylation machinery, we used tandem affinity purification (TAP), based on the largest subunit of the CPSF complex, the putative CPSF160 [Tb927.11.14560]. We generated a clonal procyclic cell line, which stably expresses CPSF160 with a C terminal PTP tag (for Western blot analysis, see Figure 5.7 A; for
its cellular distribution, Figure 5.7 B). As expected, CPSF160-PTP predominantly localizes to the nucleus, with only minor staining of the cytoplasm.

**Figure 5.7 Expression and cellular localization of *T. brucei* CPSF160-PTP.**

(A) Expression of PTP-tagged CPSF160 homolog [Tb927.11.14560; lane PTP] in *T. brucei* cells was detected by Western blotting, using polyclonal anti-protein A antibodies. As a control, *T. brucei* 427 wildtype (lane WT) cells were analyzed. Protein size marker in kDa.

(B) Nuclear localization of CPSF160-PTP. Cell lines stably expressing PTP-tagged *T. brucei* CPSF160 were fixed and stained with DAPI. The localization of tagged protein was assessed by indirect immunofluorescence, using anti-protein A antibodies (CPSF160-PTP). In addition, superimpositions of DAPI and PTP-tagged protein (merge) as well as with brightfield are shown (merge with brightfield).

Based on CPSF160-PTP, two independent tandem affinity purifications (TAP) were performed (prot C exp. 1 and 2); as control, wildtype *T. brucei* 427 cells (WT) were used (Figure 5.8; page 75). Factors co-purifying with CPSF160 were analyzed by mass spectrometry and are listed in Figure 5.8 (page 75; for raw data, see Supplementary Table 8.1; page 131-133): Eight proteins were identified by homology search as putative factors of the polyadenylation complex in trypanosomes, including five CPSF subunits (CPSF160; CPSF100; CPSF73; CPSF30; Fip1), two constituents of CstF (CstF64 hinge domain [see below]; CstF50), and the scaffold protein Symplekin. Two additional proteins were found [Tb927.11.13860; Tb927.8.4480], which are likely associated with the polyadenylation complex, since they are clearly above the peptide counts of the background control (WT). However, neither the poly(A) polymerase Tb927.7.3780 nor Tb927.3.3160 was co-purified with CPSF160-PTP. In addition, we did homology-based databank searches for trypanosomatid homologs of other eukaryotic polyadenylation factors, which had not appeared in our mass spectrometry analysis. As a result, we identified potential homologs of the 25 kDa subunit of cleavage factor I (CFIm25) [Tb927.7.1620] and cleavage factor II (CFIIm) [Tb927.6.3690] (Figure 5.8; page 75).
5. Results

Figure 5.8 Identification of components of the polyadenylation complex, based on TAP-tag purification of the nuclear localized trypanosomatid CPSF160 homolog. CPSF160-associated proteins were identified by mass spectrometry. Extracts were prepared from *T. brucei* cell lines that stably express tagged CPSF160 (in two biological replicates; exp. 1/2), followed by TAP-tag affinity purification. Proteins after the second purification step (protein C elution) were analyzed by SDS-PAGE and Coomassie staining (lanes prot C experiment 1/2). As control, *T. brucei* 427 wildtype cells were used (lane WT). Protein size markers in kDa (M). Putative protein constituents of the *T. brucei* polyadenylation complex were identified by mass spectrometry. Each protein is described by name (factor), the TriTrypDB annotation number (http://tritrypdb.org/tritrypdb/), its molecular mass (in kDa), the number of exclusive unique peptide counts obtained by mass spectrometry, and, if applicable, by literature reference. Below, the three proteins identified only by database search (PAP, CFI25m, and CFI11) are listed by name, TriTrypDB annotation number, and molecular mass.

<table>
<thead>
<tr>
<th>factor</th>
<th>annotation in TriTrypDB</th>
<th>mol. mass [kDa]</th>
<th>peptides experiment 1/2</th>
<th>reference</th>
</tr>
</thead>
</table>
| CPSF
| CPSF160 | Tb927.11.14560 | 159.5 | 83/84 |
| CPSF100 | Tb927.11.230 | 90.6 | 45/42 |
| CPSF73 | Tb927.4.1340 | 85.3 | 37/33 |
| CPSF30 | Tb927.11.12750 | 31.6 | 15/12 |
| Fip1 | Tb927.5.4320 | 31.0 | 9/9 |
| CstF
| CstF64 | Tb927.8.8210 | 14.7 | 3/7 |
| CstF50 | Tb927.6.1830 | 49.9 | 2/5 |
| Others
| Symplekin | Tb927.8.7490 | 158.0 | 79/76 |
| hyp. protein | Tb927.11.13860 | 29.7 | 15/17 |
| hyp. protein | Tb927.8.4480 | 28.6 | 15/18 |

Identified by database search

<table>
<thead>
<tr>
<th>factor</th>
<th>annotation in TriTrypDB</th>
<th>mol. mass [kDa]</th>
<th>peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAP</td>
<td>Tb927.7.3780</td>
<td>74.4</td>
<td></td>
</tr>
<tr>
<td>CFI25m</td>
<td>Tb927.7.1620</td>
<td>32.7</td>
<td></td>
</tr>
<tr>
<td>CFI11</td>
<td>Tb927.6.3690</td>
<td>46.0</td>
<td></td>
</tr>
</tbody>
</table>
In addition to MS-based homology search, we further confirmed our data by comparative InterProScan protein domain identification of the individual trypanosomatid factors and their respective human counterparts (Figure 5.9; page 78).

All CPSFs (see above) identified in our mass spectrometry analysis have a domain organization that is similar to their respective human homologs: For both CPSF160 subunits (sequence identity: 20.9%) the C-terminal CPSF A domain, whose function is unknown, could be identified. Both the human as well as the trypanosomatid CPSF100 (sequence identity 22.0%) and CPSF73 (sequence identity 39.3%) have a β-lactamase domain with a β-caspase domain (β-casp) inserted like a cassette with the active site of the protein at the interface of these two domains (Mandel et al., 2006). In higher eukaryotes CPSF73 is the endonuclease in 3’ end processing, and the residues (mainly histidines) coordinating the two zinc atoms that are required for RNA cleavage are conserved in trypanosomes as well. In contrast, CPSF100 lacks the zinc-binding residues and cannot bind zinc, arguing against nuclease activity of the protein. In addition both CPSF 100 and CPSF73 have an RMMBL domain (RNA metabolizing metallo beta lactamase) and characteristic C-terminal domains, required for in vivo core cleavage complex formation composed of CPSF100, CPSF73 and Symplekin (Michalski & Steiniger, 2015).

The domain structure of *T. brucei* CPSF30 is compared to the human protein conserved (sequence identity 31.0%), harboring five C3H1 zinc fingers and a C-terminal C2HC zinc finger. The human protein was shown to bind (in cooperation with WDR33) the canonical AAUAAA hexamer polyadenylation signal via zinc fingers two and three, with a minor contribution of the C-terminal zinc finger (Chan et al., 2014, Schönemann et al., 2014, Shimberg et al., 2016). Although, the AAUAAA polyadenylation signal was not found in trypanosomes, the conserved CPSF30 domain structure argues for a possible role in RNA binding.

Fip1 (sequence identity 12.7%), the fifth CPSF identified in our mass spectrometry analysis, contains the Fip1 domain (Fip1) present in the human homolog, which is required for the interaction with CPSF30 (zinc fingers four and five) in trypanosomes (Bercovich et al., 2009). The C-terminal arginine-aspartate-rich region (RD-rich) of the human protein that binds to U-rich RNA sequences is not present in the trypanosomatid protein but it contains an additional C3H1 zinc finger (Kaufmann et al., 2004).

All other identified proteins except for CstF64 display conserved domain architectures as well. CstF64 (sequence identity 4.9%) was only identified through a putative hinge domain present in both the trypanosomatid and human protein that was shown to interact with CstF77 and Symplekin in a mutual exclusive manner in humans (Takagaki & Manley, 2000, Ruepp et al., 2011b). The N-terminal RRM binding the U-rich DSE and the highly conserved C-terminal
domain (CstF64-C) required for polyadenylation are absent in trypanosomes (MacDonald et al., 1994, Gross & Moore, 2001, Qu et al., 2007).

Both the *T. brucei* and the human CstF50 protein (sequence identity 17.2%) contain the WD40 repeat domain, which functions as a binding platform for CstF77 interaction (Takagaki & Manley, 2000). However, the N-terminal self-dimerization domain (CstF1-dd) of the human homolog is missing.

Trypanosomatid CFIm25 (sequence identity 23.7%) and CFIm (sequence identity 26.6 %) have the same domain structure as their respective human counterparts, of which the CFIm25 NUDIX hydrolase domain is of special interest, since it binds the USE and regulates poly(A) site selection in humans (Yang et al., 2010).

Symplekin (sequence identity 12.6%) functions as a scaffold protein and the C-terminal domain (Symplekin-C) is conserved in trypanosomes as well, but the ARM repeats (armadillo-type fold) are missing.

Take together our results suggest a general conservation of polyadenylation complex constituents among eukaryotes.
5. Results

Figure 5.9 Domain conservation of polyadenylation factors.
Schematics of human polyadenylation factors and the *T. brucei* homologs. Domains were identified by InterProScan 5 scanning and are depicted in scale as colored boxes. CPSF A, cleavage and polyadenylation specificity factor, A subunit, C-terminal domain; 8-lactamase, beta-lactamase domain; 8-casp, beta-caspase domain; RMMBL, RNA metabolizing metallo beta lactamase domain (metal binding amino acid residues are colored in red); CPSF2-C, cleavage and polyadenylation specificity factor 2, C-terminal domain; CPSF73-C, cleavage and polyadenylation specificity factor 3 C-terminal domain; ZnF C3H1-type, zinc finger C3H1-type; ZnF C2HC, zinc finger type C2HC; Fip1, pre-mRNA polyadenylation factor Fip1 domain; RD-rich, arginine-aspartate-rich region; H-rich, histidine-rich region; P-rich, proline-rich region; CstF1-dd, cleavage stimulation factor subunit 1, dimerization domain; RRM, RNA recognition motif; hinge domain; CstF64-C, cleavage stimulating factor C-terminal domain; WD40 repeat, tryptophan-aspartate repeat; NUDIX, NUDIX hydrolase domain; MobB, MobB-type P-loop domain; Clp1, Pre-mRNA cleavage complex II cleavage and polyadenylation factor 1 domain; ARM repeat, armadillo-type fold; Symplekin-C, Symplekin C-terminal domain.
5.1.5 Characterization of the CPSF30 RNA-binding properties

To further analyze a possible RNA-binding function of CPSF30, suggested by the conserved domain structure (see 5.1.4), a SELEX-seq experiment was performed to identify an *in vitro* binding motif.

In brief: Recombinantly expressed and purified GST-CPSF30 and GST, as control, ([Figure 5.10](#)) was incubated with an *in vitro* T7 transcribed RNA, derived from a DNA oligonucleotide containing a random 20-mer sequence ([Figure 5.11 A](#); page 80). Bound RNA was recovered, reverse-transcribed, PCR-amplified and used again for *in vitro* T7 transcription followed by another round of selection. Finally, libraries of all five rounds of selection (for GST, only round five) were prepared and subjected to high-throughput sequencing. We obtained between 0.55 mio and 0.87 mio reads (0.86 mio for GST) containing a 18 to 20-mer sequence that were used for CPSF30 RNA-binding motif analysis.

The representation of each of the possible tetramer (256) and hexamer (4096) motifs was determined after each SELEX cycle, and the enrichment of each tetramer and hexamer was evaluated by z-score values. ([Figure 5.11 B](#) page 80) shows heat maps for the top 20 and the bottom 10 tetramers (upper panel) and hexamers (lower panel) (ordered according to their z-score after the fifth SELEX round with GST-CPSF30 protein), especially how the motifs were enriched over the five cycles (CPSF30 R1 to R5). The most enriched tetramers and hexamers were long poly(U) stretches, whereas GC-rich sequences were most depleted, arguing for an oligo(U) *in vitro* RNA-binding motif of trypanosomatid CPSF30.

In addition, for further iCLIP-based *in vivo* analysis of CPSF30 RNA-binding a clonal procyclic cell line stably expressing a C-terminal PTP-tagged CPSF30 was generated ([Figure 5.12](#); page 80).

![Quantification of E. coli-expressed GST-CPSF30 and GST.](image)

Recombinant proteins were expressed in *E.coli* BL21 DE3 star cells, bound to glutathione sepharose and analyzed by SDS-PAGE and Coomassie staining. BSA standards (1-0.2 µg) were used for quantification. Protein size markers in kDa.
Figure 5.11 Identification of the CPSF30 in vitro RNA-binding motif by SELEX-seq.

(A) Schematic representation of the SELEX-seq procedure and data analysis. Using GST-CPSF30 (GST as control) and an N20 RNA pool, CPSF30 RNA-binding sequences were enriched through five SELEX rounds. After each selection round (CPSF30 R1-R5; GST control only R5), aliquots were taken for library preparation and high-throughput sequencing, followed by motif enrichment analysis (see panel B).

(B) Motif enrichment analysis of CPSF30 RNA-binding. The upper and lower panel shows the enrichment of tetramers and hexamers during the five SELEX rounds with GST-CPSF30 (R1-R5) and GST control (R5), respectively. Only the top 20 and the bottom 10 tetramers and hexamers are represented as colour-coded heat maps, showing z-scores of the motif frequencies for each round (ordered by z-score in R5 of GST-CPSF30).

Figure 5.12 Validation of CPSF30-PTP expression.

The expression of PTP-tagged CPSF30 [Tb927.11.12750] in T. brucei cells was detected by Western blotting, using polyclonal anti-protein A antibodies (lane PTP). As control, T. brucei 427 wildtype (lane WT) cells were included. Protein size markers in kDa.
5.1.6 Polyadenylation factors are required for both polyadenylation and trans splicing

To assess the functional relevance of the proteins identified in the polyadenylation complex, we depleted each of them by inducible RNAi: Each of the individual knockdowns showed a severe growth defect in procyclic trypanosomes, with the exception of one of the two hypothetical proteins, Tb927.11.13860 (Figure 5.13; page 82). Following quality control of isolated total RNA by gel electrophoresis and silver staining, depletion of the individual factors was validated by semiquantitative RT-PCR (Figure 5.14 A and B; page 83). Afterwards the *in vivo* polyadenylation status was examined (Figure 5.15 A; page 84; for raw data, see Figure 5.16; page 85-86) as described for the two poly(A) polymerases (see 5.1.3).

Knockdown of CPSF100, CPSF73, and CPSF30 strongly reduced the overall poly(A) tail lengths. In contrast, depletion of CPSF160 and Fip1 -as part of the CPSF complex- did not cause significant changes. CstF50 depletion strongly affected polyadenylation, in contrast to CstF64 depletion. Note that the putative CstF64 analyzed here exhibits only a very low sequence similarity to its mammalian counterpart (see 5.1.4); it may represent a truncated homolog functionally different from the mammalian factor or a trypanosomatid-specific factor.

Depletion of the scaffold factor Symplekin resulted in a severe polyadenylation phenotype. In contrast, depletion of the two cleavage factors CFIm25 and CFIlm did not significantly affect poly(A) tail length. Finally, when analyzing the two unannotated proteins [Tb927.11.13860 and Tb927.8.4480], small, but significant inhibitory and stimulatory effects were observed, respectively.

As several previous studies have indicated, both major pre-mRNA processing steps, *trans* splicing and polyadenylation, are interconnected. Therefore we also analyzed *trans* splicing activities to compare them directly with the corresponding polyadenylation effects. Both spliced-leader (SL) RNA (Figure 5.15 B; page 84), which is known to accumulate upon inhibition of *trans* splicing (for example, see Tkacz et al., 2010), and a specific *trans*-spliced mRNA (PPlase) (Figure 5.15 C; page 84) were measured by quantitative RT-PCR after knockdown of the individual polyadenylation factors. Depletion of the CPSF complex subunits 160, 100, 73, and 30 resulted in spliced-leader accumulation and severe splicing deficiencies. However, for Fip1 only blocking of splicing was observed, but no effect on SL RNA accumulation (Figure 5.15 C; page 84). Knockdown of both CstF subunits 64 and 50, Symplekin, CFIlm, and the two hypothetical proteins Tb927.11.13860 and Tb927.8.4480 resulted in moderate effects on spliced-leader accumulation and in most cases also on inhibition of *trans* splicing. However, depletion of CFIm25 had neither an effect on SL RNA accumulation nor on *trans* splicing. These effects on PPlase *trans* splicing were largely confirmed with another gene, PRP8, analyzed by
Results

We conclude that the corresponding knockdown effects of most core polyadenylation factors on both polyadenylation and trans splicing, as well as splicing deficiencies after CPSF160 and Fip1 depletion, argue for a general coupling of these two major RNA processing steps.

Figure 5.13 Growth curves of procyclic trypanosomes after RNAi depletion of polyadenylation complex constituents.

Growth curves of representative cell lines, in which RNAi was induced by doxycycline (+Dox, lines with asterisks) and respective uninduced controls (-Dox, lines with triangles). The following factors were subjected to RNAi induction (TriTrypDB numbers in brackets): CPSF160 [Tb927.11.14560]; CPSF100 [Tb927.11.230]; CPSF73 [Tb927.11.12750]; CPSF30 [Tb927.11.12750]; Fip1 [Tb927.5.4320]; CstF50 [Tb927.6.1830]; CstF64 [Tb927.8.8210]; CFIm25 [Tb927.7.1620]; CFIm [Tb927.6.3690]; Symplekin [Tb927.8.7490]; Tb927.11.13860; Tb927.8.4480.
5. Results

Figure 5.14 RNAi mediated knockdown of polyadenylation complex constituents.

(A) RNA was isolated from uninduced (-) and RNAi induced cells (+), separated on a denaturing 10% polyacrylamide/urea gel and visualized by silver staining. The ribosomal sRNA1, sRNA2, 5.8S, sRNA4 and 5S RNA as well as the tRNAs are indicated. The following factors were subjected to RNAi induction (TriTrypDB numbers in brackets): CPSF160 [Tb927.11.14560]; CPSF100 [Tb927.11.230]; CPSF73 [Tb927.4.1340]; CPSF30 [Tb927.11.12750]; Fip1 [Tb927.5.4320]; CstF64 [Tb927.8.8210]; CstF50 [Tb927.6.1830]; CFIm25 [Tb927.7.1620]; CFIm [Tb927.6.3690]; Symplekin [Tb927.8.7490]; Tb927.11.13860; Tb927.8.4480.

(B) mRNA levels were analyzed by semiquantitative RT-PCR from uninduced cells (-) and after 3 days of RNAi induction (+). As control, 7SL RNA from the same samples was analyzed. M, markers (200, 300, 400, and 500 bp; for 7SL: 100 and 200 bp).
5. Results

Figure 5.15 Polyadenylation factors required for both polyadenylation and splicing in vivo.
The following factors were subjected to RNAi knockdown (TriTrypDB numbers in brackets): CPSF160
[Tb927.11.14560]; CPSF100 [Tb927.11.230]; CPSF73 [Tb927.4.1340]; CPSF30 [Tb927.11.12750]; Fip1
[Tb927.5.4320]; CstF50 [Tb927.6.1830]; CstF64 [Tb927.8.8210]; CFIm25 [Tb927.7.1620]; CFIm
[Tb927.6.3690]; Symplekin [Tb927.8.7490]; Tb927.11.13860; Tb927.8.4480. Changes in the overall
poly(A) tail length (A), in SL RNA accumulation (B), and in trans splicing of PPlase mRNA (C) upon
depletion of individual polyadenylation factors are graphically shown. Poly(A) tail lengths were analyzed
and quantitated by the “Riemann’s sum”, comparing induced and uninduced cells (for raw data, see
Figure 5.15; page 86/87). SL RNA accumulation and trans-splicing activities were determined by RT-
qPCR on SL RNA and trans spliced PPlase mRNA, respectively, and normalized to U3 RNA. For both
polyadenylation, SL RNA accumulation, and trans splicing, fold changes (in log2 values) of uninduced and
induced levels were calculated, with standard deviations derived from three biological replicates (* for P
values <0.05; ** for P values <0.01).
5. Results

A

B

CPSF160  CPSF73  CPSF30  CstF64  CstF50  Symplekin  CFim25  CFilm  Tbs27.11.13660  Tbs27.8.4480  Dox

poly(A)  -7SL

length of poly(A) tail

CPSF160  CPSF100  CPSF73  CPSF30

length of poly(A) tail
Figure 5.16 Changes in the poly(A) tail length upon depletion of 12 individual trypanosomatid polyadenylation factors.

(A and B; page 85-86) The following factors were subjected to RNAi induction (TriTrypDB numbers in brackets): CPSF160 [Tb927.11.14560]; CPSF100 [Tb927.11.230]; CPSF73 [Tb927.4.1340]; CPSF30 [Tb927.11.12750]; Fip1 [Tb927.5.4320]; CstF50 [Tb927.6.1830]; CstF64 [Tb927.8.8210]; CFIm25 [Tb927.7.1620]; CFIm [Tb927.6.3690]; Symplekin [Tb927.8.7490]; Tb927.11.13860; Tb927.8.4480.

(A) RNA from uninduced (-Dox) and RNAi induced (+Dox) cells was splint-labeled, digested with RNases A/T1, followed by separation of the $^{32}$P-labeled tails [poly(A)] on a denaturing 15% polyacrylamide/urea gel. 7SL RNA detected by RT-PCR served as loading control. Marker sizes are indicated in nucleotides.

(B) The relative intensities of knockdown and uninduced control signals were measured with ImageJ and plotted. The tail length distributions in uninduced (blue lines) and induced cells (red lines) are represented.
Figure 5.17 Changes in PRP8 mRNA trans splicing upon depletion of 12 individual trypanosomatid polyadenylation factors.

The respective splicing activities were monitored in uninduced (-Dox) and induced (+Dox) cells by semiquantitative RT-PCR. To detect the trans spliced PRP8 mRNA, SL- and PRP8 exon-specific primer were used. The primer positions and orientations are schematically represented on the right side. As control 7SL RNA was analyzed. M, marker (100, 200, 300, 400, and 500 bp). TriTrypDB numbers in brackets: CPSF160 [Tb927.11.14560]; CPSF100 [Tb927.11.230]; CPSF73 [Tb927.4.1340]; CPSF30 [Tb927.11.12750]; Fip1 [Tb927.5.4320]; CstF50 [Tb927.6.1830]; CstF64 [Tb927.8.8210]; CFIm25 [Tb927.7.1620]; CFIm [Tb927.6.3690]; Symplekin [Tb927.8.7490]; Tb927.11.13860; Tb927.8.4480.
5. Results

5.2 DRBD4-mediated regulation of mRNA polyadenylation in *Trypanosoma brucei*

5.2.1 DRBD4 binds *in vivo* preferentially to gene UTRs

To map DRBD4-RNA contacts at single-nucleotide resolution in procyclic *T. brucei* cells, we applied the iCLIP approach. Three individual experiments were performed, using a cell line expressing DRBD4 with a PTP tag (protein C epitope / TEV cleavage site / 2× protein A epitope), with wildtype *T. brucei* 427 (WT) cells as negative control. Trypanosomes were UV-crosslinked, and DRBD4-RNA complexes were purified by tandem affinity purification (TAP), separated by SDS-PAGE, and blotted onto nitrocellulose membrane (Figure 5.18). Subsequently, bound RNA was eluted and used for library preparation, followed by high-throughput sequencing.

![Figure 5.18](image)

**Figure 5.18 Representative DRBD4-PTP iCLIP experiment.**
Protein-RNA interactions were crosslinked by UV-irradiation of DRBD4-PTP-expressing cells or Tb427 wildtype cells (WT; negative control), followed by lysate preparation, limited RNase digestion, immunoprecipitation of DRBD4-PTP RNA-protein (RNP) complexes, 3’ linker ligation, tobacco-etch-virus (TEV) protease cleavage and the second affinity step (anti-protein C immunoaffinity purification). 5’-terminal 32P-labeled RNP-complexes were then gel-separated under denaturing conditions and visualized by autoradiography (left panel). In each case, RNase I amounts were titrated (as indicated). Boxed regions were cut out and subjected to RNA isolation and library preparation. Western blot analysis (right panel) of the same membrane was performed using anti-protein C antibodies including 2% input material. The gel positions of DRBD4 with the full PTP tag (PTP) and the remaining protein C epitope (P) after TEV-protease cleavage are indicated.

Approximately 3.6 mio of the 9.5 mio total DRBD4-PTP iCLIP sequences could be uniquely mapped to the *T. brucei* genome (Figure 5.19; page 89). After random-barcode filtering, the 387,577 crosslink tags from 255,063 crosslink sites were used for downstream analysis. Within the individual DRBD4 iCLIP crosslink tags, approximately 56% mapped to mRNAs, with a strong preference for the UTR regions (~42%, compared with ~14% in the coding regions). The large
proportion of not annotated iCLIP crosslink tags (~37%) can be most likely explained by the incomplete annotation of gene UTRs. In addition, the overall highly consistent distribution of iCLIP tags in the different RNA classes argues for the reproducibility of the individual experiments.

### Table

<table>
<thead>
<tr>
<th></th>
<th>sequenced reads</th>
<th>uniquely mapped reads</th>
<th>crosslink site counts</th>
<th>crosslink tags counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb427 (WT)</td>
<td>828,656</td>
<td>83,965</td>
<td>1,264</td>
<td>2,677</td>
</tr>
<tr>
<td>DRBD4 (total)</td>
<td>9,489,415</td>
<td>3,625,320</td>
<td>255,063</td>
<td>387,577</td>
</tr>
<tr>
<td>DRBD4-PTP_1</td>
<td>2,435,834</td>
<td>392,110</td>
<td>23,128</td>
<td>36,327</td>
</tr>
<tr>
<td>DRBD4-PTP_2</td>
<td>4,187,648</td>
<td>1,838,268</td>
<td>164,787</td>
<td>249,799</td>
</tr>
<tr>
<td>DRBD4-PTP_3</td>
<td>2,865,933</td>
<td>1,394,942</td>
<td>67,148</td>
<td>101,451</td>
</tr>
</tbody>
</table>

**Figure 5.19** Genome-wide mapping of DRBD4-RNA interactions in *T. brucei* by iCLIP-seq.

Summary of sequenced/mapped iCLIP reads (upper panel) and the genomic distribution of barcode-filtered tag counts (in percent; lower panel), based on three individual DRBD4-PTP experiments.

In addition to the PTP-based iCLIP-approach (see above), a second iCLIP protocol using a YFP-tag was established. A single experiment was performed with a procyclic DRBD4-YFP cell line and wildtype *T. brucei* 427 (WT) cells served as negative control. The experimental procedure is as described above, with the exception that only a one step immunoprecipitation using anti-GFP antibodies is applied.

Approximately 1.1 mio of the 2.2 mio DRBD4-YFP iCLIP sequences were uniquely mapped to the *T. brucei* genome (**Figure 5.20**; page 90). After random-barcode filtering, the 954,953 crosslink tags from 739,861 crosslink sites were used for downstream analysis. The majority of YFP-DRBD4 iCLIP tags mapped -similar to the DRBD4-PTP iCLIP (see above)- to mRNAs (68.1%), displaying again a preference for UTR regions (~40.2%, compared with ~27.9% in the coding regions). Moreover, a large proportion of not annotated iCLIP crosslink tags (~30.5%) was observed (see above). A general reproducibility between both protocols is reflected by the constant distribution of iCLIP tags in the different classes of RNAs in the PTP-based and YFP-based experiments. However, the numbers of uniquely mapped reads, crosslink site counts and crosslink tag counts in the YFP-based experiment were strongly increased, when compared to
individual DRBD-PTP experiments with a similar number of total sequenced reads (DRBD4-PTP_1 and 3). This indicates that the YFP-library was more complex, compared to the individual DRBD4-PTP libraries, which can be most likely explained by a higher YFP-immunoprecipitation efficiency, since only a single-step method was used.

![Figure 5.20 DRBD4-YFP iCLIP experiment.](image)

(A) Protein-RNA interactions were crosslinked by UV-irradiation of DRBD4-YFP expressing cells or Tb427 wildtype cells (WT; negative control), followed by lysate preparation, limited RNase digestion, immunoprecipitation of DRBD4-YFP RNA-protein (RNP) complexes and 3’ linker ligation. 5’-terminal 32P-labeled RNP-complexes were then gel-separated under denaturing conditions and visualized by autoradiography (left panel). In each case, RNase I amounts were titrated (as indicated). Boxed regions were cut out and subjected to RNA isolation and library preparation. Western blot analysis (right panel) of the same membrane was performed using anti-protein GFP antibodies including 1.5% input material. The gel position of DRBD4-YFP is indicated.

(B) Summary of sequenced/mapped iCLIP reads (upper panel) and the genomic distribution of barcode-filtered tag counts (in percent; lower panel), in the DRBD4-YFP experiment.
5.2.2 DRBD4 binds purine-rich sequences containing AUGA elements \textit{in vitro}

To further investigate the DRBD4 RNA-binding properties we next used a SELEX-seq approach to derive a minimal binding motif. Briefly, recombinant \textit{T. brucei} GST-DRBD4 (with GST protein as negative control; \textbf{Figure 5.21 A}) was incubated with RNA T7-transcribed from a DNA oligonucleotide template containing a random 20-mer sequence (\textbf{Figure 5.21 B}). Bound RNA was recovered, reverse-transcribed, PCR-amplified, and used again for \textit{in vitro} T7 transcription, followed by further rounds of selection. Finally, RNA-seq libraries were prepared from aliquots after each of the five selection rounds (for the GST control, only after the fifth round) and subjected to high-throughput sequencing.

Between 0.45 and 0.67 mio sequence tags per SELEX cycle were obtained for GST-DRBD4 (for GST control: 2.02 mio after the fifth cycle), with 20-mer sequences that were analyzed for the \textit{in vitro} RNA-binding motif of DRBD4.

\textbf{Figure 5.21 Identification of the RNA-binding motif of DRBD4 by SELEX-seq.}

\textbf{(A)} Quantification of \textit{E. coli}-expressed GST-DRBD4 and GST. Recombinant proteins were expressed in \textit{E. coli} BL21 DE3 star cells, bound to glutathione sepharose and analyzed by SDS-PAGE and Coomassie staining. BSA standards (1-0.2 µg) were used for quantification.

\textbf{(B)} Schematic representation of the SELEX-seq procedure and data analysis. Using GST-DRBD4 (GST as control) and an N\textsubscript{20} RNA pool, DRBD4 RNA-binding sequences were enriched through five SELEX rounds. After each selection round (DRBD4 R1-R5; GST control only R5), aliquots were taken for library preparation and high-throughput sequencing, followed by motif enrichment analysis (see Figure 5.22; page 93).
The representation of each of the possible tetramer (256) and hexamer (4096) motifs was determined after each cycle, and the enrichment of each tetramer and hexamer was evaluated by z-score values. **Figure 5.22** (page 93) shows heat maps for the top 20 and the bottom 10 tetramers (upper panel) and hexamers (middle panel) (ordered according to their z-score after the fifth SELEX round with GST-DRBD4 protein), especially how the motifs were enriched over the five cycles (DRBD4 R1 to R5).

Interestingly, sequences containing purines—in particular adenosines—were highly enriched during the selection process, with the AUGA tetramer being most enriched. This was further underlined by the hexamer analysis, where also AUGA flanked by two A nucleotides was most enriched. In contrast, pyrimidine-rich stretches appear to be depleted in both the tetramer and hexamer analyses.

Since a single RRM can recognize up to eight nucleotides and DRBD4 contains four RRMs, RNA binding of DRBD4 may be combinatorial and involve more than a single minimal motif. Therefore we also analyzed tetramer combinations: The eight most enriched tetramers from the final selection round with GST-DRBD4 were combined, allowing up to two nucleotides spacing between them, and ranked according to their representation in the SELEX-sequence tags. **Figure 5.22** (lower panel; page 93) shows a heat map of the top 20 and bottom 10 most enriched double-tetramer combinations, ordered according to their motif frequency (sum of 0/1/2-nts spacing). As the most enriched double-tetramer combinations AUGA-AUGA and GAAA-UGAA were identified, each without spacer (note that the latter one also contains an AUGA motif, with flanking adenosine nucleotides, as the most enriched hexamer).
Figure 5.22 Motif enrichment analysis of DRBD4 RNA binding.
The upper and middle panel shows the enrichment of tetramers and hexamers during the five SELEX rounds with GST-DRBD4 (R1-R5) and GST control (R5). Only the top 20 and the bottom 10 tetramers and hexamers are represented as colour-coded heat maps, showing z-scores of the motif frequencies for each round (ordered by z-score in R5 of GST-DRBD4). The lower panel shows the enrichment for combinations of the top eight tetramers from the upper panel (with 0/1/2 nts spacing). Only the top 20 and the bottom 10 double-tetramers are displayed as colour-coded heat map, indicating motif frequencies in the R5 round of GST-DRBD4 (ordered by the sum of frequencies of 0/1/2 nts spacing). The motifs selected for experimental validation are highlighted in bold.

To biochemically validate the binding motif, we assayed for *in vitro* binding of GST-DRBD4 to selected $^{32}$P-labeled RNA transcripts, using GST pulldown and analysis of associated RNAs by denaturing PAGE, followed by quantitation (Figure 5.23 A; page 95).

Surprisingly, the most enriched hexamer (motif #1; in two copies with an eight-nucleotide spacer in between, see Figure 5.23 A for context; page 95) and the double-tetramer combination GAAA-UGAA (motif #3; in a single copy) did not significantly bind DRBD4 above GST control levels. In contrast, the transcript with the double-AUGA sequence (motif #2) efficiently bound
and appears to represent a minimal functional binding motif. To further analyze a possible spacing tolerance of this minimal motif, which was suggested by a high motif frequency with one or two spacing nucleotides (see Figure 5.22; lower panel; page 93), adenosine nucleotides (which were overrepresented in our SELEX-derived, DRBD4-binding sequences; Figure 5.22; page 93), were inserted between the two AUGA tetramers (motifs #6-8): Compared with the minimal double-tetramer (motif #2), the insertion of two adenosine (A₂) nucleotides reduced, whereas four (A₄) or eight (A₈) adenosines enhanced binding. However, three adjacent AUGA motifs without spacing did not further increase the binding efficiency (motif #4). The specificity for and requirement of the AUGA unit was underlined by the strong effect of a single point mutation in the tetramer motif (AUGA to AUUA; compare motif #4 and #10). Moreover, another triple-tetramer combination (motif #5), with the high-affinity AUGA flanked by AAGA and AGAA, among the most enriched tetramers (see Figure 5.23 A; upper panel; page 95), bound almost as well as the triple-AUGA sequence. Surprisingly, even a simple oligo-adenosine stretch (A)₁₂ was bound with similar efficiency (motif #9).

In addition, to test a polypyrimidine-rich DRBD4 binding motif as suggested in a previous study (Stern et al., 2009), we also assayed DRBD4 binding to four T. brucei polypyrimidine tracts (Figure 5.23 B; page 95). However, the relatively high DRBD4 binding affinities we observed were in all cases similar to the GST control, arguing for unspecific background binding probably based on the GST moiety.

Taken together, our binding assays further confirmed and validated the SELEX-seq data: We conclude that DRBD4 binds in vitro RNA sequences that are purine-, in particular adenosine-rich, often harboring AUGA elements.
Figure 5.23 DRBD4 preferentially binds purine-rich sequences containing AUGA motifs. 

$^{32}$P labeled short RNA transcripts containing different (A) SELEX-derived binding motifs (#1-10) or (B) polypyrimidine tracts form *T. brucei* genes (#1: Tb927.1.2340; #2, 3: Tb927.11.3630; #4: Tb927.6.510) were incubated with GST-DRBD4 or GST (negative control), immobilized on glutathione sepharose. Bound RNA was recovered and analyzed by denaturing polyacrylamide gel electrophoresis (marker sizes in nucleotides). For each pulldown assay, 10% of the input and the total recovered RNA were loaded. The respective pulldown efficiencies including standard deviations were calculated from three replicates.
5. Results

5.2.3 DRBD4 depletion affects polyadenylation site usage

In addition to the reported functions of the trypanosome DRBD4 in mRNA stability and splicing (Stern et al., 2009), DRBD4 was identified in a targeted RNAi screen for genes that were necessary for the instability of developmentally regulated mRNAs. In the course of this work, it was found that RNAi depletion of DRBD4 resulted in an altered poly(A) site location in the GPI-PLC (Tb927.2.6000) mRNA (see 5.2.4). As the human homolog PTB1 can also modulate polyadenylation efficiency (Castelo-Branco et al., 2004, Izquierdo et al., 2005, Fred & Welsh, 2009) we assayed for a regulatory role of DRBD4 in trypanosomatid 3´ end processing of mRNAs. In a procyclic T. brucei cell line expressing endogenous YFP-tagged DRBD4, depletion was initiated for three days by doxycycline-inducible RNAi and monitored through flow cytometry (Figure 5.24 A); the cells continued to proliferate throughout the time course (Figure 5.24 B).

![Figure 5.24 Validation of RNAi-mediated DRBD4 depletion.](image)

**Figure 5.24 Validation of RNAi-mediated DRBD4 depletion.**

(A) Flow cytometry measurements of YPF in cell lines expressing tet repressor (pSPR2, gray) and an additional endogenous YFP-tagged DRBD4 allele (p3217, green). DRBD4-YFP levels were measured in a time course after 18, 42, 66 and 90 hours of DRBD4-RNAi (p4376) induction (blue; +Dox), and uninduced cells served as control (red; -Dox). FL1-H indicates the relative intensity of the YFP fluorescence.

(B) Growth curve of the DRBD4-RNAi line (as in A), in which RNAi was induced by doxycycline (red line with asterisks), and the respective uninduced control (blue line with triangles).
Total RNA was isolated from cells after RNAi and from uninduced control, followed by poly(A)$^+$ RNA selection, library preparation for 3´ end mapping and RNA-seq (Table 5.1). Based on 6.86 mio and 10.48 mio total sequenced reads in control and knockdown cells, respectively, polyadenylation sites in both samples were mapped using the SLaP mapper tool: Based on 1.19 mio (in control) and 1.70 mio poly(A)-site-spanning reads (in knockdown cells), we identified in total 152,628 (control) and 200,141 (knockdown) individual polyadenylation sites. First, poly(A) sites from the uninduced control cells were clustered (to give 8,962 poly(A) site clusters) and second, the poly(A)-site reads in each cluster were summarized and compared with those in DRBD4-knockdown cells: As a result, we identified 234 down- and 248 up-regulated poly(A) site clusters (Table 5.1).

Table 5.1 DRBD-mediated poly(A) site choice.
Summary of poly(A)-site-spanning reads, poly(A) sites, poly(A) site clusters, and up-/down-regulated clusters, based on RNA-seq analysis of DRBD4 knockdown (kd) and control (ctr) cells.

<table>
<thead>
<tr>
<th></th>
<th>ctr</th>
<th>kd</th>
</tr>
</thead>
<tbody>
<tr>
<td>total sequenced reads</td>
<td>6,863,039</td>
<td>10,482,225</td>
</tr>
<tr>
<td>poly(A)-site-spanning reads</td>
<td>1,186,940</td>
<td>1,699,656</td>
</tr>
<tr>
<td>poly(A) sites</td>
<td>152,628</td>
<td>200,141</td>
</tr>
</tbody>
</table>

First, poly(A) sites from the uninduced control cells were clustered (to give 8,962 poly(A) site clusters) and second, the poly(A)-site reads in each cluster were summarized and compared with those in DRBD4-knockdown cells: As a result, we identified 234 down- and 248 up-regulated poly(A) site clusters (Table 5.1).

5.2.4 DRBD4-regulated poly(A) site choice: integrating iCLIP and RNA-seq data
Next, we combined our DRBD4-PTP iCLIP and RNA-seq datasets, with the aim to correlate in vivo RNA binding and DRBD4-dependent poly(A) site choice. Based on the 234 down- and 248 up-regulated poly(A) site clusters, iCLIP crosslink tag counts in the region between -200 to +200 nts of the regulated poly(A) site clusters were summarized. Likewise, non-regulated poly(A) site clusters with similar expression levels were selected as background group (see Materials and methods for details). Figure 5.25 (page 98, for DRBD4-YFP iCLIP see, Supplementary Figure 8.2; page 130) shows the accumulated distribution of DRBD4-PTP iCLIP tag counts (as smooth line in 21-nts windows) around the regulated poly(A) sites; a poly(A) site was defined as the position with the highest number of poly(A) site tags within the cluster, with the background group (dashed line) plotted for comparison. As a result, we found an enrichment of DRBD4 binding in the upstream region of regulated poly(A) sites (-20 to -140 nts for down-regulated, -100 to -180 nts for up-regulated sites).
In sum, these data provide evidence for a correlation between DRBD4 \textit{in vivo} binding and DRBD4-dependent changes in poly(A) site selection, arguing for a DRBD4-mediated regulation of this process.

\textbf{Figure 5.25 DRBD4 RNA binding around regulated poly(A) sites.}
Accumulated DRBD4-PTP iCLIP crosslink-tag counts were plotted in the region between -200 to +200 nts relative to the poly(A) site of the down- and up-regulated clusters (top: down-regulated n=234; background n=936; bottom: up-regulated n=248; background n=992) and are depicted as solid lines. As background control, unaffected polyadenylation sites were used (dashed lines). Crosslink-tag counts were plotted with fitted values (21-nts windows) on the Y-axis and positions on the X-axis.

To validate individual cases of DRBD4-dependent poly(A) site choice, we analyzed by Northern blotting changes in transcript levels after DRBD4 knockdown (at 24, 48, and 72 hours; \textbf{Figure 5.26}; page 100; for sequences and crosslink sites, see \textbf{Figure 5.27}; page 101). We initially tested two individual genes, Tb927.8.760 and Tb927.8.770, each containing a single poly(A) site, which is down-regulated after DRBD4 depletion (based on our RNA-seq read numbers: for Tb927.8.760, from 692 to 470 reads; for Tb927.8.770, from 223 to 128 reads) and Northern blot analysis confirmed lower transcript levels (\textbf{Figure 5.26 A}; top and middle; page 100).

In contrast, Tb927.11.5440 is more complex since there are two reciprocally used poly(A) sites, separated by 350 nucleotides (tandem 3’ UTR APA; see 3.4): the distal one, which is down-regulated after DRBD4 depletion (200 to 69 reads), and a proximal one, which is up-regulated
(46 to 177 reads), causing a shift from a long to a short isoform (Figure 5.26 A; bottom; page 100). This shift in isoform levels, due to differential poly(A) site choice, could be nicely confirmed by Northern blotting, which showed after DRBD4 knockdown a decrease of the long and a parallel, approximately two-fold increase of the short isoform. Interestingly, in the bloodstream form, the opposite can be observed, that the short Tb927.11.5440 isoform is predominantly expressed (Figure 5.28; page 102).

In addition, we tested two genes with DRBD4 crosslink sites upstream of a poly(A) site that was up-regulated after DRBD4 knockdown (Figure 5.26 B; page 100; for sequences and crosslink sites, see Figure 5.27; page 101): First, Tb927.6.1470 contains two poly(A) sites with a 960-nucleotide distance (tandem 3´ UTR APA; see 3.4), of which the proximal site is up-regulated (62 to 270 reads) and the distal one down-regulated (52 to 32 reads). Northern blot analysis clearly confirmed this strong and reciprocal shift in poly(A) site use.

In the second example, Tb927.2.6000 encoding a developmentally regulated glycosylphosphatidylinositol-phospholipase C (GPI-PLC), DRBD4 depletion activated a cryptic poly(A) site (proximal; 0 to 27 reads), whereas the distal polyadenylation site was not strongly affected (18 to 25 reads). The GPI-PLC mRNA is normally unstable in procyclic cells (Webb et al., 2005a) but the shortening of the 3´ UTR resulting from use of the cryptic polyadenylation sites removed the cis-elements necessary to most of the instability (Webb et al., 2005b). The activation of the cryptic poly(A) site results in the expression of a transcript shortened by 2.5 kb (Figure 5.26 B; page 100), as validated by Northern blotting, which showed the respective band only after knockdown for 48 and 72 hours. The long transcript [using the distal poly(A) site] appears in three closely spaced bands, most likely due to alternative trans splicing; in combination, these long isoforms did not change significantly over the knockdown time course.

Taken together, our data clearly show that DRBD4 binds upstream of polyadenylation sites and is required for both their activation and repression. For the first time this provides evidence for a trans-acting RNA-binding protein regulating polyadenylation in trypanosomes.
5. Results

A

**Tb927.8.760**  
200 nts (%AG: 60.5)

### # CLIP tags

| # reads | ctr 692 | kd 470 |

### # reads

<table>
<thead>
<tr>
<th>(A)₁₀₀</th>
<th>0 24 48 72</th>
<th>RNAi [h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.8.760</td>
<td>1 0.9 0.7 0.7</td>
<td>rel. expression</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>0.5 0.3 0.3</td>
<td></td>
</tr>
</tbody>
</table>

**Tb927.8.770**  
200 nts (%AG: 66.5)

### # CLIP tags

| # reads | ctr 233 | kd 128 |

### # reads

<table>
<thead>
<tr>
<th>(A)₁₀₀</th>
<th>0 24 48 72</th>
<th>RNAi [h]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.8.770</td>
<td>1 0.8 0.8 0.8</td>
<td>rel. expression</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>0.5 0.3 0.3</td>
<td></td>
</tr>
</tbody>
</table>

**Tb927.11.5440**  
200 nts (%AG: 61.5)

### # CLIP tags

| # reads | ctr 46 | kd 177 |

### # reads

<table>
<thead>
<tr>
<th>(A)₁₀₀</th>
<th>0.35 kb</th>
<th>(A)₁₀₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>prox.</td>
<td>1 1.2 1.4 2.0</td>
<td>dist.</td>
</tr>
<tr>
<td>long</td>
<td>short</td>
<td></td>
</tr>
</tbody>
</table>

### # reads

<table>
<thead>
<tr>
<th>long</th>
<th>short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.11.5440</td>
<td>1 1.0 1.2 1.4 2.0</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>0.5 0.3 0.2</td>
</tr>
<tr>
<td>rel. expression</td>
<td></td>
</tr>
</tbody>
</table>

B

**Tb927.6.1470**  
200 nts (%AG: 67.5)

### # CLIP tags

| # reads | ctr 62 | kd 270 |

### # reads

<table>
<thead>
<tr>
<th>(A)₁₀₀</th>
<th>0.96 kb</th>
<th>(A)₁₀₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>prox.</td>
<td>1 1.2 1.4 2.4</td>
<td>dist.</td>
</tr>
<tr>
<td>long</td>
<td>short</td>
<td></td>
</tr>
</tbody>
</table>

### # reads

<table>
<thead>
<tr>
<th>long</th>
<th>short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.6.1470</td>
<td>1 1.0 1.2 1.4 2.4</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>0.5 0.3 0.2</td>
</tr>
<tr>
<td>rel. expression</td>
<td></td>
</tr>
</tbody>
</table>

**Tb927.2.6000**  
200 nts (%AG: 51.5)

### # CLIP tags

| # reads | ctr 0 | kd 27 |

### # reads

<table>
<thead>
<tr>
<th>(A)₁₀₀</th>
<th>2.5 kb</th>
<th>(A)₁₀₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>prox.</td>
<td>1 1.0 1.2 1.4 1.6 1.8 2.0 2.4</td>
<td>dist.</td>
</tr>
<tr>
<td>long</td>
<td>short</td>
<td></td>
</tr>
</tbody>
</table>

### # reads

<table>
<thead>
<tr>
<th>long</th>
<th>short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tb927.2.6000</td>
<td>1 1.0 1.2 1.4 1.6 1.8 2.0 2.4</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>0.9 0.7 0.8</td>
</tr>
<tr>
<td>rel. expression</td>
<td></td>
</tr>
</tbody>
</table>

100
5. Results

Figure 5.26 DRBD4 binding modulates alternative poly(A) site usage.
For each gene the DRBD4 iCLIP tag distribution, DRBD4-dependent alternative poly(A) site usage (based on RNA-seq read numbers), and validation by Northern blot analysis are shown together (page 100; panel A, three examples down-regulated after DRBD4 knockdown: Tb927.8.760; Tb927.8.770; Tb927.11.5440; panel B, two up-regulated examples: Tb927.6.1470; Tb927.2.6000).
First, in each case the distribution of DRBD4 iCLIP tags within 200 nts upstream of the DRBD4-regulated poly(A) site [(A)n] is graphically represented, including the AG content of this region (in %). Gene bodies are marked as blue boxes, with distances between proximal (prox.) and distal (dist.) poly(A) sites indicated. Second, below the respective poly(A) sites the RNA-seq-based read counts are listed (DRBD4 knockdown at 0 versus 72 h). Third, for RNA-seq validation, DRBD4-dependent poly(A) site choice was directly analyzed by Northern blotting, showing four time points of DRBD4 knockdown (0-72 h) and using 18S rRNA as a loading control and for normalization. Below, the relative mRNA levels are listed.

<table>
<thead>
<tr>
<th>nts upstream of poly(A) site</th>
<th>mRNA sequence (% of purine crosslinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-200 to -151</td>
<td>GUUGAAGUGCAGCUUGCUUAGAUAUGGACGCCUUAG</td>
</tr>
<tr>
<td>-150 to -101</td>
<td>GAACTAAAGAAGGGUGAUGUAUAGGAGAGAAAAGUG</td>
</tr>
<tr>
<td>-100 to -51</td>
<td>AUAACUGCGUGAAUAGAUAGGAGCGUGAAUAGAUA</td>
</tr>
<tr>
<td>-50 to 0</td>
<td>GAAUCUCCUGGCAUCUGUAAGAAGGAGCUUAG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nts upstream of poly(A) site</th>
<th>mRNA sequence (% of purine crosslinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-200 to -151</td>
<td>GGUUGUAGGAGGCAGGCUUGCUUAGAUAUGGACG</td>
</tr>
<tr>
<td>-150 to -101</td>
<td>AAACUAGGAGCGAAACACACAGCACAGCAAG</td>
</tr>
<tr>
<td>-100 to -51</td>
<td>AAGAAGAAGAAGGUGAAGAAGAAGGUGAAGAAG</td>
</tr>
<tr>
<td>-50 to 0</td>
<td>GGAUACCAAAAGAGAGGUGGUGAGGAG</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nts upstream of poly(A) site</th>
<th>mRNA sequence (% of purine crosslinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-200 to -151</td>
<td>GUUGGCGAGTGATCAGCUGCUUGCUUAGAUAUGG</td>
</tr>
<tr>
<td>-150 to -101</td>
<td>UGUGAAGGAGCCAAUUGCGGUGUUAGUAAUAGGAG</td>
</tr>
<tr>
<td>-100 to -51</td>
<td>CAAAAGCGAAAAAGGCGAACGCUUCGCUUUAG</td>
</tr>
<tr>
<td>-50 to 0</td>
<td>GUGACGCCCGCCUCACAGCUUAAUGGAG</td>
</tr>
</tbody>
</table>

Figure 5.27 mRNA sequences upstream of regulated poly(A) sites.
mRNA sequences with purine nucleotides (marked in grey) upstream of DRBD4-regulated poly(A) sites. The DRBD4 iCLIP crosslink sites are indicated as bold letters and the respective superscripts indicate the number of crosslink site counts. The number of crosslinked purine nucleotides (relative to the total crosslinks) in given percent.
5. Results

**Figure 5.28** Life-cycle dependent expression of Tb927.11.5440 isoforms. Expression levels of the long and short Tb927.11.5440 mRNA in wildtype bloodstream (BS) and procyclic (PC) trypanosomes after DRBD4 knockdown (0-72 hours). Transcript isoforms were detected by Northern blotting, and 18S rRNA served as loading control.

### 5.2.5 DRBD4 is associated with polyadenylation complex constituents

To get initial insights in the DRBD4-dependent regulation of polyadenylation we identified DRBD4-associated proteins. Two independent tandem affinity purifications (TAP) of DRBD4-PTP were performed, and co-purified factors were analyzed by mass spectrometry (**Figure 5.29**; page 103). Among the approximately 200 factors (data not shown) with more than one unique peptide count (as sum of both experiments), we identified five proteins of the *T. brucei* polyadenylation complex, which were also purified by CPSF160-PTP TAP (see 5.1.4): Three subunits of the CPSF complex (CPSF160, CPSF100 and CPSF73), one constituent of the CstF (CstF50) and the scaffold protein Symplekin. In addition, we also co-purified the polyadenylation factor PABP2, the homolog of the human PABPN1 protein, which displayed in comparison to the polyadenylation complex constituents a high number of unique peptides. Interestingly, also the functional poly(A) polymerase [Tb927.7.3780], which was not purified by TAP of CPSF160-PTP, was associated with DRBD4. However, in contrast to the polyadenylation complex (see 5.1.4), DRBD4 does not form large, stable protein complexes, as indicated by a high number of co-purified proteins with a very low unique peptide count. In sum, these data of DRBD4-associated polyadenylation factors provides further evidence for a DRBD4-dependent regulation of polyadenylation in trypanosomes.
Figure 5.29 Identification of DRBD4-associated proteins by mass spectrometry.

Extracts were prepared from *T. brucei* cell lines that stably express tagged DRBD4 (in two biological replicates; exp 1/2), followed by TAP-tag affinity purification. Proteins after the second purification step (protein C elution) were analyzed by SDS-PAGE, Coomassie staining and Western blotting using polyclonal anti-protein A antibodies (lanes prot C exp 1/2). Protein size markers in kDa (M). Protein constituents involved in polyadenylation as well as putative factors were identified by mass spectrometry. Each protein is described by name (factor), the TriTrypDB annotation number (http://tritrypdb.org/tritrypdb/), its molecular mass (in kDa), the number of exclusive unique peptide counts obtained by mass spectrometry, and, if applicable, by literature reference.
6 Discussion

6.1 The poly(A) polymerase of T. brucei

Two genes Tb927.3.3160 and Tb927.7.3780 code for putative poly(A) polymerases in T. brucei, the first of which is peculiar in that it contains an intron requiring additional cis splicing for mRNA maturation. Both proteins share strong similarities in their domain structure with other eukaryotic poly(A) polymerases (see 5.1.1): The tripartite domain architecture of classical poly(A) polymerases consisting of the N-terminal catalytic domain, the central domain (also referred as PAP domain), and the C-terminal RNA-binding domain, which is conserved in both trypanosomatid paralogs (Martin et al., 2000). Both Tb927.3.3160 and Tb927.7.3780 lack a clearly defined nuclear localization signal (NLS); however, some basic residues can be found near the C-terminus of the Tb927.7.3780 protein, which may act as an NLS (Supplementary Figure 8.1; page 128-129). Interestingly, only Tb927.7.3780 is located in the nucleus, whereas the putative poly(A) polymerase paralog Tb927.3.3160 has a cytoplasmic localization, specifically within distinct spots (see 5.1.1). Despite the fact that only the Tb927.7.3780 protein showed in vitro as well as in vivo polyadenylation activity, this argues for functions of the Tb927.3.3160 protein other than 3´ end processing of pre-mRNAs (see 5.1.2 and 5.1.3). However, due to our negative results in in vitro polyadenylation, we cannot rule out that Tb927.3.3160 may represent a functional poly(A) polymerase because additional factors might be required for activity even if Mg²⁺ is replaced by Mn²⁺ (Christofori & Keller, 1988). Moreover, we cannot exclude that Tb927.3.3160 has a limited, specialized set of targets, which our in vivo polyadenylation assay could not detect. For example, Tb927.3.3160 could function in cytoplasmic polyadenylation or uridylation regulating translation or mRNA stability (Norbury, 2013). During cytoplasmic polyadenylation a poly(A) tail is re-added to a transcript, which was exported from the nucleus with a very short poly(A) tail preventing translation since PABP (poly(A)-binding protein) cannot bind. Cytoplasmic polyadenylation depends on two cis-acting elements, the PAS, which is bound by the CPSF complex, and a less conserved cytoplasmic polyadenylation element (CPE) with a UUUUUUA₁₃U motif, which is bound by the CPE-binding protein. As a result, re-addition of the poly(A) tail activates the transcript for translation by the ribosome. In contrast, cytoplasmic uridylation of polyadenylated mRNAs leads to decapping, transcriptional silencing and degradation of the mRNA.
6.2 Importance of cis splicing in trypanosomes

An open question is the importance of cis splicing in trypanosomes, since only two genes contain a single intron with a highly conserved sequence in trypanosomes and require cis splicing for mRNA maturation (Tschudi & Ullu, 1990, Mair et al., 2000, Liang et al., 2003, Berriman et al., 2005, Kolev et al., 2010, Siegel et al., 2010): A putative poly(A) polymerase, Tb927.3.3160, which is currently classified as non-functional in 3’ end processing (see 5.1.2 and 5.1.3) and an ATP-dependent DEAD box helicase Tb927.8.1510. However, RNAi-based depletion of Tb927.3.3160 (see 5.1.3) or Tb927.8.1510 had no effect on cell growth (Alsford et al., 2011). A recent study from our group supports the hypothesis that cis splicing may be an evolutionary relic (Preußer et al., 2014): Depletion of the U1 snRNP-specific U1C protein had no effect on cell growth and neither affected U1 snRNA levels nor the integrity of the U1 snRNP. U1C knockdown inhibited cis, but not trans splicing, which is in line with U1 snRNA binding at the 5’ cis splice site of both cis-spliced genes. Moreover, cis splicing was blocked with an antisense morpholino oligonucleotide that specifically binds the 5’ end of the U1 snRNA. After U1C depletion recovery from starvation stress was clearly impeded, indicating that cis splicing becomes essential only under stress conditions. Interestingly, Tb927.7.3780 is the major functional poly(A) polymerase in 3’ end processing, and for the intron-containing ATP-dependent DEAD box helicase Tb927.8.1510 an intronless putative counterpart may exist (Tb927.10.6630). However, this does not exclude that the two introns may fulfill an important, still unknown function, especially since the sequence is highly conserved in trypanosomes.

6.3 The trypanosomatid polyadenylation complex

Our identification and initial characterization of the components of the trypanosomatid polyadenylation complex provides an important initial step in understanding the evolution of mRNA processing mechanisms. There are considerable differences in the cis-acting polyadenylation elements between trypanosomes, yeast, and mammals. The mammalian polyadenylation signal consists of three elements (see 3.3): First, the highly conserved AAUAAA hexanucleotide PAS typically located 15 to 30 nts upstream of the cleavage site, which is in most cases a CA dinucleotide (Sheets et al., 1990, Beaudoin et al., 2000, Hu et al., 2005); second, the GU- (YGUGUUYY; Y=pyrimidine) or U-rich (UUUU) DSE within 20 nts downstream of the cleavage site (Hart et al., 1985, McLauchlan et al., 1985, Gil & Proudfoot, 1987, MacDonald et al., 1994, Hu et al., 2005); and third, the USE with a UGUA sequence, which is positioned 40 to 100 nts upstream of the cleavage site (Hu et al., 2005).

In contrast, the yeast signal is somewhat different in both its sequence and organization (Zhao et al., 1999, Mandel et al., 2008). It is less conserved and a minimum three elements are
needed. The UA-efficiency element (UAUAUA) is located at variable distances upstream of the cleavage site and is required for activation of the downstream positioning element. For functioning of the efficiency element, the first and the fifth U are most critical. However, this element only improves the efficiency of the cleavage reaction but is not essential. The A-rich positioning element is located approximately 20 nts upstream of the polyadenylation site and directs the cleavage site position. Mutations in this element disrupt only the position of the cleavage site, but not the efficiency of 3’ end processing, indicating that the function is less critical when compared to its mammalian PAS counterpart. Polyadenylation occurs in most cases at Y(A)n sequences (Y=pyrimidine) and the polyadenylation site is in comparison to humans less well defined, resulting in polyadenylation site clusters.

In trypanosomes the only known cis-acting sequence element is the polypyrimidine tract upstream of a trans splice site, which affects polyadenylation of the upstream gene and trans splicing of the downstream gene (LeBowitz et al., 1993, Matthews et al., 1994, Benz et al., 2005). The polyadenylation site is similar to yeast (see above) less well defined, resulting in polyadenylation site clusters as well (Siegel et al., 2011). These differences of cis-acting polyadenylation elements between trypanosomes, yeast, and mammals raised the question whether the protein factors involved in polyadenylation are correspondingly divergent in trypanosomes. For example, for some mammalian polyadenylation factors, such as CstF50, no counterpart could be identified in yeast (Takagaki & Manley, 1997).

In trypanosomes, only Fip1 and CPSF30 had previously been identified and functionally described so far (Hendriks et al., 2003, Bercovich et al., 2009). Specifically, CPSF30 is required for polycistronic mRNA processing (Hendriks et al., 2003). In addition, the CPSF73 ortholog associates with the U1 snRNP specific U1A protein in trypanosomes (Tkacz et al., 2010). Here we identified eight additional factors as part of the trypanosomatid polyadenylation machinery, based on mass spectrometry and database search: CPSF160, CPSF100, CPSF73, CstF64, CstF50, Symplekin, Tb927.11.13860 and Tb927.8.4480 (see 5.1.4). However, the poly(A) polymerase was not co-purified, which is in line with previous observations, suggesting that the poly(A) polymerase is not tightly associated with other processing factors; instead, there may be a dynamic or transient association with the core polyadenylation complex (Shi et al., 2009). Even if the sequence similarity between these trypanosome polyadenylation factors and their human counterparts is low, they share for the most part their domain architectures. This becomes particularly obvious when comparing the conservation of the subunits of the CPSF complex (see 5.1.4). For example, the CPSF73 subunit, which attracted special attention as the functional endonuclease in the mammalian polyadenylation reaction, shares specific, catalytically important residues with the trypanosome homolog: The zinc-binding residues (mainly histidines) are
6. Discussion

conserved in the trypanosomatid CPSF73, indicating that CPSF73 likely also acts as the endonuclease during polyadenylation in trypanosomes (Mandel et al., 2006). Moreover, the domain structures of T. brucei CPSF30 and its mammalian homolog are strongly conserved, harboring five C3H1 zinc fingers and a C-terminal C2HC zinc finger. Human CPSF30 (together with WDR33) is required for PAS binding, and it was shown that CPSF30 binds poly(U) sequences in vitro (Barabino et al., 1997, Chan et al., 2014, Schönemann et al., 2014, Shimberg et al., 2016). These in vitro binding properties for poly(U) sequences were observed for trypanosomatid CPSF30 in our SELEX-seq experiment as well, arguing for similar RNA-binding properties of both proteins (see 5.1.5).

Some of the new trypanosome factors display only low sequence conservation. The putative trypanosome CstF64 strikingly differs from the human counterpart (compare the molecular masses of 14 kDa vs. 64 kDa for the T. brucei and the human CstF64, respectively) and could only be identified through a putative hinge domain required for CstF64 interaction with CstF77 and Symplekin (Takagaki & Manley, 2000, Ruepp et al., 2011b). The otherwise conserved N-terminal RRM domain, however, is completely missing in trypanosomes. Therefore the trypanosomatid CstF64 may have retained only function as a scaffolding protein required for the recruitment of other 3´ end processing factors. Moreover, we cannot exclude that the functions have been moved to another, still unknown factor or may be not essential in trypanosomes anymore.

Finally, some known mammalian polyadenylation factors could be identified neither in our mass spectrometry approach nor through extensive database homology search: As an example, we did not find a trypanosome homolog of WDR33, a subunit of the CPSF complex. Independent studies had recently identified WDR33 as the factor that together with CPSF30 recognizes the AAUAAA polyadenylation signal (Chan et al., 2014, Schönemann et al., 2014, Shimberg et al., 2016). Since trypanosomes lack the classical hexameric polyadenylation signal, this appears plausible. In sum, our results argue for a general conservation of polyadenylation factors among eukaryotes.

6.4 Linkage of polyadenylation and trans splicing

Depletion of the individual factors of the polyadenylation complex identified in our study showed severe growth defects and inhibition of polyadenylation in most cases (see 5.1.6). Interestingly, the two hypothetical proteins, Tb927.11.13860 and Tb927.8.4480, of which no known homolog or conserved domain structure could be identified, activate or repress polyadenylation, respectively. Such effects have been described in higher eukaryotes for factors that are not constituents of the core polyadenylation complex, but are required for regulation of
polyadenylation. As an example, 14-3-3ε interacts directly with the C-terminus of the poly(A) polymerase and inhibits the polyadenylation activity, whereas the transcription activator GAL4 stimulates polyadenylation via an interaction with PAF1c (Kim et al., 2003, Nagaike et al., 2011). In addition, we observed trans splicing defects consistent with close coupling of both mRNA processing reactions in trypanosomes. Interestingly, we did not detect for all factors effects on polyadenylation and trans splicing: For example, knockdown of cleavage factor I m25 (CFIm25) neither affected poly(A) tail length nor splicing. In higher eukaryotes CFIm25 is involved in the regulation of alternative polyadenylation, especially in the repression of proximal polyadenylation sites leading to longer 3´ UTRs (Gruber et al., 2012). Since alternative polyadenylation is common in trypanosomes as well, CFIm25 identified here, which also contains a NUDIX domain required for USE binding in humans, is a potential regulator of this process (Kolev et al., 2010, Siegel et al., 2010, Yang et al., 2010).

In other eukaryotes, a direct linkage between spliceosomal and 3´ end processing components had been demonstrated by several studies (Vagner et al., 2000, Kyburz et al., 2006, Millevoi et al., 2006). For example, U2AF65, which binds the polypyrimidine tract during splicing, interacts with CFIm59 (via an arginine-serine-rich region in both proteins) and stimulates 3´ end cleavage and polyadenylation. Moreover, the poly(A) polymerase interacts with U2AF65 and increases the binding affinity of the latter one to the polypyrimidine tract, which in turn enhances splicing. In addition, the U2 snRNP interacts via SF3b with the CPSF complex, and CPSF stimulates splicing, whereas the U2 snRNP activates mRNA cleavage. In trypanosomes, so far only the polypyrimidine tract is known to affect trans splicing and polyadenylation of two adjacent genes (LeBowitz et al., 1993, Matthews et al., 1994, Benz et al., 2005). Until recently no factor could be identified, which directly links the polyadenylation complex and the trans-spliceosome in trypanosomes. This is in line with our result that no known spliceosomal factor co-purified with CPSF160. However, a recent study provided first evidence that the polyadenylation factor CPSF73 copurifies with the U1A protein, a specific component of the spliceosomal U1 snRNP (Tkacz et al., 2010). The U1A protein was previously shown to autoregulate its own expression by binding of two U1A proteins to its own mRNA, regulating poly(A) polymerase activity in higher eukaryotes (Gunderson et al., 1994). In addition, the U1A protein interacts with CPSF160 and can generally stimulate polyadenylation (Lutz et al., 1996). Moreover, the whole U1 snRNP can inhibit polyadenylation due to an interaction between U1-70K and the poly(A) polymerase and is also involved in blocking cryptic polyadenylation signals in introns, thereby preventing premature cleavage and polyadenylation (Gunderson et al., 1998, Kaida et al., 2010, Berg et al., 2012). Note that the spliceosomal U1 snRNP was thought to be required only for cis splicing, not for trans splicing in trypanosomes. However, we recently reported that the U1 snRNP may also
associate with the trans-spliceosome, indicating a physical linkage between cis- and trans-splicing (Preußer et al., 2014). Taken together this suggests that the U1 snRNP functions in both splicing and polyadenylation in trypanosomes, likely linking both processes as described for other eukaryotes before.

6.5 Characterization of the DRBD4 RNA binding motif
We initially characterized in vivo RNA binding of DRBD4 by iCLIP, which demonstrated a strong preference for UTR regions (see 5.2.1). Earlier studies had mapped DRBD3, a close paralog of DRBD4, in 3’ UTR regions, underlining the important regulatory role of gene UTRs (Das et al., 2015). The DRBD4 RNA-binding motif was further evaluated by in vitro SELEX-seq, followed by GST pulldown assays of short transcripts for validation: The DRBD4 RNA-binding motif is purine-, especially adenosine-rich, containing AUGA elements (see 5.2.2). In contrast, polypyrimidine stretches were deplet during the SELEX rounds of selection.

This was surprising, because human PTB binds, as indicated by the name, polypyrimidine stretches: First, two studies identified UCUUC and CUCUCU RNA-binding motifs, and second, follow-up experiments identified the RNA-binding motif of each RRM with four different consensus motifs (RRM1: YCU; RRM2: CU(N)N, RRM3: YCUNN, RRM4: YCN (1995, Perez et al., 1997, Oberstrass et al., 2005, Reid et al., 2009). Third, in vivo CLIP delineated a UYUYU consensus motif for PTB (Xue et al., 2009). In addition, a PTB-like, pyrimidine-rich binding motif had been identified for the second PTB homolog in trypanosomes, DRBD3 (Das et al., 2015).

However, one has to consider that the sequence homology between DRBD4 and PTB is low and subtle changes in amino acids involved in RNA binding may change specificity (Maris et al., 2005)

6.6 DRBD4-dependent poly(A) site choice
In trypanosomes gene expression is controlled mainly by post-transcriptional mechanisms (Clayton, 2014). Recent genome-wide approaches showed that most genes contain several trans splice and polyadenylation sites, indicating that the regulation of both processes provides another, still not very well explored level of post-transcriptional gene regulation in trypanosomes (Kolev et al., 2010, Nilsson et al., 2010, Siegel et al., 2010, Veitch et al., 2010).

We focussed on DRBD4 binding upstream of activated and repressed polyadenylation sites, as an attempt to derive a predictive map for this factor. Our results provide evidence for a regulatory role of DRBD4 in poly(A) site choice in T. brucei (see 5.2.3 and 5.2.4), although we cannot recognize distinct features of DRBD4 binding sites or positioning that would associate with poly(A) site activation versus repression. Position-dependent effects on alternative splicing have been described for splicing regulators in humans (Rossbach et al., 2014): hnRNP L binds
at repressed exons preferential upstream of the 3´ splice site (peak at position -40), whereas hnRNP L activates exon splicing when bound in the intron further downstream of the 5´ splice site (positions +25 to +200). In sum, it was proposed that hnRNP L represses 3´ splice sites by blocking the recognition of the polypyrimidine tract and 3´ splice site region, and activates 5´ splice sites from downstream intronic positions.

However, in case of DRBD4-regulated polyadenylation additional protein factors may be required, or there are distinctive changes in the sequence elements involved. Interestingly, among the DRBD4-associated proteins we found several constituents of the trypanosomatid polyadenylation complex (CPSF160, CPSF100, CPSF73, CstF50 and Symplekin) including the functional poly(A) polymerase [Tb927.7.3780] (see 5.2.5). Moreover, trypanosomatid PABP2 the homolog of human PABPN1, which stimulates the poly(A) polymerase for efficient polyadenylation and also regulates APA, was co-purified (Wahle, 1991a, Bienroth et al., 1993, Wahle, 1995b, Janal et al., 2012). The association of DRBD4 and PABP2 may provide a first protein factor that could be involved in a combinatorial regulation of polyadenylation as observed for human PTB (see below).

However, further analyses are required to characterize the biochemical mechanisms and protein interactions involved in DRBD4-dependent regulation of polyadenylation. The repression of a polyadenylation site, as observed for a cryptic site in the 3´ UTR of Tb927.2.6000 (see 5.2.4) can be mediated in principle by blocking the assembly of single or multiple constituents of the polyadenylation complex. As an example, PTB and CstF64 can compete for binding to the downstream GU/U-rich sequence element (Castelo-Branco et al., 2004). In addition, it was reported that the U1 snRNP can block cryptic polyadenylation signals in introns, thereby preventing premature cleavage and polyadenylation (Kaida et al., 2010, Berg et al., 2012). Stimulatory effects on polyadenylation site choice, as observed in case of the gene Tb927.11.5440 (see 5.2.4), may be caused by recruitment of polyadenylation complex constituents like CstF or the poly(A) polymerase or by directly stimulating poly(A) polymerase activity as observed for PTB (Millevoi et al., 2009).

6.7 Alternative polyadenylation in T. brucei
In other eukaryotes, alternative polyadenylation is often associated with proliferation and cellular differentiation (Elkon et al., 2013). For example, activation of T-cell proliferation is linked to proximal poly(A) site usage as well as general shortening of 3´ UTRs in cancer cells (Sandberg et al., 2008, Mayr & Bartel, 2009, Lin et al., 2012, Morris et al., 2012). Most likely this correlates with the deletion of binding sites for miRNAs or trans-acting proteins that can negatively affect mRNA stability or translation efficiency. Moreover, many tissues show a specific bias in poly(A)
site choice, for instance a preference for distal poly(A) sites in brain and nervous system (Zhang et al., 2005). In procyclic trypanosomes, mainly the distal of the two polyadenylation sites in the transcript of Tb927.11.5440 (malic enzyme) is used, resulting in an almost exclusive expression of the long mRNA isoform: Upon DRBD4 depletion, the distal poly(A) site is down-regulated, whereas the proximal one is up-regulated, causing a shift to the short isoform (see 5.2.4). Interestingly, in the bloodstream form, the opposite can be observed, that the short Tb927.11.5440 isoform is predominantly expressed (see 5.2.4). In addition, a previous study showed that expression of Tb927.11.5440 is developmentally regulated, with malic enzyme levels higher in the procyclic than the bloodstream form (Leroux et al., 2011). Therefore the DRBD4-dependent poly(A) site choice and isoform expression in procyclic cells may be directly linked to differential regulation of developmental genes in the trypanosomatid life-cycle.

6.8 Further perspectives
Here we initially identified the functional poly(A) polymerase (see 5.1.1, 5.1.2 and 5.1.3) and the constituents of the polyadenylation complex in T. brucei (see 5.1.4). Moreover, we could show that depletion of most polyadenylation factors impaired both polyadenylation and trans splicing in vivo, arguing for close coupling of these processes (see 5.1.6). However, further work is required to study the polyadenylation complex in trypanosomes in more details. For example, RNA binding of the complex remains still elusive, since the canonical cis-acting elements (PAS and DSE) and the WDR33 protein, which directly binds the PAS, are missing in trypanosomes (Proudfoot, 2011, Chan et al., 2014, Schönemann et al., 2014). Moreover, trypanosomatid CstF64 lacks the N-terminal RRM required for DSE recognition and was only identified through a putative hinge domain (see 5.1.4) (Takagaki et al., 1992, MacDonald et al., 1994, Takagaki & Manley, 1997, Perez Canadillas & Varani, 2003). Most likely the best candidate to analyze RNA binding of the T. brucei polyadenylation complex is CPSF30, since the mammalian protein is involved in PAS binding and the domain structure is highly conserved between the human and the trypanosomatid proteins (Chan et al., 2014, Schönemann et al., 2014, Shimberg et al., 2016). We initially identified a poly(U) in vitro binding motif for T. brucei CPSF30 in our SELEX-seq experiment, which was also observed for the mammalian protein (Barabino et al., 1997). However, this motif needs to be further analyzed biochemically by GST pulldown experiments of various RNA transcripts. To generally investigate in vivo RNA binding of the polyadenylation complex, CPSF30 iCLIP experiments can be used, and correlated with poly(A) sites identified in various RNA-seq datasets.
CPSF73 is the endonuclease in the cleavage and polyadenylation reaction and the domain structures of both the human and the trypanosomatid proteins are highly conserved, including the zinc-binding residues (mainly histidines), which are required for cleavage activity (Mandel et al., 2006). To assay the trypanosomatid protein for cleavage activity, CPSF73 can be recombinantly expressed, purified and tested for in vitro cleavage. In addition, the conserved residues for zinc-binding can be mutated to further analyze the catalytic mechanism.

Moreover, the protein-protein interactions between the identified factors of the T. brucei polyadenylation complex remain elusive, but they have been well characterized for the mammalian system (see 3.3) (Chan et al., 2011, Xiang et al., 2014). Since the domain structures of most mammalian and trypanosomatid proteins are highly conserved (see 5.1.4) it appears likely that the interaction between the proteins is conserved as well. Indeed, the interaction between Fip1 and CPSF30 has been described in both trypanosomes and mammals (Bercovich et al., 2009, Kaufmann et al., 2004). To further identify the protein-protein interactions in the trypanosomatid polyadenylation complex, the factors can be recombinantly expressed, purified and tested for in vitro binding in pulldown experiments.

Furthermore, we identified the trypanosomatid PTB homolog DRBD4 as first trans-acting regulator of polyadenylation. DRBD4 binds upstream of regulated polyadenylation sites, has a purine, in particular adenosine, -rich binding motif and is required for both activation and repression of polyadenylation sites affecting general transcript and isoform expression levels (see 5.2.1, 5.2.2, 5.2.3 and 5.2.4).

To further analyze the DRBD4 RNA-binding motif and address the influence of single RRMs, truncated versions of the protein lacking individual RRMs can be recombinantly expressed and used in both SELEX-seq and pulldown assays.

We analyzed the DRBD4-dependent poly(A) site choice only in procyclic trypanosomes and have evidence for a life-cycle-dependent regulation, as observed for the differential isoform expression of Tb927.11.5440 (malic enzyme). For further characterization of a life-cycle-dependent regulation, DRBD4 binding sites need to be identified in the bloodstream form by iCLIP and correlated with poly(A) site choice after DRBD4 depletion or overexpression.

Another putative regulator of poly(A) site choice is the trypanosomatid CFIm25 homolog, since the mammalian protein is involved in the repression of proximal polyadenylation sites leading to longer 3’ UTRs (Gruber et al., 2012). To assess CFIm25-dependent poly(A) site choice in T. brucei, the same combinatorial iCLIP/RNA-seq strategy as applied here can be used.

The cross-regulation between trans splicing and polyadenylation is still an open question in trypanosomes. To analyze a trans splice-dependent polyadenylation site choice, splicing can be blocked by various methods (knockdown of splicing factors, antisense morpholino
oligonucleotides for against specific snRNAs and Sinefungin treatment) and poly(A) site usage can be assayed by RNA-seq.
7 References


7. References


References


References


## 8 Supplementary Information

### 8.1 Figures and Tables

#### A

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPPFVTQGQTQPQP</td>
<td>EKTLESMAIR</td>
</tr>
<tr>
<td></td>
<td>PVIKNSIKLRLNR 74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVQRHSLN</td>
<td>TDBQWARAHPGCSLGA-VAELKERAVALKERSVTDIDPQPSEP----ELKYNSNGVS</td>
</tr>
<tr>
<td></td>
<td>WRESIEKNLPSVIENVGKTETGNYGTWGTKDRCLVAMHRHDSKPETSYDKKDCSEVKLRA 148</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVAEQVLYNESEMPFYTPAFFAPAYPAHASSRPGCWEALYSETPTVCGLTLART---IPLEEM 336</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPRFSSMPSKSKCIKCCPFEPAVAGASSPGKGEALYSETPTVCGLTLART---IPLEEM</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GESCLDS</td>
<td>EELMVKKSSPFALLRSSAGKRFHHEFFTTS</td>
</tr>
<tr>
<td></td>
<td>452</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSGTSSEPS</td>
<td>EKTPFAISPPPKPTSVSVSTRLWNPFRSSGNGATSGNAATKIPFPIGVKRTSHPKE 658</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tb927.3.3160</th>
<th>PAPOLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESKKKTXTETELANLTSEKTARVTETSTTQSETIQTAASLLASCQKTSSTLDSDIPALPANPI</td>
<td>732</td>
</tr>
</tbody>
</table>

---

8. Supplement
Supplementary Figure 8.1 Alignment of the human poly(A) polymerase (PAPOLA) to the two putative T. brucei poly(A) polymerases, (A) Tb927.3.3160, and (B) Tb927.7.3780.

Residues (page 128-129) that are conserved or have a similar physicochemical character in more than 50% of sequences are highlighted in black and gray, respectively. The predicted secondary structures of PAP from T. brucei and H. sapiens are shown above and below the alignments, respectively. α-helices are represented as tubes, and β-strands as arrows, while disordered regions are shown as “~”.

---

8. Supplement
Figure 8.2 DRBD4 RNA binding around regulated poly(A) sites.
Accumulated DRBD4-YFP iCLIP crosslink-tag counts were plotted in the region between -200 to +200 nts relative to the poly(A) site of the down- and up-regulated clusters (top: down-regulated n=234; background n=936; bottom: up-regulated n=248; background n=992) and are depicted as solid lines. As background control, unaffected polyadenylation sites were used (dashed lines). Crosslink-tag counts were plotted with fitted values (21-nts windows) on the Y-axis and positions on the X-axis.
Supplementary Table 8.1 CPSF160-associated proteins identified by mass spectrometry after tandem affinity purification.

Extracts were prepared from *T. brucei* cell lines that stably express tagged CPSF160 (in two biological replicates; experiment 1/2), followed by TAP-tag affinity purification. Proteins after the second purification step (protein C elution) were identified by mass spectrometry. Each protein is described by name (factor), the TriTrypDB annotation number (http://tritrypdb.org/tritrypdb/), its molecular mass (in kDa), the number of exclusive unique peptide counts obtained by mass spectrometry. As control, *T. brucei* wildtype cells were used.

<table>
<thead>
<tr>
<th>Database Name: the T_brucie database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version: unknown</td>
</tr>
<tr>
<td>Taxonomy: All Entries</td>
</tr>
<tr>
<td>Number of Proteins: 38057</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search Engine Set: 1 Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search Engine: Mascot</td>
</tr>
<tr>
<td>Version: 2.3.02</td>
</tr>
<tr>
<td>Samples: All Samples</td>
</tr>
<tr>
<td>Fragment Tolerance: 0.60 Da (Monoisotopic)</td>
</tr>
<tr>
<td>Parent Tolerance: 10.0 PPM (Monoisotopic)</td>
</tr>
<tr>
<td>Fixed Modifications:</td>
</tr>
<tr>
<td>Variable Modifications: +16 on M (Oxidation), +57 on C (Carbamidomethyl)</td>
</tr>
<tr>
<td>Database: the T_brucie database (unknown version, 38057 entries)</td>
</tr>
<tr>
<td>Digestion Enzyme: Trypsin</td>
</tr>
<tr>
<td>Max Missed Cleavages: 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scaffold: Version: Scaffold_4.4.1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modification Metadata Set: 1541 modifications</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comment:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Grouping Strategy: Experiment-wide grouping with binary peptide-protein weights</td>
</tr>
<tr>
<td>Peptide Thresholds: 95.0% minimum</td>
</tr>
<tr>
<td>Protein Thresholds: 95.0% minimum and 3 peptides minimum</td>
</tr>
<tr>
<td>Peptide FDR: 0.7% (Prophet)</td>
</tr>
<tr>
<td>Protein FDR: 0.0% (Prophet)</td>
</tr>
<tr>
<td>#</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>27</td>
</tr>
<tr>
<td>28</td>
</tr>
<tr>
<td>29</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>31</td>
</tr>
<tr>
<td>32</td>
</tr>
<tr>
<td>33</td>
</tr>
<tr>
<td>34</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>36</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>37</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>39</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>42</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>45</td>
</tr>
<tr>
<td>46</td>
</tr>
<tr>
<td>47</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>51</td>
</tr>
</tbody>
</table>
8.2 Abbreviations and Symbols

\[\sim\] approximately
\[<\] less than
\[>\] more than
\[+\] plus
\[-\] minus
\[/\] percent
\[\times\] times
\[\times g\] times gravity
\[\degree C\] degree Celsius
\[\mu\] micro
\[2′\] bonds to the C-2 carbon of ribose in nucleic acids
\[3′\] directionality in nucleic acids: in the direction of the C-3 ribose carbon
\[5′\] directionality in nucleic acids: in the direction of the C-5 ribose carbon
\[\alpha\] alpha
\[\beta\] beta
\[\epsilon\] epsilon
\[\gamma\] gamma
\[\Delta\] delta, without / lacking
\[A\] adenosine
\[aa\] amino acids
\[AcTEV\] tobacco etch virus protease
\[APA\] alternative polyadenylation
\[APS\] ammonium persulfate
\[ARM\] armadillo-type fold
\[As\] anti-sense
\[ATP\] adenosine triphosphate
\[\beta\text{-}\text{casp}\] beta-caspase domain
\[Bdf3\] bomodomain factor 3 protein
\[bp\] base pairs
\[BP\] branch point
\[BSA\] bovine serum albumin
\[C\] cytidine
\[cDNA\] complementary DNA
\[CF\] cleavage factor
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ci</td>
<td>Curie</td>
</tr>
<tr>
<td>CLIP</td>
<td>cross-linking and immunoprecipitation</td>
</tr>
<tr>
<td>CPSF</td>
<td>cleavage and polyadenylation specificity factor</td>
</tr>
<tr>
<td>CstF</td>
<td>cleavage stimulation factor</td>
</tr>
<tr>
<td>CstF1-dd</td>
<td>cleavage stimulation factor subunit 1, dimerization domain</td>
</tr>
<tr>
<td>CstF64-C</td>
<td>cleavage stimulating factor C-terminal domain</td>
</tr>
<tr>
<td>C_T</td>
<td>cycle of threshold</td>
</tr>
<tr>
<td>CTD</td>
<td>carboxyl-terminal domain of RNA polymerase II</td>
</tr>
<tr>
<td>C-terminal</td>
<td>carboxyl-terminal</td>
</tr>
<tr>
<td>CTP</td>
<td>cytidine triphosphate</td>
</tr>
<tr>
<td>ctr</td>
<td>control</td>
</tr>
<tr>
<td>DAPI</td>
<td>4′,6-diamidino-2-phenylindole</td>
</tr>
<tr>
<td>dCTP</td>
<td>deoxycytidine triphosphate</td>
</tr>
<tr>
<td>DIG</td>
<td>digoxigenin</td>
</tr>
<tr>
<td>DMPC</td>
<td>dimethyl pyrocarbonate</td>
</tr>
<tr>
<td>DMSO</td>
<td>dimethyl sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleoside triphosphate</td>
</tr>
<tr>
<td>Dox</td>
<td>doxycyclin</td>
</tr>
<tr>
<td>DRBD</td>
<td>double-stranded RNA-binding domain</td>
</tr>
<tr>
<td>DSE</td>
<td>downstream sequence element</td>
</tr>
<tr>
<td>dT</td>
<td>deoxythymidine triphosphate</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>E. coli</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>e.g.</td>
<td><em>exempli gratia</em> (for the sake of example)</td>
</tr>
<tr>
<td>EGTA</td>
<td>ethyleneglycoltetraacetic acid</td>
</tr>
<tr>
<td><em>et al.</em></td>
<td><em>et alii</em> (and others)</td>
</tr>
<tr>
<td>f</td>
<td>femto</td>
</tr>
<tr>
<td>FBS</td>
<td>fetal bovine serum</td>
</tr>
<tr>
<td>fw.</td>
<td>forward (primer)</td>
</tr>
<tr>
<td>g</td>
<td>gram</td>
</tr>
<tr>
<td>G</td>
<td>guanosine</td>
</tr>
<tr>
<td>GST</td>
<td>glutathione S-transferase</td>
</tr>
<tr>
<td>GTP</td>
<td>guanosine triphosphate</td>
</tr>
</tbody>
</table>
h
HEPES
His
hnRNP
hPAPOLA
HPLC
H. sapiens
iCLIP
IgG
IPTG
k
kb
kd
kDa
l
L-
LB
LC-MS/MS
M
m
M
mA
min
miRNA
mRNA
MS
(n)
n
N
NCBI
NLS
NP-40
nPTB
N-terminal
nts
hour(s)
N-2-hydroxyethylpiperazine
polyhistidine-tag
heterogeneous ribonucleoprotein particle
human poly(A) polymerase alpha
high-performance liquid chromatography
Homo sapiens
individual-nucleotide resolution cross-linking and immunoprecipitation
immunoglobulin G
isopropyl-β-D-thiogalactopyranosid
kilo
kilobases
knockdown
kilodalton
liter
L-confirmation
Luria/Miller
liquid chromatography-tandem mass spectrometry
marker
milli
molar
milliampere
minute(s)
micro RNA
messenger RNA
mass spectrometry
poly
nano
random / any nucleotide
National center for biotechnology information
nuclear localization signal
Nonidet P-40
neuronal polypyrimidine tract binding protein
amino-terminal
nucleotides
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUDIX</td>
<td>NUDIX hydrolase domain</td>
</tr>
<tr>
<td>OH</td>
<td>hydroxyl (-group)</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>p</td>
<td>pico</td>
</tr>
<tr>
<td>P, p</td>
<td>phosphate</td>
</tr>
<tr>
<td>PABPN1</td>
<td>nuclear poly(A)-binding protein 1</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>PAP</td>
<td>poly(A) polymerase</td>
</tr>
<tr>
<td>PAS</td>
<td>polyadenylation signal</td>
</tr>
<tr>
<td>PBS</td>
<td>phosphate-buffered saline</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PK</td>
<td>proteinase K</td>
</tr>
<tr>
<td>PNK</td>
<td>polynucleotide kinase</td>
</tr>
<tr>
<td>Pol II</td>
<td>DNA-dependent RNA-polymerases II</td>
</tr>
<tr>
<td>PPIase</td>
<td>prolyl isomerase</td>
</tr>
<tr>
<td>PPT</td>
<td>polypyrimidine tract</td>
</tr>
<tr>
<td>PRP8</td>
<td>pre-mRNA processing factor 8</td>
</tr>
<tr>
<td>PTB</td>
<td>polypyrimidine tract binding protein</td>
</tr>
<tr>
<td>PTP</td>
<td>protein C epitope / TEV cleavage site / 2× protein A epitope</td>
</tr>
<tr>
<td>PTU</td>
<td>polycistronic transcription unit</td>
</tr>
<tr>
<td>PVA</td>
<td>polyvinyl alcohol</td>
</tr>
<tr>
<td>®</td>
<td>registered trade mark</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>RNP</td>
<td>ribonucleoprotein (-particle)</td>
</tr>
<tr>
<td>RRM</td>
<td>RNA recognition motif</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>reverse transcription</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>quantitative RT-PCR</td>
</tr>
<tr>
<td>Rv.</td>
<td>reverse (primer)</td>
</tr>
<tr>
<td>S</td>
<td>Svedberg; sedimentation coefficient</td>
</tr>
<tr>
<td>SAP</td>
<td>shrimp alkaline phosphatase</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulfate</td>
</tr>
<tr>
<td>sec</td>
<td>second</td>
</tr>
<tr>
<td>SELEX-seq</td>
<td>systematic evolution of ligands by exponential enrichment-sequencing</td>
</tr>
</tbody>
</table>
seq sequencing
SL spliced-leader
snRNA small nuclear RNA
snRNP small nuclear ribonucleoprotein (-particle)
ss splice site
SSR strand switch region
T thymidine
t time point
T4 enterobacteria phage T4
T7 bacteriophage T7
TAP tandem affinity purification
TBE Tris-Borate-EDTA
T. brucei Trypanosoma brucei
TEMED N,N,N',N'-tetramethylenediamine
TM Trade Mark
Tris Tris-hydroxymethylaminomethane
tRNA transfer RNA
TSS transcription start site
TTS transcription termination site
Tween 20 Polyoxyethylene (20) sorbitan monolaurate
U uridine
U unit
USE upstream sequence element
UTP uridine triphosphate
UTR untranslated region
UV ultraviolet
V volt
V A, G or C
W watt
WD40 repeat tryptophan-aspartate repeat
WT wildtype
Y pyrimidine
The curriculum vitae was removed from the electronic version of the paper.
10 Danksagung

Ich danke zunächst Prof. Dr. Albrecht Bindereif für die Möglichkeit diese Arbeit in seinem Labor anfertigen zu können. Vor allem möchte ich mich für die sehr gute Betreuung bei allen theoretischen und praktischen Fragestellungen bedanken. Außerdem möchte ich mich herzlich bei meinem IRTG Co-Supervisor Prof. Dr. Roland Hartmann bedanken, der zudem das Zweitgutachten für diese Arbeit angefertigt hat.


Moreover, I want to thank Prof. Dr. Mark Carrington for the opportunity to participate in the DRBD4 project and to visit his laboratory in Cambridge (UK). In addition, I want to thank Janina from the Carrington lab, who helped me with Glyoxal-Northern blotting and performed essential experiments. Finally, I want to thank Dr. Steven Kelly, who analyzed the DRBD4 RNA-seq data.

Großer Dank gilt natürlich auch allen anderen Kollegen aus dem Labor, ohne die es nicht so viel Spaß gemacht hätte, wenn es einmal nicht so rund lief: Silke, die das Labor am Laufen erhält und für alles einen guten Rat hat, Oliver, Tim, Isabelle, Stefan und den Ehemaligen Heinrich und Patrick. Außerdem danke ich natürlich auch dem IRTG, das mir unter anderem den Laboraufenthalt in England sowie die Teilnahme an zahlreichen Konferenzen und Fortbildungen ermöglicht hat.

Vielen Dank auch allen Freunden aus benachbarten Arbeitsgruppen: Dennis, Lennart, Bernhard und vielen anderen für all die lustigen Abende.

Ein ganz besonderer Dank gilt natürlich meiner gesamten Familie, die mich immer unterstützt hat.

Und zu guter Letzt danke an: RSC, Fightclub, BK, BB, BG, Pizza-Nord sowie das zugehörige PL, UVP, Marböbchen, Ritzcarlton, Abaum, Sowiesx, Brick, Klim, Börse, Express, Affen, Ulo, Scara, Harlem, Aramäer Klaus, Tante Emma, Tegue und die Göttin der Weisheit: Öttexia -nebst ihrer anderen zahlreichen Erscheinungen.
11. Eidesstattliche Erklärung


Regarding the collaborative contributions of other scientists: PTP-tagged poly(A) polymerase and CPSF160 cell lines were generated by Christian Preußer. Recombinant poly(A) polymerases were expressed, purified and assayed for *in vitro* activity by Christian Preußer. RNAi-mediated knockdown of the poly(A) polymerases and following *in vivo* polyadenylation assays were performed by Christian Preußer. CPSF160-PTP TAP was done by Christian Preußer and mass spectrometry was carried out by Monika Raabe and Henning Urlaub (MPI Göttingen). DRBD4 knockdown and RNA-seq, including data analysis were performed by Marc Carrington and Steven Kelly (University of Cambridge; University of Oxford, respectively). Analysis of iCLIP and SELEX datasets as well as the integration of iCLIP and RNA-seq datasets were done by Lee-Hsueh Hung.

________________________

Gießen, den 7.7.2016

Henrik Koch