



**Development of tools for the study of gene regulation in**  
***Trypanosoma brucei***

**Entwicklung neuer Methoden zur Untersuchung der**  
**Genregulation in *Trypanosoma brucei***

Doctoral thesis for a doctoral degree  
at the Graduate School of Life Sciences,  
Julius-Maximilians-Universität Würzburg,

Section Infection and Immunity

Submitted by

**Juan José Vásquez Ospina**

From

**Bogotá, Colombia**

Würzburg, 2016

Submitted on: .....

(Office stamp)

**Members of the *Promotionskomitee*:**

Chairperson: Prof. Dr. Markus Engstler

Primary Supervisor: Dr. Tim Nicolai Siegel

Supervisor (Second): Prof. Dr. Christian Janzen

Supervisor (Third): Dr. Ana Eulalio

Supervisor (Fourth): Priv. Doz. Dr. rer. nat. Sandra Hake

Date of Public Defence: .....

Date of Receipt of Certificates: .....

## **Affidavit**

I hereby confirm that my thesis entitled “Development of tools for the study of gene regulation in *Trypanosoma brucei*” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, date

Signature

## **Eidesstattliche Erklärung**

Hiermit erkläre ich an Eides statt, die Dissertation “Entwicklung neuer Methoden zur Untersuchung der Genregulation in *Trypanosoma brucei*” eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorlegen hat.

Ort, Datum

Unterschrift

# Development of tools for the study of gene regulation in *Trypanosoma brucei*

## Summary

The protozoan parasite *Trypanosoma brucei* is the causal agent of sleeping sickness and besides its epidemiological importance it has been used as model organism for the study of many aspects of cellular and molecular biology especially the post-transcriptional control of gene expression.

Several studies in the last 30 years have shown the importance of mRNA processing and stability for gene regulation. In *T. brucei* genes are unusually arranged in polycistronic transcription units (PTUs) and a coupled process of trans-splicing and polyadenylation produces the mature mRNAs. Both processes, mRNA processing and stability, cannot completely explain the control of gene expression in the different life cycle stages analyzed in *T. brucei* so far.

In recent years, the relevance of expression regulation at the level of translation has become evident in other eukaryotes. Therefore, in the first part of my thesis I studied the impact of translational regulation by means of a genome-wide ribosome profiling approach. My data suggest that translational efficiencies vary between life cycle stages of the parasite as well as between genes within one life cycle stage. Furthermore, using ribosome profiling I was able to identify many new putative un-annotated coding sequences and to evaluate the coding potential of upstream open reading frames (uORF). Comparing my results with previously published proteomic and RNA interference (RNAi) target sequencing



(RIT-seq) datasets allowed me to validate some of the new coding sequences and to evaluate their relevance for the fitness of the parasite.

In the second part of my thesis I used the transcriptomic and translomic profiles obtained from the ribosome profiling analysis for the identification of putative non-coding RNAs (ncRNAs). These results led to the analysis of the coding potential in the regions upstream and downstream of the expressed variant surface glycoprotein (*VSG*), which is outlined in the third part of the results section. The region upstream of the *VSG*, the co-transposed region (CTR), has been implicated in an increase of the *in situ* switching rate upon its deletion. The ribosome profiling results indicated moderate transcription but not translation in this region. These results raised the possibility that the CTR may be transcribed into ncRNA. Therefore, in the third part of my thesis, I performed a primary characterization of the CTR-derived transcripts based on northern blotting and RACE. The results suggested the presence of a unique transcript species of about 1,200 nucleotides (nt) and polyadenylated at the 3'-end of the sequence.

The deletion of the CTR sequence promoting and increase of the *in situ* switching rates was performed around 20 years ago by means of inserting reporter genes. With the recent development of endonuclease-based tools for genome editing, it is now possible to delete sequences in a marker-free way. In the fourth part of my thesis, I show the results on the implementation of the highly efficient genome-editing CRISPR-Cas9 system in *T. brucei* using episomes. As a proof of principle, I inserted the sequence coding for the enhanced green fluorescent protein (eGFP) at the end of the SCD6 coding sequence (CDS).

Fluorescent cells were observed as early as two days after transfection. Therefore, after the successful set up of the CRISPR-Cas9 system it will be possible to modify genomic regions with more relevance for the biology of the parasite, such as the substitution of codons present in gene tandem arrays.

The implementation of ribosome profiling in *T. brucei* opens the opportunity for the study of translational regulation in a genome-wide scale, the re-annotation of the currently available genome, the search for new putative coding sequences, the detection of putative ncRNAs, the evaluation of the coding potential in uORFs and the role of untranslated regions (UTRs) in the regulation of translation. In turn, the implementation of the CRISPR-Cas9 system offers the possibility to manipulate the genome of the parasite at a nucleotide resolution and without the need of including resistant makers. The CRISPR-Cas9 system is a powerful tool for editing ncRNAs, UTRs, multicopy gene families and CDSs keeping their endogenous UTRs. Moreover, the system can be used for the modification of both alleles after just one round of transfection and of codons coding for amino acids carrying post-translational modifications (PTMs) among other possibilities.

# Entwicklung neuer Methoden zur Untersuchung der Genregulation in *Trypanosoma brucei*

## Zusammenfassung

*Trypanosoma brucei* ist nicht nur als Erreger der Schlafkrankheit von großer epidemiologischer Bedeutung, sondern dient auch der Zell- und Molekularbiologie – insbesondere zur Erforschung der Genregulation auf posttranskriptionaler Ebene – als wichtiger Modellorganismus.

In den vergangenen 30 Jahren konnten mehrere Forschungsarbeiten zeigen, dass mRNA-Stabilität und –Prozessierung maßgeblich zur Regulation der Genexpression beitragen. Anders als in den meisten Eukaryoten sind die Gene in *T. brucei* in polycistronischen Transkriptionseinheiten (PTUs) angeordnet. Die reife mRNA entsteht aus dem polycistronischen Transkript in einem gekoppelten Prozess aus Trans-splicing und paralleler Polyadenylierung.

Beide Vorgänge allein, mRNA-Stabilität und –Prozessierung, reichen nicht aus, um die Regulation der Genexpression in *T. brucei* vollständig zu erklären und zusätzliche Mechanismen müssen wirksam sein.

Daher habe ich im ersten Teil meiner hier vorliegenden Doktorarbeit die Genregulation auf Ebene der Translation mittels genomweitem Ribosome Profiling untersucht. Die dabei gewonnenen Daten deuten darauf hin, dass die Translationseffizienzen nicht nur zwischen prozyklischen- und Blutstromformen des Parasiten differieren, sondern auch die Gene innerhalb eines Stadiums verschieden effizient translatiert werden. Zudem war es mir mit diesem Ansatz

möglich, neue, noch nicht annotierte kodierende Sequenzen zu identifizieren und das Kodierungspotenzial der jeweils vorgelagerten offenen Leseraster (ORFs) zu evaluieren. Mithilfe bereits veröffentlichter Proteom- und RNA Interferenz-Studien (RIT-seq) konnte ich einige der neu identifizierten kodierenden Sequenzen validieren und deren Bedeutung für die Fitness des Parasiten bestimmen.

Im zweiten Teil der Arbeit wurden die ermittelten Translations- und Transkriptionsprofile miteinander verglichen, um auf diese Weise mögliche nicht-kodierende RNAs (ncRNAs) zu identifizieren. Dies führte zu einer eingehenderen Betrachtung der Kodierungspotenziale der dem exprimierten variablen Oberflächenproteins (VSG) vor- und nachgeschalteten Regionen. In früheren Arbeiten wurde bereits beschrieben, dass eine Deletion der dem VSG vorgelagerten, sogenannten co-transposed region (CTR), vermehrt zu einer Aktivierung einer alternativen VSG Expressionsseite (*in situ* switches) führt. Ribosome Profiling zeigte, dass eben jede Regionen zwar moderat transkribiert, jedoch nicht translatiert werden. Da diese Ergebnisse vermuten ließen, dass die CTR für eine ncRNA kodiert, hab ich im dritten Teil meiner Arbeit die CTR Transkripte mittels Northern Blot und RACE weiter charakterisiert. Auf diese Weise konnte ich spezifische, 1200 Nukleotide (nt) lange und am 3'-Ende polyadenylierte Transkripte nachweisen.

Die bereits erwähnte Deletion der CTR verbunden mit einer erhöhten Rate an *in situ* switches wurde vor etwa 20 Jahren durch Insertion von Reportergenen durchgeführt. Heute ist es möglich mithilfe von Endonukleasen Genome ohne solche Marker zu editieren. So beschreibt der vierte Teil der Arbeit

die Konstruktion von Episomen zur Etablierung und Anwendung des CRISPR-Cas9 Systems in *T. brucei*. Als Machbarkeitsnachweis wurde die kodierende Sequenz des grün fluoreszierenden Proteins (eGFP) am Ende des SCD6 Gens als Fusionsprotein inseriert. Grün fluoreszierende Zellen konnten bereits zwei Tage nach der Transfektion nachgewiesen werden. Nachdem CRISPR-Cas9 erfolgreich in *T. brucei* etabliert werden konnte, werde ich im Folgenden weitere relevante Regionen im Genom modifizieren und beispielsweise die Deletion zweier Histonvarianten durchführen.

Die Ribosome Profiling Studie in *T. brucei* erlaubt es uns, genomweit Genregulation auf Ebene der Translation zu analysieren, das uns zurzeit vorliegende Genom zu re-annotieren, neue kodierende Sequenzen wie auch ncRNAs zu identifizieren und den Einfluss nicht-kodierender Sequenzen auf die Translation zu untersuchen. Gleichzeitig ermöglicht die Etablierung des CRISPR-Cas9 Systems in *T. brucei* eine hochpräzise Manipulation des Genoms ohne den Einsatz von Resistenzmarkern. Auf diese Weise ist es möglich, Gene zu modifizieren und dabei die zugehörigen untranslatierten Bereiche (UTRs) zu erhalten, aber auch ncRNAs, UTRs und mehrfache Kopien eines Gens (gleichzeitig) zu editieren. Ebenso können einzelne Kodons in der Sequenz und somit posttranslational modifizierte Aminosäuren im Genprodukt verändert werden, was uns weitere Möglichkeiten zur Erforschung der Genregulation eröffnet.

to PVCdA

## **Acknowledgments**

I want to thank Dr. Tim Nicolai Siegel for giving me the opportunity to work in his laboratory, which provided the conditions to make me a much better scientist and was decisive for pursuing my career in the field of tool development. Additionally, I want to thank him for the useful corrections and advice when proofreading this thesis.

I am very grateful to the members of my supervisory committee consisting of Prof. Dr. Christian Janzen, Dr. Ana Eulalio and Priv. Doz. Dr. rer. nat. Sandra Hake for their constructive criticism, insightful comments and encouragement that contribute to my development as a scientist. I also want to thank Prof. Dr. Markus Engstler for serving as the chairperson on my thesis defense.

I want to thank the laboratories of Drs. Christian Janzen, Ana Eulalio, Susanne Kramer, Jörg Vogel and Cynthia Sharma for providing me technical support and sharing reagents and equipment.

I am thankful as well all the members of the Siegel laboratory for their contribution during my doctoral studies. Especially to MSc. Laura Müller for helping me with the translations from English to German included in this document.

I thank the ZINF/IMIB for the fantastic infrastructure that allowed me to perform all the necessary experiments to accomplish my work, Hilde Merkert for technical support and Monika Meece, Jessica Dickerson and Isabell Saal for their permanent support in administrative issues.

I am grateful to the members of the GSLS for their academic guidance during all these years at the University of Würzburg.

Last but not least, I thank the people that constantly and unconditionally support me during these three and a half years of intense work: my wonderful wife Leonor, my beloved parents Patricia and José and my great friend Juan Carlos. You made this possible.

I want also to thank Catalina, Alex, Leoni and Uta for being such a good friends and for the great time we spent together.



# Table of Contents (Inhaltsverzeichnis)

<b>Summary (Zusammenfassung)</b> .....	<b>iv</b>
<b>Acknowledgements ...</b> .....	<b>xi</b>
<b>Table of Contents (Inhaltsverzeichnis)</b> .....	<b>xiii</b>
<b>List of Figures</b> .....	<b>xx</b>
<b>List of Tables</b> .....	<b>xxiii</b>
<b>List of Abbreviations</b> .....	<b>xxiv</b>
<b>Chapter 1: General Introduction</b> .....	<b>1</b>
<b>Trypanosoma brucei: general overview</b> .....	<b>1</b>
African Trypanosomiasis .....	2
Drug discovery .....	3
Antigenic variation .....	4
The co-transposed region (CTR) and its role in <i>in situ</i> switching .....	8
Long non-coding RNAs .....	12
Genome organization .....	13
The <i>T. brucei</i> life cycle .....	16
<b>Regulation of gene expression in <i>T. brucei</i></b> .....	<b>18</b>
Regulation in RNA polymerase II transcribed genes .....	18
RNA maturation .....	20
RNA stability .....	22
The CRISPR-Cas9 system as a tool for genome editing .....	24
Translational efficiency .....	31

Ribosome profiling for the study of gene expression regulation at the level of translation .....	33
Protein stability .....	37
Regulation in RNA polymerase I transcribed genes .....	39
<b>Aims</b> .....	<b>41</b>
<b>Chapter 2: Materials and Methods</b> .....	<b>44</b>
<b>Ribosome profiling</b> .....	<b>44</b>
<i>T. brucei</i> culture and cell harvest .....	45
Nuclease digestion and monosome purification .....	46
Total RNA extraction .....	48
PolyA-enrichment and fragmentation of undigested RNA .....	49
Fragmented RNA and footprint purification .....	49
Preparation of RNA footprint and mRNA sequencing libraries .....	53
RNA dephosphorylation of the 3'-end .....	53
Ligation reaction .....	54
Reverse transcription .....	55
Circularization reaction .....	57
PCR amplification and barcode addition .....	58
Data analysis .....	60
Pre-processing and mapping of reads .....	60
Calculation of abundance, translational efficiency and read 5'-end periodicity.....	61
Definition of uORFs and novel CDSs .....	61
Calculation of ribosome release score .....	62

Identification of novel putative CDSs important for parasite's fitness .....	62
Analysis of proteomics data .....	63
<b>Characterization of a ncRNA transcribed from the CTR .....</b>	<b>64</b>
Northern blot .....	64
Synthesis of <i>VSG</i> and CTR probes by PCR .....	64
Probe radiolabelling using the DECAprime™ II Kit (Ambion) .....	66
Preparation of template DNA .....	66
Probe synthesis reaction .....	66
Isolation of total RNA .....	67
Culture and harvest of BF Lister 427 .....	67
RNA purification .....	68
Preparation of agarose gels for RNA .....	68
Sample preparation .....	68
Electrophoresis .....	69
RNA transfer to a nylon membrane .....	69
Northern hybridization .....	71
Characterization of the CTR ends by circular RACE PCR .....	72
Isolation of total RNA .....	72
RACE using circularized RNA .....	73
Decapping the RNA .....	73
Circularizing the RNA .....	74
Reverse transcription of the circularized RNA into cDNA .....	74
PCR reaction using the Kapa HiFi Ready Mix .....	75
Nested PCR using the Kapa HiFi Ready Mix .....	76

<b>CRISPR-Cas9 system .....</b>	<b>78</b>
Tagging the 3'-end of the SCD6 CDS with the eGFP CDS .....	78
Removal of the luciferase CDS and a vector sequence upstream of the luciferase start codon .....	80
Insertion of the puromycin resistant cassette instead of the G418 resistant cassette for the generation of pEV_puro .....	81
Generation of pEV_Cas9v1_puro .....	81
Generation of pEV_Cas9v2_puro .....	82
Generation of pEV_SCD6_GFP_gBlock .....	82
Insertion of the HH ribozyme, sgRNAs protospacers, the chimeric guide RNA scaffold and the HDV ribozyme sequences in pEV_SCD6_gBlock .....	84
Sequential transfection of wild type PF cells with pEV_Cas9v2_puro and pEV_sgRNA1_SCD6_GFP, pEV_sgRNA2_SCD6_GFP or pEV_sgRNA3_SCD6_GFP .....	85
Co-transfection of wild type PF cells with pEV_Cas9v2_puro and pEV_sgRNA1_SCD6_10_GFP .....	86
Generation of a SCD6-GFP fusion protein using the pMOTag3G and 4G tagging system .....	86
Evaluation of the nuclear localization of Cas9v1 and Cas9v2 by immunofluorescence (IF) .....	87
Evaluation of genome editing and allelic efficiency by PCR .....	89
Evaluation of GFP positive cells using flow cytometry (MACS quant) .....	90
Evaluation of GFP positive cells by microscopy .....	90
Fluorescence microscopy .....	90

Confocal microscopy .....	91
Whole genome sequencing .....	91
Off-target analysis .....	94
Generation of a HSVTK-negative selection system in cells containing the episome pEV_Cas9v2_puro .....	95
Evaluation of episome integration in the genome of the parasite by Southern blot .....	96
Synthesis of probes by PCR .....	96
Genomic DNA extraction and digestion .....	97
Electrophoresis .....	98
Preparation of the gel for transfer .....	98
Transfer of the DNA to a nylon membrane .....	98
Southern hybridization .....	99
<b>Oligonucleotides, primers and cell lines .....</b>	<b>101</b>
Ribosome profiling .....	101
Northern blot .....	102
Circular RACE PCR .....	102
CRISPR-Cas9 system .....	103
<b>Reagents, equipment and consumables .....</b>	<b>105</b>
Ribosome profiling .....	105
Characterization of the CTR .....	106
CRISPR-Cas9 system .....	107

<b>Chapter 3: Implementation of ribosome profiling in two life cycle stages of <i>T. brucei</i> for the study of translational regulation .....</b>	<b>109</b>
<b>Introduction .....</b>	<b>109</b>
<b>Results .....</b>	<b>112</b>
Ribosome footprints indicate ribosome position at sub-codon resolution .....	112
Unambiguous identification of translation initiation sites .....	119
Evaluation of translational efficiency by ribosome profiling .....	121
uORFs may regulate translational efficiency .....	127
Ribosome footprints allow the identification of previously un-annotated CDSs .....	131
Ribosome profiling newly identified CDSs may be important for parasite fitness .....	135
<b>Discussion .....</b>	<b>137</b>
 <b>Chapter 4: Identification and characterization of a ncRNA transcribed from the CTR in the active BES of the Lister 427 strain .....</b>	 <b>143</b>
<b>Introduction .....</b>	<b>143</b>
<b>Results .....</b>	<b>146</b>
CTRs are transcribed into long non-coding RNAs (lncRNAs) .....	146
<b>Discussion .....</b>	<b>151</b>

<b>Chapter 5: Implementation of the CRISPR-Cas9 system for genome editing in <i>T. brucei</i></b> .....	<b>161</b>
<b>Introduction</b> .....	<b>161</b>
<b>Results</b> .....	<b>164</b>
The Cas9, sgRNA and repair sequence of the CRISPR-Cas9 system were cloned in two episomes for their implementation in PF of <i>T. brucei</i> .....	164
Efficient CRISPR-Cas9 mediated gene editing in <i>T. brucei</i> .....	168
Episome stability in <i>T. brucei</i> .....	175
Off-target mutation analysis .....	177
<b>Discussion</b> .....	<b>179</b>
<b>Chapter 6: Conclusions and future directions</b> .....	<b>188</b>
<b>Ribosome profiling enables a genome-wide translation analysis in different life cycle stages of <i>T. brucei</i></b> .....	<b>188</b>
<b>A polyadenylated, non-spliced ncRNA is transcribed from the CTR in the Lister 427 strain of <i>T. brucei</i></b> .....	<b>191</b>
<b>The CRISPR-Cas9 system allows the efficient genome editing in procyclic forms</b> .....	<b>197</b>
<b>List of collaborators</b> .....	<b>200</b>
<b>References (Literaturverzeichnis)</b> .....	<b>201</b>
<b>Curriculum vitae</b> .....	<b>245</b>

## List of Figures

<b>Figure 1.</b> Antigenic variation in <i>T. brucei</i> .....	4
<b>Figure 2.</b> Electron microscopy evidences the dense VSG coat covering the cell membrane outer surface in <i>T. brucei</i> .....	5
<b>Figure 3.</b> Main mechanisms of VSG switching in <i>T. brucei</i> .....	7
<b>Figure 4.</b> Localization of the co-transposed region (CTR) in the BES of <i>T. brucei</i> .....	10
<b>Figure 5.</b> Life cycle of <i>T. brucei</i> .....	17
<b>Figure 6.</b> Representation of the RNA-guided Cas9 nuclease .....	28
<b>Figure 7.</b> Flow chart illustrating the aims of the thesis.....	43
<b>Figure 8.</b> Establishment of a ribo-seq protocol .....	45
<b>Figure 9.</b> Representative gels from intermediate product purification in the library generation process .....	51
<b>Figure 10.</b> Schematic view of the CTR and the corresponding regions from where the CTR-U and CTR-D probes were synthesized .....	65
<b>Figure 11.</b> Schematic illustration of a circular RACE PCR .....	77
<b>Figure 12.</b> Episomes used for tagging the SCD6 CDS with the eGFP sequence using the CRISPR-Cas9 system .....	79
<b>Figure 13.</b> Repair sequence for the insertion of the eGFP sequence at the 3'-end of the SCD6 CDS and the three sgRNAs evaluated .....	84
<b>Figure 14.</b> Separation of sonicated gDNA in an agarose gel for library preparation .....	92
<b>Figure 15.</b> Sucrose density gradient profiles of undigested and RNaseI-digested samples from PF and BF .....	113



<b>Figure 16.</b> Pair-wise comparison of ribosome footprint density in BF and PF	.114
<b>Figure 17.</b> Ribosome footprints are enriched specifically across CDSs	..... 115
<b>Figure 18.</b> Alignment of the 5'-nucleotide from ribosome footprint reads	..... 116
<b>Figure 19.</b> Ribosome footprints reveal coding sequences at sub-codon resolution	..... 117
<b>Figure 20.</b> Meta-gene analysis of ribosome density	..... 118
<b>Figure 21.</b> RNA-seq and ribosome profiles for the two intron-containing genes	..... 119
<b>Figure 22.</b> Ribosome profiling results of two likely mis-annotated CDSs	..... 121
<b>Figure 23.</b> Translational regulation in PF and BF	..... 123
<b>Figure 24.</b> Life cycle regulation of translational efficiency	..... 125
<b>Figure 25.</b> Characteristic 3-nt periodicity of ribosome profiling footprints in uORF	..... 129
<b>Figure 26.</b> Ribosome footprints allows the identification of un-annotated CDSs	..... 134
<b>Figure 27.</b> Ribosome profiling and RIT-seq profiles of two previously un- annotated putative CDSs in BF	..... 137
<b>Figure 28.</b> Representation of the RNA-seq and ribosome profiling data for the region encompassing the 70-bp repeat region and the telomere repeats in the BES1 of Lister 427	..... 147
<b>Figure 29.</b> Characterization of the ncRNA transcribed from the CTR in the active BES of Lister 427	..... 148
<b>Figure 30.</b> The 5'-end of the CTR transcript is capped	..... 150
<b>Figure 31.</b> Cellular localization of a human codon adapted SpCas9 expressed in PF	..... 165

<b>Figure 32.</b> Processing of the sgRNA by the HH and HDV ribozymes .....	167
<b>Figure 33.</b> Localization of SCD6-eGFP expressed in cells sequentially transfected with episomes including the CRISPR-Cas9 system .....	169
<b>Figure 34.</b> Flow cytometry analysis of PF transfected with pEV_Cas9v2_puro and pEV_sgRNA1_SCD6_GFP .....	172
<b>Figure 35.</b> Evaluation of the allelic editing frequency by PCR .....	174
<b>Figure 36.</b> Testing the episome stability in PF by Southern blot .....	176

## List of Tables

<b>Table 1.</b> Relevant information about the sequencing process of the six ribosome profiling libraries generated in this study .....	60
<b>Table 2.</b> Relevant information about the sequencing process of the two libraries generated to evaluate the on and off-target effects of the CRISPR-Cas9 system .....	93
<b>Table 3.</b> Developmentally regulated translation in BF and PF .....	126
<b>Table 4.</b> Translational efficiency of genes preceded or not by uORF .....	130
<b>Table 5.</b> Percentage of fluorescent cells three weeks after the second round of transfection .....	171
<b>Table 6.</b> List of putative off-target sites and their location regarding indels for the sgRNA1 (SCD6-GFP) .....	178

## List of Abbreviations

bp base pair  
BES bloodstream form expression site  
BF bloodstream forms  
DAPI 4,6-diamidino-2-phenylindole  
DNA deoxyribonucleic acid  
DSB double stranded break  
eGFP enhanced green fluorescent protein  
ESAG Expression site associated genes  
ESB expression site body  
gDNA genomic DNA  
HAT Human African trypanosomiasis  
H2AZ a variant form of H2A  
H2BV histone H2B variant  
H3V histone H3 variant  
H4 histone H4 unmodified  
H4K4 H4 unmodified at lysine 4  
H4K4ac H4 acetylated at lysine 4  
H4K10ac- H4 acetylated at lysine 10  
H4V histone H4 variant  
HAT trypanosome histone acetyltransferases  
HAT trypanosome histone acetyltransferase 3  
HDR homology directed repair  
HR homologous recombination  
IF immunofluorescence  
indel insertion/deletion mutations  
ISWI imitation SWI2  
KO gene knockout  
lncRNA long non-coding RNA  
MACS magnetic activated cell sorting  
Min minute(s)  
MMEJ microhomology-mediated end joining  
ncRNA non-coding RNA  
NGS next generation sequencing  
NHEJ non-homologous end joining  
nt nucleotide  
ORF open reading frame  
PAM protospacer adjacent motif  
PABP poly(A)-binding protein  
PARP procyclic acidic repetitive proteins  
PAS polyadenylation site  
PCR polymerase chain reaction  
PF procyclic forms  
poly(A) polyadenine  
PTM post-translational modifications  
PTU polycistronic transcription unit  
Ribo-seq ribosome profiling  
RNA ribonucleic acid

RNAi RNA-interference  
RNA-seq RNA sequencing  
rpkm reads mapped per kilobase per million non-structural RNA reads  
rRNA: ribosomal RNA  
RS repair sequence  
SAS splice acceptor site  
Sec second(s)  
sgRNA chimeric single-guide RNA  
SL spliced leader  
snRNA: small nuclear RNA  
SpCas9 *Streptococcus pyogenes* Cas9 endonuclease  
SSR strand-switch region  
*T. brucei Trypanosoma brucei*  
TbSir2rp Sir2-related protein  
tRNA: transfer RNA  
TSS transcription start site  
TTS transcription termination site  
UTR untranslated region  
VDR *VSG* downstream region  
VSG variant surface glycoprotein  
VSG-2 preferred name  
VSG-221 local name for the same VSG-2

## Chapter 1. General Introduction

This introduction contains an overview about general aspects on the biology of *T. brucei*, a brief summary of the current knowledge on the regulation of gene expression in the parasite, and an introduction about ribosome profiling, long ncRNAs (lncRNAs), the CTR and the CRISPR-Cas9 system for genome editing.

### ***Trypanosoma brucei*: general overview**

*Trypanosoma brucei* has been classified as a member of the taxon: superkingdom Eukaryota; kingdom Excavata; phylum Euglenozoa; order Kinetoplastida; family Trypanosomatidae; genus Trypanosoma; subgenus Trypanozoon; species *T. brucei*. The Kinetoplastida is one of the earliest diverging lineages among eukaryotes (Fernandes et al., 1993) and for this reason the organisms belonging to this order present dissimilar characteristics compared to the more recently diverged lineages of eukaryotes. The lack of regulation at the level of transcription initiation, the RNA-editing in the mitochondrion (Benne et al., 1986), the presence of an RNAi silencing pathway (Ngô et al., 1998) that has been fundamental for the down regulation of genes for experimental purposes, and the trans-splicing of mRNA (De Lange et al., 1984; Parsons et al., 1984) are among the several characteristics that make *T. brucei* an interesting model organism.

## **African Trypanosomiasis**

*T. brucei* was first observed by David Bruce in 1894 in the blood of a cow in South Africa (Bruce, 1895) and for this reason the species was later named in his honor in 1899.

Human African trypanosomiasis (HAT), also known as sleeping sickness, is a disease caused by two subspecies of *T. brucei*: *T. b. gambiense* and *T. b. rhodesiense*. The first subspecies causes chronic infections while *T. b. rhodesiense* causes acute infections (WHO, 2015). A third subspecies, *T. b. brucei*, cannot infect humans but it is the etiological agent of nagana in animals, which includes cattle and therefore causes important economical losses. The occurrence of HAT is restricted to the distribution of tsetse flies in 36 sub-Saharan African countries and *T. b. gambiense* causes over 98% of the reported cases (WHO, 2015).

Since the disease was described, several epidemics have occurred in Africa over the last century: between 1896 and 1906, mostly in Uganda and the Congo Basin, in 1920 and the most recent from 1970 to the late 1990s (Fèvre et al., 2008; Lozano et al., 2012). Since then, permanent control efforts have reduced the number of new cases to 6,314 recorded events in 2013 and it is expected that HAT may be eliminated as a public health problem by 2020 (WHO, 2015).

The prevalent symptoms in the patient are divided in two stages. In the first stage of the disease, when the trypanosomes are restricted to the blood and lymphatic systems, the symptoms include fever, headache, joint pain and itching. In the second stage, when the trypanosomes invade the central nervous system,

the symptoms are characterized by neurological and endocrinal complications. In acute infections caused by *T. b. rhodesiense*, the untreated patient can die within months whereas those infected with the chronic form (*T. b. gambiense*) can survive for several years (WHO, 2015).

Although HAT is not anymore a public health issue for many of the countries where the tsetse fly is endemic, the livestock disease is still causing an economic burden for the farmers throughout sub-Saharan Africa, and evidence of drug resistance is becoming a cause of concern for the governments in those areas (WHO, 2015).

### **Drug discovery**

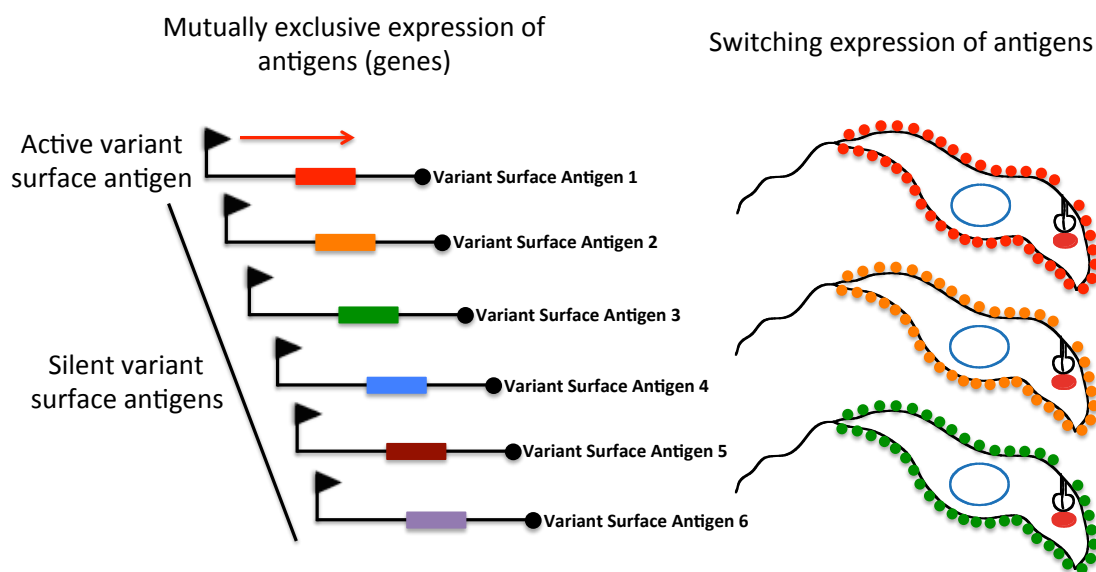
At present there are five drugs available for the chemotherapy of HAT (Steverding, 2010). Suramin, pentamidine and melarsoprol were developed in the first half of the XX century. Suramin is still in use for the therapy of early-stage caused by *T. b. rhodesiense* and pentamidine is still used for treatment of the first stage of *T. b. gambiense* sleeping sickness. Melarsoprol was introduced in 1949 and is the only effective drug for the second stage of *T. b. rhodesiense* sleeping sickness. Eflornithine was originally developed in the 1970s as an anti-cancer drug and in 1990 approved for use against *T. b. gambiense*. NECT (nifurtimox eflornithine combination therapy) was launched in 2009 and reduces the time of treatment from 14 to 10 days as well as the number of intravenous doses (Yun et al., 2010). Nevertheless, the possibility of appearance of resistant strains cannot be ruled out. For this reason the elucidation of the physiology and specifically the regulation of gene expression in *T. brucei* can be



valuable for the development of new therapeutic strategies that in turn can reduce the secondary effects produced by the currently available drugs mentioned above.

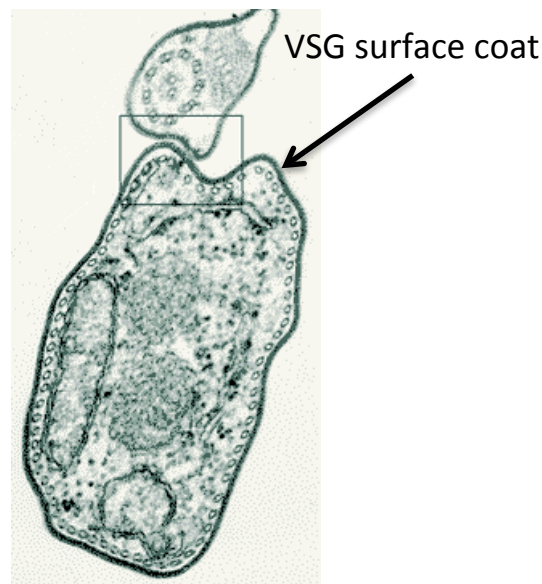
## Antigenic variation

The *T. brucei* form present in the bloodstream of the mammalian host avoids destruction by the immune response through their capacity for antigenic variation (reviewed in Horn, 2014). Antigenic variation requires two mechanisms: one that ensures the mutually exclusive expression of only one antigen from a large family of antigens; and one that allows the parasite to switch the expression of antigens periodically (Deitsch et al., 2009) (Figure 1).



**Figure 1.** Antigenic variation in *T. brucei*. Antigenic variation comprises two processes in *T. brucei*: the mutually exclusive expression of single antigens from a large pool of antigen (left) and the parasite's ability to switch the expression of antigens periodically (right).

In the bloodstream form (BF) stage of *T. brucei*, the main antigens that induce an immune response are the VSGs (Vickerman, 1969) (Figure 2).



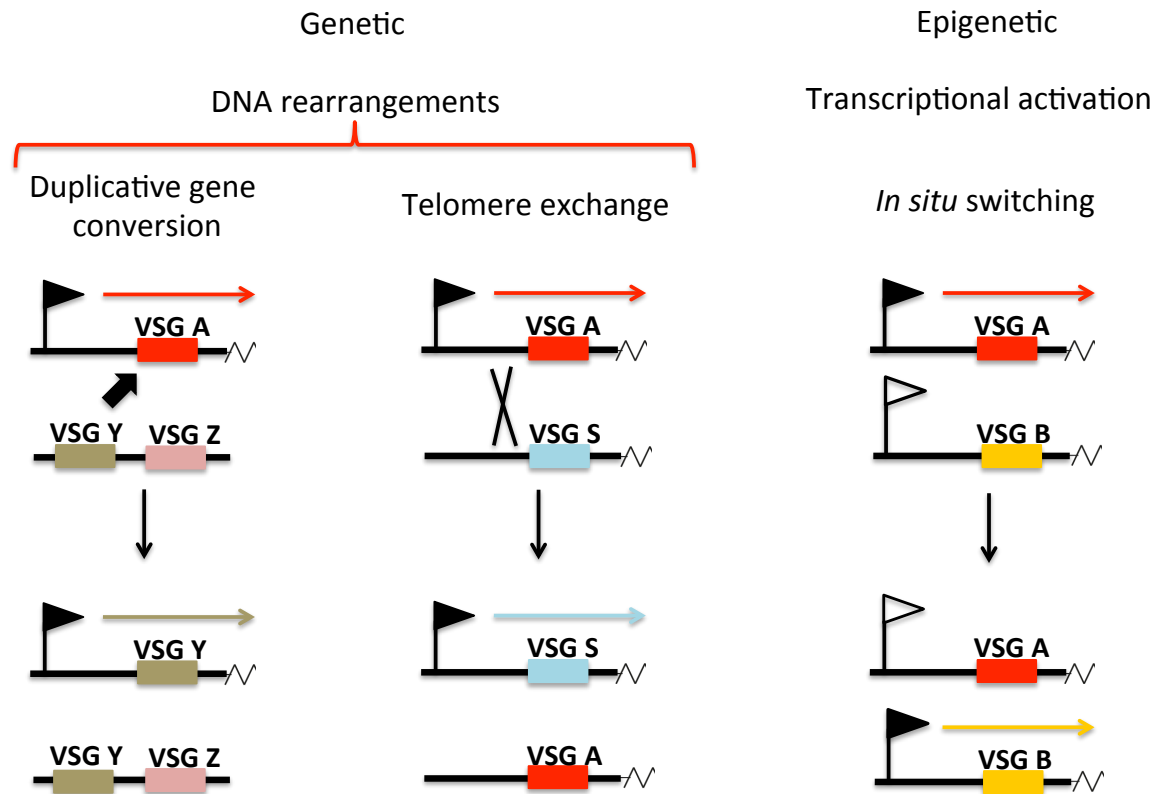
**Figure 2.** Electron microscopy evidences the dense VSG coat covering the cell membrane outer surface in *T. brucei* (Vickerman, 1969).

These proteins present in the coat of the parasite were purified by Cross (Cross, 1975) and the crystal structure to a resolution of 2.9 Å was published in 1990 (Freymann et al., 1990). The N-terminal domain provides the antigenic diversity and the C-terminal domain interacts with the cell membrane through its anchoring via glycosylphosphatidylinositol (Ferguson et al., 1988). The fact that there are ~ 10 million tightly packaged VSG molecules on the surface of the parasite (Cross, 1975), their rapid recycling (Engstler et al., 2004) and their fluidity (Engstler et al., 2007) makes it difficult for the host immune response to clear the infection.

Depending on the strain analyzed, there are between 15 to 20 bloodstream form expression sites (BES). In the specific case of Lister 427, 15 of these BES have been identified and cloned (Hertz-Fowler et al., 2008). BESs are

subtelomeric transcription units composed of an RNA polymerase I promoter, Expression Site Associated Genes (*ESAGs*) and a *VSG* gene upstream of the telomeric regions. The promoter is located around 60 kbp upstream of the *VSG* (Johnson, 1987) and recruits RNA polymerase I (Günzl et al., 2003), but its sequence is different of the rDNA promoter (Zomerdijk et al., 1991).

The main mechanisms of *VSG* switching in *T. brucei* are duplicative gene conversion, telomere exchange, segmental gene conversion and expression site switching (Rudenko, 2011) (Figure 3).



**Figure 3.** Main mechanisms of VSG switching in *T. brucei*. The filled boxes indicate VSGs. The filled flags indicate the promoter in the active BES and the non-filled flags promoters in silenced BES. Transcription is indicated with an arrow with the same color corresponding to the expressed VSG. The zigzagged symbol at the end of the BES corresponds to the telomere. Left. Duplicative gene conversion. Any VSG can be copied into the active BES and the originally expressed VSG is lost. Center. Telomere exchange. A VSG switch mediated by a telomere exchange involves DNA recombination between two telomeres. Right. *In situ* switching. A new BES is transcriptionally activated and in turn the old one is silenced.

The first three mechanisms are dependent of genetic rearrangements mediated by homologous recombination (HR) (McCulloch and Barry, 1999) and the last one by transcriptional activation. In general, VSGs are flanked by common sequences: upstream by 0.2–7.1 kb long stretch of 70 bp repeats and downstream by the region encoding the VSG C-terminal domain or the VSG 3'-untranslated region (UTR) (Marcello and Barry, 2007). According to the most

recent estimates, the *VSG* reservoir comprises more than 2,000 sequences of mainly incomplete *VSGs* (pseudogenes) (Cross et al., 2014), which by segmental gene conversion can generate 'mosaic' *VSGs* important for parasite survival in chronic infections (Hall et al., 2013). This discovery will help to understand the order of appearance of the different *VSG* antigens.

The mutually exclusive activation of only one BES occurs in an extranucleolar RNA polymerase I transcription site that has been named the expression site body (ESB) (Navarro and Gull, 2001). This structure provides a positional and structural explanation for the monoallelic exclusion but the epigenetic mechanisms controlling the activation of the BES and the silencing of the others are just beginning to be elucidated (Hughes et al., 2007; Povelones et al., 2012; Wang et al., 2010; Yang et al., 2009).

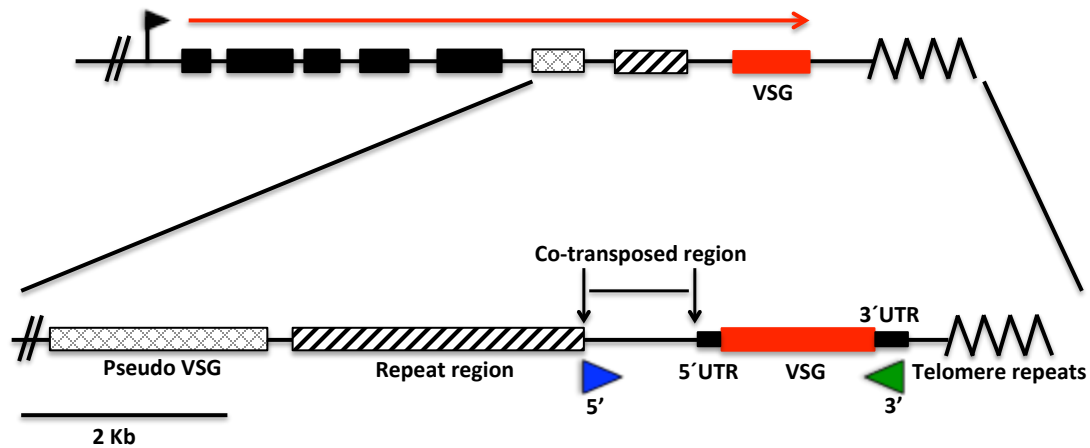
Even though the mechanism of antigenic variation is by far the most studied field of research in trypanosomes, at present we just have an outline of the machinery controlling the mutually expression of *VSGs* and switching.

### **The co-transposed region (CTR) and its role in *in situ* switching**

The process regulating *VSG* expression operates at the level of transcription initiation, elongation, RNA maturation, stability, and translation efficiency (Vanhamme and Pays, 1995). As mentioned above, in *T. brucei* only one *VSG* from a pool of around 2,000 sequences is expressed at any given time, conforming the mechanism of mutually exclusive expression of antigens. The exact mechanism of antigenic variation in *T. brucei* is at present unknown but there are two basic ways

by which *T. brucei* achieves *VSG* gene switching. They include genetic rearrangements by duplicative gene conversion, telomere exchange and segmental gene conversion or by transcriptional activation of a new BES also known as *in situ* switching (Rudenko, 2011). As the mechanism of antigenic variation suggest, once a new BES is activated, the previous one has to be shut down because the host immune response may kill parasites faster if they express a mixed coat of VSGs. In laboratory-adapted BF, the rate of *in situ* switching has been estimated in around  $1 \times 10^{-7}$  per generation (Lamont et al., 1986).

Apparently, each individual *VSG* in the reservoir is flanked by a co-transposed region (CTR) (Aline et al., 1985; Campbell et al., 1984; Florent et al., 1987; Liu et al., 1983). It has been aptly called the co-transposed region because the 5'- and 3'- boundaries during a duplicative gene conversion event are located within the 70-bp repeat region and either in the *VSG* 3'-coding region or 3'-UTR (Figure 4), which in both cases contain highly conserved sequences. Therefore, after each gene conversion event, the CTR is co-transposed together with the *VSG* sequence. Furthermore, the CTRs present in the BESs present different lengths and DNA sequence.



**Figure 4.** Localization of the co-transposed region (CTR) in the BES of *T. brucei*. Top: schematic view of the BES. The flag represents the BES promoter, the parallel lines indicate the limit of the region shown, the black boxes represent *ESAGs*, the outlined diamond box a pseudo *VSG* present in the BES1 in Lister 427, the wide upward diagonal box the 70-bp repeat region, the red box the *VSG* CDS, the zigzagged lines the telomeres and the red line the direction of transcription in the active BES. Bottom: the region where is located the CTR was zoomed up. The conventions are the same as above. In addition, the smaller black boxes represent the *VSG* UTRs and the blue and green triangles the boundaries of the region translocated when a duplicative gene conversion event occurs.

In 1997, Davis *et al.* deleted part of the CTR sequence in the *T. brucei* strain Lister 427 causing a significant increase in the rate of *in situ* switching (Davies *et al.*, 1997). Disruption of most of the CTR sequence in two active BESs (expressing VSG-221 or VSG-118) caused an increase higher than 100-fold to about  $1.3 \times 10^{-4}$  switching events per generation. When they disrupted the entire CTR, the 5'-UTR and the start codon of VSG-221, a new BES was activated after a few cell divisions even though another *VSG* was inserted into the active BES upstream of the CTR deletion. The deletion of the 70-bp repeats and/or the *VSG* pseudogene present upstream of the CTR, the introduction of a reporter gene

(luciferase or *VSG*-117) or a resistant marker did not affect the BES stability.

In summary, that was the first report indicating that the editing of the *T. brucei* genome can increase the BES switching. In the paper, the authors speculate about the putative role of the CTR as a stabilizing agent of the chromatin arrangement in the active BES. Even though most of the cells continued to use the active BES some of the cells switched to a new BES by the requirement of some factor present in the CTR upstream of the expressed *VSG*. They suggest as well that previously reported CTR-derived transcripts may fulfill the stabilizing role (Aline et al., 1989; Scholler et al., 1988). Apparently, each *VSG* is preceded by a specific CTR that may contain motifs enabling the folding of the corresponding transcript to have regulatory activity over a yet unknown target or affecting the DNA structure in a defined region in the active BES. In 1986, it was reported that mung-bean nuclease preferentially cleaves the DNA between 300–800-bp upstream of the 5'-end of the *VSG* CDS indicating that the nuclease hypersensitivity was originated by an anomaly in the DNA coiling or some difference in the chromatin structure (Brown et al., 1986). The stabilizing element(s) may interact with nuclear matrix proteins for the localization of the active BES in the ESB, allowing a high level of gene transcription. Finally, Davis *et al.* hypothesized that the stabilizing element present in the CTR may prevent the heterochromatinization of the active BES.



The results published by Scholler *et al.* point to the presence of lncRNAs transcribed from the CTR in *T. brucei* but Aline *et al.* showed that some of them were associated with polysomes. Therefore, with the ribosome profiling results obtained in my thesis I was able to determine whether the transcripts from the CTR in the active BES of Lister 427 correspond to lncRNAs or whether they were translated to proteins.

### **Long non-coding RNAs**

In mammals thousands of long transcripts that do not encode proteins are located in intergenic, intronic regions or antisense to protein-coding genes (Cheng *et al.*, 2005; Consortium *et al.*, 2005a, 2005b). According to The Encyclopedia of DNA Elements (ENCODE), at least 80% of the human and mouse genomes are transcribed and can be grouped in any of these three categories (Consortium, 2007; Rosenbloom *et al.*, 2012).

lncRNAs are operationally defined as non-protein-coding RNAs that are >200 nucleotides in length (Morris and Mattick, 2014). Below this cutoff are included all known classes of small RNAs (Mercer *et al.*, 2009). Basically, a transcript can be classified as non-coding if contains open reading frames (ORFs) of less than 100 codons. ncRNAs are categorized as intronic, large intergenic non-coding RNA (lincRNA) or antisense according to experimental studies, mechanistic expectations or to distinguish them from protein-coding genes (Morris and Mattick, 2014). However, these three subclasses are related only to genomic origin and

for classificatory purposes but there is not a structural difference between them.

Taking into account that sometimes lncRNAs present low evolutionary conservation and low levels of expression, it is possible that they just correspond to transcriptional noise without any physiological function. Nevertheless, the lack of primary sequence conservation does not necessarily mean lack of function (Pang et al., 2006) and these transcripts can still present some kind of structural conservation (Johnsson et al., 2014; Smith et al., 2013).

### **Genome organization**

*T. brucei* belongs to the order of Kinetoplastid organisms together with other parasites like *Trypanosoma cruzi* and *Leishmania* species. This group is characterized by the presence of a flagellum and an organelle called kinetoplast from which the order derives its name. A kinetoplast is a network of circular DNA contained inside a large mitochondrion.

The nuclear DNA is composed of 11 pairs of megabase-size chromosomes, 5 intermediate-size chromosomes (200 to 900 kb) and ~100 minichromosomes (30 to 150 kb) (Daniels et al., 2010). The proposed function of the minichromosomes is to increase the pool of telomeric *VSG* genes for antigenic variation. The MBCs contains around 9,068 predicted genes, including 900 pseudogenes and 1,700 *T. brucei*-specific genes (Berriman et al., 2005). The genes located in the MBCs are organized as long, non-overlapping PTUs (Berriman et al., 2005). Contiguous PTUs may be located on the same DNA

strand or on different strands, separated by convergent or divergent strand switching regions (SSRs) and sometimes interrupted by tRNA, snRNA, siRNA or rRNA genes. Genes in PTUs are constitutively expressed and in general they are functionally unrelated. Analysis of the parasite transcriptome revealed that RNA polymerase II transcription initiates bidirectionally from putative pol II transcription start sites (TSS) at divergent SSRs, as well as at internal locations in the PTU array (Kolev et al., 2010).

In addition to transcribing rRNA genes, RNA polymerase I transcribes as well procyclin (also known as procyclic acidic repetitive proteins [PARP]) and *VSG* genes. The structure of the BES containing the *VSG* genes was already introduced in the section “Antigenic variation”. Nevertheless, it is important to mention that in addition to the *VSG* genes, there are several *ESAGs* located in the BES and only a few of them have been properly characterized. In the Lister 427 strain they present a similar topology on each BES with only minor differences (Hertz-Fowler et al., 2008). For example, *ESAG3*, *ESAG5* and *ESAG11* encode membrane-associated or targeted proteins (Pays et al., 2001); *ESAG4* is a putative transmembrane receptor with adenylate cyclase activity (Paindavoine et al., 1992); *ESAG6* and *ESAG7* are subunits of a receptor for the uptake of transferrin from the host (Ligtenberg et al., 1994) and *ESAG8* is a putative nuclear DNA-binding protein that accumulates in the nucleolus (Hoek et al., 2000). Outside the BES, the genome of *T. brucei* harbors copies of *ESAGs* containing subtle differences that could provide growth advantages in different host sera (Steverding, 2003).

The replication machinery for nuclear DNA is unusual regarding the formation of the early replication complex and the presence of multiple widely spaced origins of replication probably related to the gene organization of the PTUs in these parasites (Tiengwe et al., 2012).

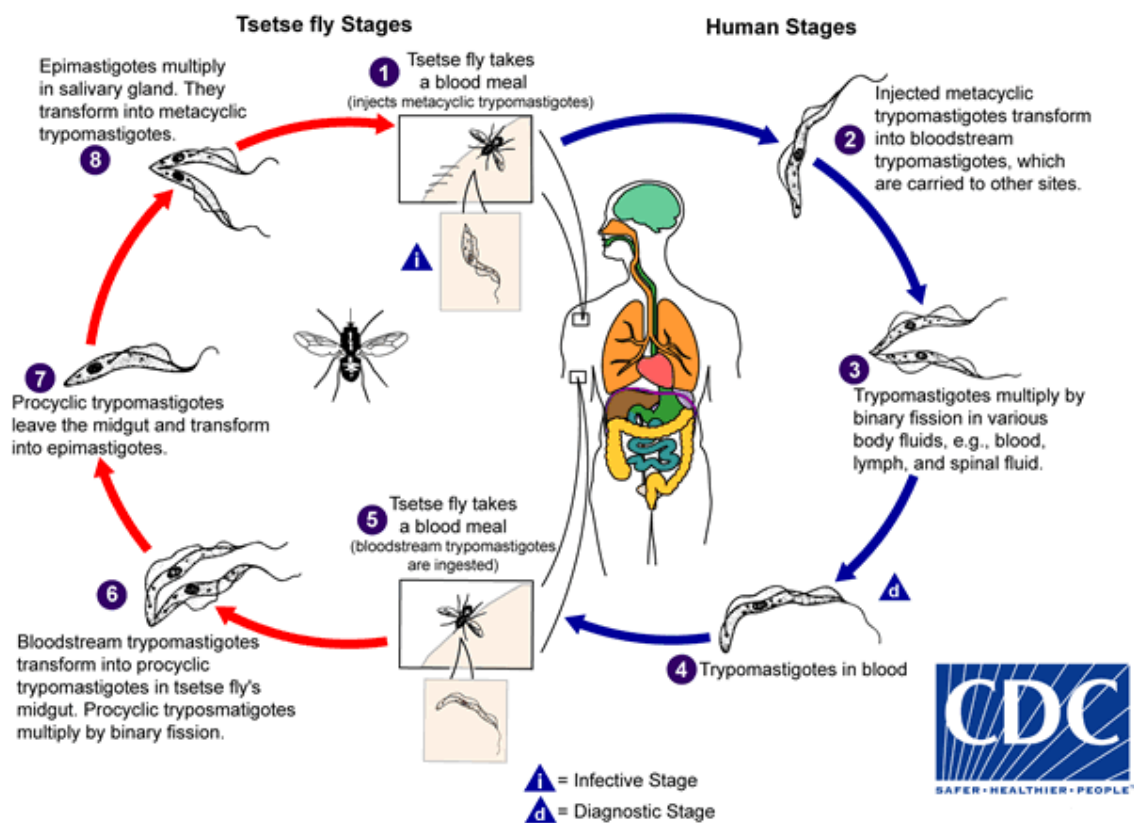
The most commonly *T. brucei* strains used in molecular and cell biology research are TREU 927/4 and Lister 427. TREU 927/4 is pleomorphic, which makes it competent for progression through the different stages of the parasite life cycle (van Deursen et al., 2001) and its genome was formally published as a draft genome in July 2005 (Berriman et al., 2005). Lister 427 is a monomorphic strain that grows fast *in vitro* and presents a low frequency of antigen switching, characteristics that make it attractive for research groups studying antigenic variation (Wirtz et al., 1999). The failure of Lister 427 to complete the natural developmental cycle in the vector was the key factor for its elimination as a candidate for the *T. brucei* genome project (Berriman et al., 2005).

Even though both strains belong to the same species, several analyses have pointed out important differences in the genome organization. The Lister 427 genome contains approximately 16.5 Mb more megabase chromosomal DNA than the 26 Mb TREU 927/4 genome and its chromosomes are in general larger than their homologues in TREU 927/4 (Melville et al., 2000). In addition, the sequences of some BES are not found on the same chromosomes in each strain (Melville et al., 2000), which have complicated the study of these important regions in Lister 427.

## The *T. brucei* life cycle

The infection of the mammalian host by *T. brucei* occurs after the bite of tsetse flies (genus *Glossina*) (reviewed in Fenn and Matthews, 2007; Matthews et al., 2004) (see Figure 5). Once this happens, the parasites present in the bloodstream multiply rapidly as morphologically long slender forms and as a consequence, the parasitemia increases, accumulating the so-called stumpy induction factor (SIF). SIF is a factor or group of factors of unknown nature that stimulates the transition of slender forms to short stumpy forms (Vassella et al., 1997). Nevertheless, some of the components of the signaling pathway required for slender-to-stumpy differentiation have been characterized in recent years (Mony et al., 2014). Both types of BF express the VSGs but the stumpy cells are arrested in the G1 phase of the cell cycle (Matthews and Gull, 1994; Shapiro et al., 1984). When taken up in a tsetse blood meal, the stumpy forms differentiate to procyclic forms (PF) in a process that can be simulated *in vitro* by a reduction in temperature from 37 °C to 27 °C and by the addition of cis-aconitate and/or citrate (Ziegelbauer et al., 1990). PF migrate to the tsetse midgut, where they can proliferate and express a different set of coat proteins called the procyclic surface antigens EP and GPEET. EP and GPEET are expressed in early PF forms and only the EP procyclin in late PF (Vassella et al., 2001). Then, PF migrate to the salivary glands and differentiate into epimastigote forms after an asymmetric division and by expressing a new surface protein called BARP (Sharma et al., 2008; VanDenAbbeele et al., 1999). It is in this life cycle stage that *T. brucei* can undergo meiosis and sexual reproduction (Peacock et al., 2011). Epimastigotes in turn differentiate into the infective metacyclic forms, which are arrested in their

division and express a more reduced set of metacyclic VSGs. Once the fly takes a blood meal, they differentiate into the proliferative slender BF (Roditi and Liniger, 2002). The length of the life cycle varies from one to several weeks depending of the different trypanosome strains.



**Figure 5.** Life cycle of *T. brucei*. See the text for a detailed explanation. Image taken from <http://www.cdc.gov/parasites/sleepingsickness/biology.html>.

This complex life cycle raises the question of how the different *T. brucei* life cycle stages differentially regulate their gene expression in order to adapt to several stresses and stimuli experimented in the vector and the mammalian host.

## **Regulation of gene expression in *T. brucei***

### **Regulation in RNA polymerase II transcribed genes**

The *T. brucei* genome transcribed by both, RNA polymerase II and RNA polymerase I, is organized in PTUs. RNA polymerase II transcribes the majority of mRNAs with the exception of procyclin, *VSG* and *ESAG* genes that are transcribed by RNA polymerase I.

The lack of evidence for promoter motifs upstream of the PTUs transcribed by RNA polymerase II and the fact that these units are thought to be constitutively transcribed, suggest that the gene expression regulation in trypanosomes is post-transcriptional (Jensen et al., 2009; Kolev et al., 2010; Koumandou et al., 2008).

In addition, the transcription machinery differs from the canonical ones present in other eukaryotes such as *Plasmodium falciparum*, *Saccharomyces cerevisiae* and *Homo sapiens*. The *T. brucei* genome contains homologs of 12-15 subunits present in human and *S. cerevisiae* RNA polymerases (Palenchar and Bellofatto, 2006). However, there are important differences such as the largest subunit of the *T. brucei* RNA polymerase II that has a unique carboxyl-terminal domain and the two isoforms of the RPB5 and RPB6 subunits, which are usually single proteins in other eukaryotes (Palenchar and Bellofatto, 2006). Another interesting example is the transcription factor IIB, which is essential for spliced leader (SL) RNA gene transcription and cell viability, but its primary sequence is highly divergent in comparison to TFIIB in other eukaryotes (Ibrahim et al., 2009).

Within PTUs, the transcription is initiated at TSS marked by the presence of the histone variants H2A.Z, H2B.V, the histone modification H4K10ac (histone 4 acetylated at lysine 10) and the bromodomain protein BDF3 (Siegel et al., 2009). In turn, transcription termination at the end of PTUs is marked by the enrichment of the histone variants H3.V and H4.V (Siegel et al., 2009). Base J is present only in BF and is localized within regions flanking the RNA polymerase II transcribed PTUs throughout the genome as well as in repetitive DNA elements in the telomeres and in the 50, 70, and 177 bp repeat units (Cliffe et al., 2010). This base is present in all Kinetoplastids and corresponds to the thymidine residue  $\beta$ -D-glucosyl-hydroxymethyluracil, the only hyper-modified nucleotide identified in eukaryotes (van Leeuwen et al., 2000), which is produced in trypanosomes by the action of the enzymes JBP1 and JBP2.

For the reasons mentioned above, it is assumed that the regulation of genes transcribed by RNA polymerase II occurs at the post-transcriptional level by means of differential trans-splicing and polyadenylation efficiency, RNA stability mediated mainly by 3'-UTRs, translational regulation and protein turnover. It has been shown as well that the sites for RNA polymerase II transcription start and termination are delimited by histone PTMs, histone variants and base J and might confer access to the transcription machinery to an opened chromatin structure. These marks may lead to the remodeling of the chromatin by ATP-dependent remodeling complexes, histone acetyltransferases (HATs), deacetylases, methyltransferases and histone chaperones. Therefore, the effect of the *T. brucei* epigenome in the control of gene expression has been an



intense focus of research in recent years but its implications are beyond the scope of this thesis.

### **RNA maturation**

For most eukaryotic genes, the pre-mRNA that is transcribed from a gene sequence is processed before it becomes a mature mRNA. One of the processing mechanisms is RNA splicing, which involves the removal of certain sequences referred to as intervening sequences, or introns (Clancy, 2008; Hastings and Krainer, 2001; Reed, 2000). The mRNA thus consists of the remaining sequences, or exons, which are connected to one another through the splicing process. Several splicing mechanisms have been described, such as cis-, trans- and alternative splicing among others (Clancy, 2008). Trans-splicing of leader sequences onto the 5'-ends of mRNAs has been found in organisms such as nematodes and protozoa like *T. brucei*. In this parasite, the immature transcripts from the PTUs are co-transcriptionally processed by a coupled process including the addition of the SL (trans-splicing) and polyadenylation at the 3'-end (Liang et al., 2003; Matthews et al., 1994). The SL is a 39-nt sequence that is trans-spliced from a 135-nt transcript at the 5'-end of the immature mRNAs.

Trans-splicing is important for the maturation of mRNAs from primary transcripts and for the addition of the SL sequence that contains the cap structure required for the initiation of translation (Perry et al., 1987). In turn, polyadenylation confers RNA stability and provides a

binding site for poly(A)-binding proteins (PABPs), which are important for translation (Jackson et al., 2010).

Alternative trans-splicing of immature transcripts for the production of proteins with different characteristics and functions is an important post-transcriptional process for the regulation of gene expression in the parasite (Kolev et al., 2010; Nilsson et al., 2010; Siegel et al., 2010). In addition, this process can lead to a shortening or lengthening of the UTRs, modifying regulatory elements important for RNA stability or translation. The combined sequencing data from three studies enabled the mapping of >32,000 unique 5'- splice-acceptor sites (SAS) of >8,900 genes and the average number of per gene was 2.8. For example, using spliced leader trapping, Nilsson *et al.* detected the 5'-SAS of 85% of the annotated protein-coding genes, finding that up to 40% of transcripts were differentially expressed. They discovered as well more than 2,500 alternative splicing events that in some cases were developmentally regulated (Nilsson et al., 2010). In turn, >50,000 polyadenylation sites (PAS) for >8,000 genes have been mapped (Kolev et al., 2010; Siegel et al., 2010), which implies that the number of alternative polyadenylation sites was greater than for 5'-SAS.

In addition, variations in trans-splicing efficiency are important in the regulation of specific mRNAs. The process of trans-splicing requires conserved 5'-splice sites, branch points, poly(Y) tracts and 3'-splice sites. Siegel *et al.* found that the position of the poly(Y) tract relative to the 3'-splice sites is important and the presence of an AC dinucleotide at

positions -3 and -4 can lead to a 20-fold decrease in trans-splicing (Siegel et al., 2005). Even though the heterogeneity of trans-splicing and polyadenylation sites makes it difficult to measure the rate of RNA processing at a transcriptome-wide level, results from Fadda *et al.* suggest differences in the rate that many trypanosome mRNAs are spliced (Fadda et al., 2014).

### **RNA stability**

The levels of cellular mRNAs are determined by the rates of transcription, RNA maturation and decay. The regulation of mRNA decay rates is an important factor in determining the expression pattern for many genes in yeast and human cells (Wilusz and Wilusz, 2004) and between 5-8% of human genes contain AU-rich 3'-UTRs elements that may control mRNA turnover (Bakheet et al., 2003). In addition, the enzymes and general pathways of mRNA degradation have been characterized in both yeast and mammalian cells and the current focus of research in these model organisms is how that process can be regulated at multiple levels.

In *T. brucei*, differential mRNA stability plays a dominant role in the post-transcriptional regulation of gene expression (Fadda et al., 2014) and as in yeast and mammalian cells, a large part of the mRNA decay machinery has already been characterized and follows the classic eukaryotic pathway (Clayton, 2014). CAF1-NOT complex deadenylates the mRNA (Erben et al., 2014) and this reaction is followed by XRNA-

mediated 5'-3' degradation (Li et al., 2006) or by exosome-mediated 3'-5' degradation (Clayton and Estevez, 2010). There is also a deadenylation-independent pathway that is initiated by 5' degradation (Fadda et al., 2013; Schwede et al., 2009).

Studies of mRNA degradation in *T. brucei* are based on transcription inhibition. These studies have demonstrated that for most mRNAs, deadenylation precedes degradation (Fadda et al., 2013; Manful et al., 2011) but also that the 5'-3' exoribonuclease XRNA was important for degradation of very unstable mRNAs (Manful et al., 2011). In the most recent study, mRNA decay kinetics were measured in BF and PF by transcription inhibition and RNA sequencing (RNA-seq), revealing complex degradation kinetics and suggesting a role for co-transcriptional degradation in determining mRNA levels (Fadda et al., 2014).

The procyclin mRNA has been one of the best-studied *T. brucei* transcripts and several studies in the last decade of the past century demonstrated the importance of the 3'-UTR in conferring stability to the transcript in PF. By means of reporter assays, positive and negative motifs controlling the mRNA stability were found in the 3'-UTR and therefore the abundance of the procyclin transcript (Furger et al., 1997; Hotz et al., 1997). Since then, several publications have shown the role of 3'-UTRs in the stability of differentially expressed genes transcribed by RNA polymerase II, but the identification of conserved motifs has been in general elusive (Droll et al., 2013; Erben et al., 2014; Fernández-Moya et al., 2014; Walrad et al., 2012). This could be partially explained by the

action of different RNA binding proteins and combinations of them regulating the complex mRNA metabolism in *T. brucei* (Kolev et al., 2014). Taking into account that the procyclin mRNA has been the focus of research in this topic in the last 20 years, it is not surprising that several of the factors that regulate its fate have already been described (Mani et al., 2011; Walrad et al., 2009).

The importance of 3'-UTRs for the regulation of transcription (Mignone et al., 2002) and of 5'- and 3'-UTRs for translation (Jackson et al., 2010) has been shown in other eukaryotes. There are several examples of the role of 3'-UTRs on the stability of mRNAs in *T. brucei* and it exists the possibility that both 5'- and 3'-UTRs exert regulation on translation (see the section about translational efficiency below). However, the possibility of editing the UTRs in *T. brucei* without the inclusion of resistant markers was not possible using the current tools available. For this reason, the implementation of endonuclease-based tools like the CRISPR-Cas9 system for genome editing are fundamental to address questions regarding the role of the UTRs in the regulation of gene expression in *T. brucei*. In the next section I will introduce the system and the numerous possibilities it offers for the manipulation of genomes.

### **The CRISPR-Cas9 system as a tool for genome editing**

Genome engineering is the process of targeting and modifying a genome, its contexts (e.g., epigenetic marks), or its outputs (e.g., transcripts) (Hsu et al., 2014). Engineering biological

systems opens many possibilities for applications in biotechnology, medicine, drug development and basic science. Now it is possible to directly edit DNA sequences in their endogenous context unraveling the function of specific genes and regulatory elements. In biotechnology, the manipulation of genetic elements and their regulatory partners can be used in the biofuel industry, the development of new drugs and as a therapeutic system for treating genetic disorders (Tebas et al., 2014).

The four main genome editing technologies includes the RNA-guided CRISPR-Cas nuclease system, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and meganucleases derived from microbial mobile genetic elements (Hsu et al., 2014). These technologies are based on programmable sequence-specific endonucleases that can be applied even in biological systems that have not previously been genetically tractable (Carlson et al., 2012; Geurts et al., 2009; Watanabe et al., 2012). ZFNs and TALENs have in common endonuclease catalytic domains present in DNA-binding proteins that produce targeted DNA double-stranded breaks (DSBs) at specific genomic loci. The CRISPR-Cas9 system in turn is based on small RNAs that guide the Cas9 nuclease to the target DNA sequence by Watson-Crick base pairing (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012). The later system is comparatively easier to implement, more specific, efficient and

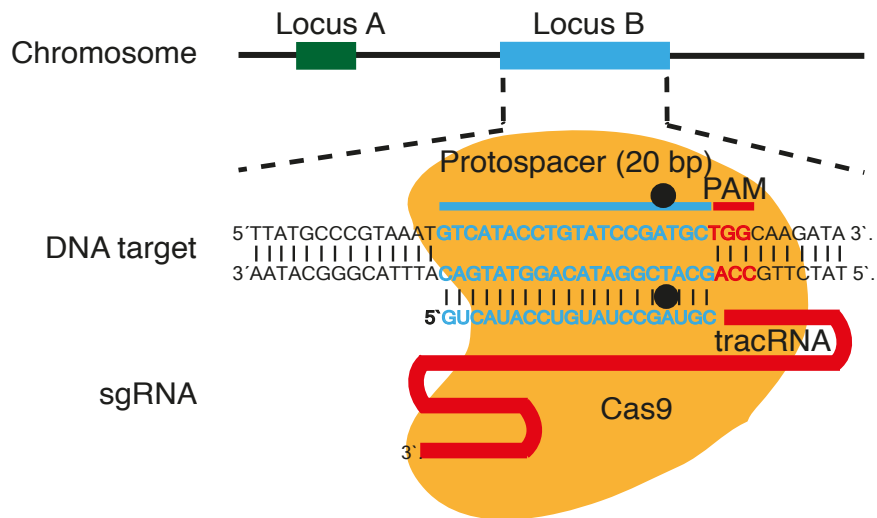
amenable for high-throughput and multiplexed gene editing. The CRISPR-Cas9 system has been adopted in several laboratories worldwide thanks to open-source distributors such as Addgene and online forums and platforms for the design of the so-called chimeric single-guide RNA (sgRNA).

CRISPR (clustered regularly interspaced short palindromic repeats) is a microbial adaptive immune system based in RNA-guided nucleases for cleaving foreign genetic elements (Bhaya et al., 2011; Garneau et al., 2010). Three types (I–III) of CRISPR systems have been described in >40% of sequenced bacteria and 90% of archaea (Mojica et al., 2000). In general, these systems include a cluster of CRISPR-associated (Cas) genes, noncoding RNAs and a distinctive array of repetitive elements (Ran et al., 2013a). The repetitive elements are interspaced by short variable sequences denominated protospacers that are derived from exogenous DNA and that constitute the CRISPR RNA (crRNA) array (Makarova et al., 2011). In addition, each protospacer is associated with a protospacer adjacent motif (PAM) in the target sequence (Ran et al., 2013a). The guide sequence usually corresponds to phage sequences that constitutes the natural mechanism for CRISPR antiviral defense and can be substituted by the sequence of interest to be targeted.

At present, the type II CRISPR system is the best characterized and most commonly applied system for editing

genomes (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012; Saprunauskas et al., 2011). In 2013, two simultaneous studies engineered type II CRISPR systems from *Streptococcus thermophilus* (Cong et al., 2013) and *Streptococcus pyogenes* (Cong et al., 2013; Mali et al., 2013) to achieve genome editing in mammalian cells. The latter consists of the Cas9 endonuclease, the crRNA array and the trans-activating crRNA (tracrRNA) that participates in the processing of the crRNA array. In this system, each crRNA unit includes a 20-nt guide sequence that guides Cas9 to the target sequence. In the most used CRISPR-Cas system derived from *S. pyogenes*, the target sequence must be flanked at the 3' end by a 5'-NGG PAM to mediate Cas9 cleavage at ~3-bp upstream of the PAM (Jinek et al., 2012) (Figure 6). The *S. pyogenes* system has been successfully used in human cell lines, bacteria, zebrafish, yeast, mouse, fruit fly, roundworm, rat, common crops, pig, and monkey (Sander and Joung, 2014) as well that parasitic protozoa like *P. falciparum* (Ghorbal et al., 2014; Wagner et al., 2014), *Toxoplasma gondii* (Shen et al., 2014; Sidik et al., 2014), *T. cruzi* (Lander et al., 2015; Peng et al., 2015) and *L. major* (Sollelis et al., 2015) and *L. donovani* (Zhang and Matlashewski, 2015).





**Figure 6.** Representation of the RNA-guided Cas9 nuclease. The Cas9 endonuclease from *S. pyogenes* spCas9 (yellow) is targeted to a genomic DNA (gDNA) target (locus B) by a sgRNA consisting of a 20-nt guide protospacer sequence (blue) and the tracrRNA scaffold (red). The guide sequence pairs with the top strand of the DNA target (blue bar representing locus B) upstream of a 5'-NGG PAM (red line). Cas9 mediates a DSB ~3 bp upstream of the PAM (black circle).

Computational and metagenomic analyses of bacteria and archaea species have led to the discovery of Cas9 with different PAMs specificities increasing the targeting range of the CRISPR-Cas9 system (Ran et al., 2013a).

In order to improve the efficiency of the system, the Cas9 can be codon-optimized as has been done in human systems (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013). In addition, the crRNA and tracrRNA are fused together to create the sgRNA (Jinek et al., 2012).

Cas9 produces a DSB at a target genomic locus which can undergo two pathways for DNA damage repair: the high-fidelity

homology directed repair (HDR) pathway and the error-prone non-homologous end joining (NHEJ) (Ran et al., 2013a). HDR occurs at lower frequencies than NHEJ but can be used to generate precise editing at a target locus in the presence of a supplied repair template. The basic requirement for the repair template is the presence of homology arms flanking the repair sequence (RS) and can be delivered either as a plasmid or as single-stranded DNA oligonucleotides. The efficiency of this pathway depends of the cell type, the genomic context and the repair template (Saleh-Gohari and Helleday, 2004).

On the other hand, the NHEJ works in the absence of a repair template and the DSBs are repaired in the form of insertion/deletion (indel) mutations. This pathway is used preferentially when the accuracy is not so important as is the case of gene knockouts (KOs) or for introducing long deletions in the genome (Cong et al., 2013). In *T. brucei*, the NHEJ pathway is inefficient and therefore the HDR pathway is used preferentially (Burton et al., 2007).

The nuclease domains of Cas9, HNH and RuvC, have been mutated for additional functions (Saprunauskas et al., 2011). Upon introducing point mutations that inactivate the RuvC or HNH nuclease domains, SpCas9 can be converted into a DNA “nickase” (Cas9n) that creates a single stranded break which is going to be preferentially repaired by HDR in order to decrease the frequency

of off-target DSBs (Gasiunas et al., 2012; Jinek et al., 2012; Sapranaukas et al., 2011). Using a pair of sgRNAs, Cas9n can simultaneously cut both strands to produce a DSB, increasing the specificity of editing (Ran et al., 2013b). Another possibility is to inactivate both catalytic domains generating a dead Cas9 (dCas9) to enable transcriptional regulation, recruitment of fluorescent protein labels or chromatin-modifying enzymes to specific genomic loci. It is even possible to use truncated Cas9 versions to rearrange the three-dimensional organization of the genome (Ran et al., 2013a). Recently, a Cas9 endonuclease encoded by *Francisella novicida* (FnCas9) was guided to mRNAs by a tracrRNA and a small CRISPR/Cas-associated RNA (scaRNA) decreasing the mRNA stability (Sampson et al., 2013). To complement this finding, the RNA-targeting CRISPR immune system present in *Pyrococcus furiosus* and *Sulfolobus solfataricus* includes the Cas RAMP module (Cmr) proteins that interact with small CRISPR RNAs targeting and cleaving complementary RNA instead of DNA (Hale et al., 2012). Therefore, these systems expand the possibilities for targeting RNAs and not only DNA sequences.

Among the limitations of the CRISPR-Cas9 system stand the specificity of the PAM sequence, which depending on the genome can occur more or less frequently. For example, SpCas9 requires a 5'-NGG PAM sequence that can be found on average every 8–12 bp in the human genome (Cong et al., 2013; Hsu et al., 2013). The

other limitation is the potential for off-target mutagenesis that nevertheless can be predicted, experimentally determined and minimized. The extent a given sequence exhibits off-target activity depends of the Cas9 concentration and the abundance of similar sequences in the target genome. The Cas9 concentration can be carefully titrated and, to minimize off-target activity, it is important to assess the probability a protospacer may present off-target sites. As a rule, more than three mismatches are not tolerated and the 8–14 bp on the 3'-end of the protospacer are less tolerant to mismatches than the bases in the 5'-end (Hsu et al., 2013). Off-target indel analysis can be performed using ClustalW, Geneious, VarScan2 or custom sequence analysis scripts (Koboldt et al., 2012; Ran et al., 2013a). It is important to mention that off-target integration is less likely when long homology arms flanking the donor cassette are used in the repair templates.

### **Translational efficiency**

In the last few years the importance of translational efficiency for the regulation of gene expression has become apparent in eukaryotes such as yeast (de Godoy et al., 2008; Lu et al., 2007), humans (Nagaraj et al., 2014), plants (Juntawong et al., 2014), *P. falciparum* (Caro et al., 2014) and *L. infantum* (McNicoll et al., 2006). There are several stages where the translation process can be regulated: mechanisms that involve 5'-end-dependent initiation such as formation of 43S preinitiation complexes, attachment of 43S complexes to mRNA, ribosome scanning of mRNA 5'-

UTRs, initiation codon recognition, commitment of ribosomes to a start codon, ribosomal subunit joining and reinitiation after a short uORF (Jackson et al., 2010). Translation initiation factors and RNA-binding proteins binding to polyadenine (Poly(A)) tails, 3'-UTRs such as PABP and proteins belonging to the Pumilio family are among the elements involved in the regulation of translation (For a review see Jackson et al., 2010).

Regarding *T. brucei*, there is some evidence suggesting that translational regulation might play a role controlling protein levels. First, Butter *et al.* found only moderate correlation (Pearson coefficient 0.57) between the transcriptome and the proteome in a comparative study in *T. brucei* using stable isotope labeling by amino acids in cell culture (Butter et al., 2013). This result suggests a strong regulatory role of translation efficiency and protein stability on the *T. brucei* protein levels. Second, the genome of *T. brucei* and other Kinetoplastids encode at least six eIF4E and four eIF4G homologues with different affinities for cap-binding complexes (Freire et al., 2011, 2014). The two isoforms of the eIF4A helicase function as a general translation factor and as part of the exon junction machinery after splicing respectively (Dhalia et al., 2006). Thus, it has been speculated that different combinations of these isoforms can exert regulation for translation initiation. Additionally, the sequences in the 5'- and 3'-UTRs as well as those surrounding the first AUG initiation codon can play a role in translational regulation in other eukaryotes (Jackson et al., 2010). Currently there is a lack of evidence that UTRs regulate translation in Trypanosomatids, but it is possible that revealing

information come up soon given the evidence present in other organisms and to the development of more advanced technologies in the field of translational regulation of gene expression such as ribosome profiling. Finally, there is evidence for the enrichment of favored codons in highly expressed genes. It has been shown that synonymous codon bias is correlated with tRNA gene copy number and with protein expression level in *T. brucei*, *T. cruzi* and *L. major* (Horn, 2008).

Therefore, it is important to determine whether translational regulation plays a relevant role in the regulation of gene expression in *T. brucei*. For this reason I implemented ribosome profiling for the genome-wide study of gene expression regulation at the level of translation in BF and PF.

### **Ribosome profiling for the study of gene expression regulation at the level of translation**

During the translation process the ribosome binds the mature mRNA template and then reads it for protein synthesis. Translation is of pivotal importance in gene expression control and is tightly regulated to avoid the production of toxic proteins and because is energetically costly (Brar and Weissman, 2015). Although the understanding of the structure and function of the ribosome and the other molecules associated with it is very advanced, the knowledge about the regulation of translation is more limited (Brar and Weissman, 2015).

Before the implementation of ribosome profiling in 2009 (Ingolia et al., 2009), monitoring gene expression on a genome-wide scale was done by measuring mRNA levels using microarrays or RNA-seq. Ribosome profiling is defined as the deep sequencing of ribosome-protected mRNA fragments, allowing global monitoring of translation *in vivo* (Brar and Weissman, 2015). The method has been extensively used to study the regulation of gene expression under different biological conditions and the mechanisms of protein synthesis as well as for the correct annotation of coding regions.

The basic principle of this tool is that ribosomes involved in translation protect mRNAs of about 26-34 nt from RNase activity (Steitz, 1969; Wolin and Walter, 1988). Sequencing of these ribosome-protected fragments or footprints provides information about the position of the ribosome once the translation has been stopped. The density of the footprints sequenced reflects the number of ribosomes at a given position and, assuming that the average translation elongation rate is similar for all genes, ribosome profiling allows the measurement of the rate of protein synthesis. In addition, the footprint positions indicate the identity of the translation products. Analysis of the datasets retrieved permits the discovery of putative new CDSs, the identification of uORFs and regulatory pauses (Brar and Weissman, 2015; Brar et al., 2012; Ingolia et al., 2011). Recently, the monitoring of

translation mediated by subsets of ribosomes has been possible based in their localization in the cell or in their interaction with molecular partners (Brar and Weissman, 2015).

Ribosome profiling requires the immobilization of ribosomes to inhibit translation; RNase digestion to produce the footprints; isolation of the monosome fraction; purification of RNA; generation of a strand-specific library; next-generation sequencing and mapping of the resulting reads to a reference genome (Ingolia et al., 2012). Often, mRNA libraries are generated in parallel to determine the cell transcriptome. The comparison between the abundance of mRNAs and footprints can yield genome-wide information on translational efficiency (Ingolia et al., 2009).

Ribosome profiling has been implemented in viruses, bacteria, yeast, protozoa, nematodes, flies, plants, fish, mice, human cells, mitochondria and chloroplasts (Andreev et al., 2015; Bazzini et al., 2012; Brar et al., 2012; Ingolia et al., 2009, 2011; Juntawong et al., 2014; Oh et al., 2011; Rooijers et al., 2013; Stadler and Fire, 2011; Stern-Ginossar et al., 2012; Zoschke et al., 2013).

The first part of my thesis describes the implementation of ribosome profiling in *T. brucei*. This was the first report of genome-wide regulation of translation in protozoan parasites (Vasquez et al., 2014). After this pioneering publication two other papers have been published in *T. brucei* (Jensen et al., 2014; Parsons et al.,



2015), one in *T. cruzi* (Smircich et al., 2015) and one in *P. falciparum* (Caro et al., 2014). Using the resulting dataset I was able to detect several putative ncRNAs including transcripts derived from the CTR.

The main advantages of ribosome profiling are the sensitivity and precision in quantifying the rate of protein synthesis. The sensitivity results from the depth of sampling, which facilitated the detection and quantification of low abundance footprints. In turn, the precision results from the positional information provided by the sub-codon resolution of the technique. The range of detection of transcripts being translated depends on the variability found for transcripts with low numbers of sequencing reads. Currently, ribosome profiling provides more quantitative and positional information than complementary methods like pulsed label-based mass spectrometry, analyses of transcript distributions on polysome gradients and 35S Met-based metabolic labeling (Brar and Weissman, 2015).

Furthermore, the precision of ribosome footprint positions allows the detection of codon periodicity (Ingolia et al., 2009) and to identify ribosomal frameshifting (alternative translation of an mRNA sequence by changing the ORF), stop codon readthrough, ribosome pausing, translation initiation at non-AUG codons and uORF translation (Brar et al., 2012; Gerashchenko et al., 2012; Ingolia et al., 2009, 2011; Li et al., 2012; Liu et al., 2013).

Although ribosome profiling has opened the possibility of obtaining instantaneous measurements of the translational profile of a cell, there are some limitations that should be considered such as the inference of protein synthesis rates, the presence of contaminating footprint-sized fragments that co-migrate in sucrose gradients with the ribosomes and the mapping of ambiguous reads.

In summary, the main applications of ribosome profiling are the quantification of protein synthesis and the identification of the full set of ribosome-translated polypeptides in a cell, including sets of translated non-canonical coding regions in 5'-UTRs and uORFs that may mediate translational regulation (Ingolia, 2014).

### **Protein stability**

The last level of gene expression regulation is at the protein level, either by post-translational modifications (PTMs) affecting protein activity and binding capacities or protein turnover via proteosomal and lysosomal degradation. PTMs of proteins are important for the cellular localization of proteins, the regulation of protein function and protein complex formation (Mocanu et al., 2009). Proteins can be affected by hundreds of different modifications and combinations of them, which change their physical and biochemical properties and increase the complexity of the cell proteome. Of these modifications, the most common and naturally occurring are cleavage, acetylation, formylation, methionine

oxidation, phosphorylation, ubiquitination, and glycosylation (Mocanu et al., 2009). PTMs diversify and extend protein function beyond the amino acid sequence allowing eukaryotic cells to regulate their physiological states. These modifications are usually analyzed by western blot and mass spectrometry. Although recent advances in mass spectrometry have facilitated the study of PTMs, the current technology is still far of the power and resolution offered by technologies like next generation sequencing (NGS) used for the analysis of nucleic acids.

The progress in the study of gene regulation at the protein level in *T. brucei* is not as complete as in the case of trans-splicing efficiency and RNA stability, but some information is available. For example, the ubiquitin-mediated protein degradation pathway has been described (Li et al., 2002) together with its relation with the developmental regulation of invariant surface protein expression (Leung et al., 2011). There are also several comparative proteomic studies between BF and PF that highlight the importance of translational regulation and/or protein stability in gene regulation (Butter et al., 2013; Gunasekera et al., 2012; Urbaniak et al., 2012) but none of them include ribosome profiling datasets that allow to determine the individual importance of these two levels of regulation.

In summary, all these different levels of post-transcriptional regulation for RNA polymerase II transcribed genes allow the parasite to cope with the challenges is exposed in extremely dissimilar environments such as the bloodstream in the mammalian host and the digestive tract in the tsetse fly. These different interconnected regulatory events are integrated in a very precise

way for generating physiological responses that differs in each life cycle stage and maybe even in different trypanosome strains and species. Nevertheless, the trypanosome scientific community is just starting to unravel how this gene regulation is coordinated in such fine-tuned manner.

### **Regulation in RNA polymerase I transcribed genes**

In *T. brucei*, RNA polymerase I transcribes *VSG* and procyclin genes in addition to the conventional transcription of rRNA genes that occurs in other eukaryotes (Günzl et al., 2003). They are organized also in PTUs and the initiation of transcription occurs at well-defined promoters unlike RNA polymerase II transcribed genes. *VSG* and procyclin genes are trans-spliced and polyadenylated in a similar fashion as pre-RNAs transcribed by RNA polymerase II (Nilsson et al., 2010). As was mentioned in the section “RNA stability”, procyclin genes have been the target of intensive research regarding the role of 3'-UTRs in transcript stability and developmental regulation (Furger et al., 1997; Hotz et al., 1997).

However, RNA polymerase I transcribed *VSG* genes present further regulation in addition to the post-transcriptional mechanisms mentioned above. Research done in the last 15 years has shown the involvement of epigenetic mechanisms for the regulation of gene expression. The best-studied expression unit in *T. brucei* is the active BES. The active BES contains less nucleosomes compared to the silent BESs, phenomenon that can be explained by the differences in transcriptional activity (Stanne and Rudenko, 2010). In addition, the protein TbTDP1 has been found enriched at the active BES (Narayanan and

Rudenko, 2013). This protein promotes the decondensation of chromatin, facilitating the access of regulatory chromatin factors (Reeves, 2010; Štros, 2010). The recruitment of chromatin remodeling factors has become evident in recent studies focused in the epigenetic regulation at BESs. One of these remodelers, TbISWI, repress the RNA polymerase I transcription in the BESs (Stanne et al., 2011). Other findings suggest that the histone deacetylases TbHDAC1 and TbHDAC3 play a role in the BESs promoter silencing (Wang et al., 2010). A telomere binding protein TbRAP1, mediates the silencing of BESs (Yang et al., 2009), as well as other protein present in the telomeres, the silent information regulator 2 related protein (TbSIR2RP1), which participates in the generation of a silencing gradient from the telomeres to the promoter of the BESs (Rudenko, 2010). In addition to the proteins bound to telomeres, the length of these regions has been correlated to the frequency of switching. The shorter they are the higher the switching frequency, a phenomenon that happens in strains recently isolated from the field (Hovel-Miner et al., 2012). It is worth to mention that Base J is enriched at silent BESs, especially towards the telomere region. All these studies have pinpointed the relevance of the telomere regions for the regulation of *VSG* expression.

One of the questions regarding how BF trypanosomes transcribe only one *VSG* gene of a multi-allelic family of *VSG* expression site loci has partially addressed by the identification of the extranucleolar ESB (Navarro and Gull, 2001). The ESB is only present in BF and contains the active BES where RNA polymerase I transcribes the single *VSG* expressed in a particular time. Navarro and Gull have described the ESB as a discrete region where the recruitment of

that single active BES and the concomitant exclusion of the other silenced BES occurs, which in turn define the mechanism responsible for the mono-allelic expression of *VSGs* (Navarro and Gull, 2001).

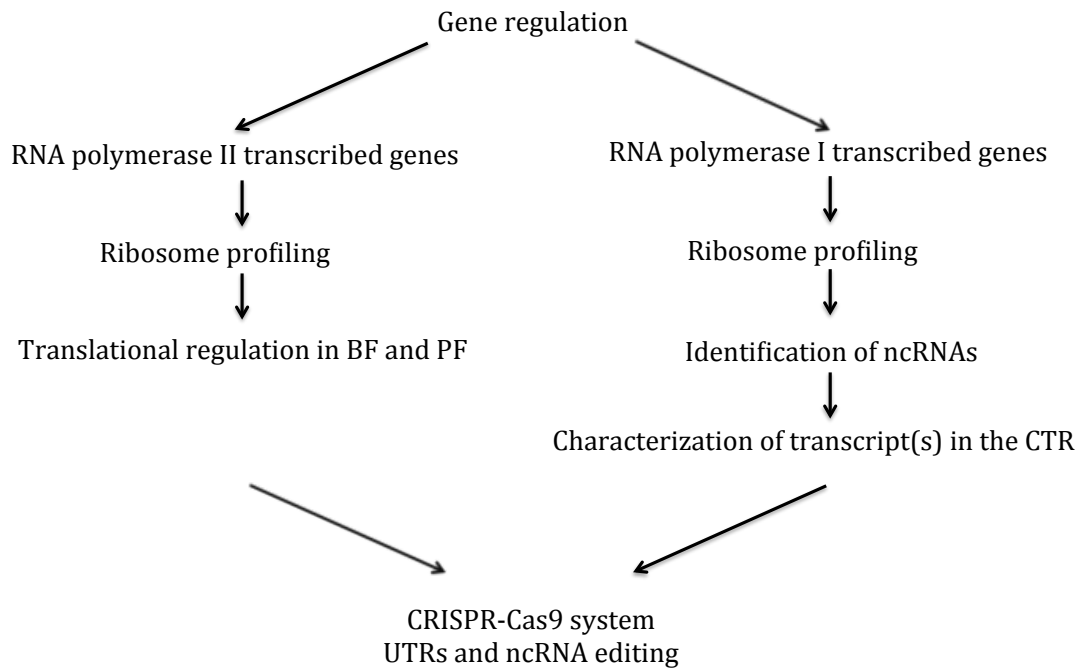
In the section titled “The co-transposed region (CTR) has a role in *in situ* switching”, I mentioned the association between the presence of the CTR and the stability of the active BES. It is possible to hypothesize the mechanism(s) responsible for that stabilizing effect. One way could be at the DNA sequence level, by providing binding-sites for proteins that keep the active BES in an open chromatin conformation and thereby allow for active transcription of the *VSG* or by anchoring the active BES to the ESB. Other way could be at the RNA level, where the CTR-derived ncRNA confers the stabilizing function, either in *cis* by promoting the transcription from the active BES or in *trans* by repressing the other BES.

## **Aims**

In *T. brucei*, the regulation of gene expression occurs mainly at the post-transcriptional level. Post-transcriptional mechanisms such as trans-splicing efficiency and RNA stability have been the focus of active research and there is a relative complete comprehension of these processes. In contrast, there are few studies addressing the role of translational regulation and protein turnover in the regulation of gene expression. Therefore, the first question I was interested in answering in my thesis was if the regulation of translation plays a significant role in the regulation of gene expression in the two life cycle stages of the parasite (Figure 7). In order to answer this question I implemented ribosome

profiling in BF and PF to evaluate on a genome-wide scale the degree of gene regulation due to translational regulation.

In addition to evaluate the translational efficiency in two life cycle stages, the ribosome profiling approach allowed the detection of putative ncRNAs in a genome-wide fashion. One of these putative ncRNAs was detected upstream of the *VSG* locus in the active BES of the Lister 427 BF, a region generally referred to as co-transposed region (CTR). It has been suggested that this region may contain a stabilizing effect on the active BES (Davies et al., 1997). Therefore, in the second part of my thesis I characterized a ncRNA transcribed from the CTR (Figure 7). As part of the CTR characterization, I was interested in mapping the precise location of the stabilizing element by systematically deleting regions of the CTR. Therefore, the third part of my thesis deals with the implementation of the CRISPR-Cas9 technology in *T. brucei* in order to delete these regions without introducing resistant markers that can alter other regions of the BES (Figure 7). The CRISPR-Cas9 system can be used as well for editing sequences involved in translational regulation, as is the case of UTRs.



**Figure 7.** Flow chart illustrating the aims of the thesis. The first aim corresponded to the genome-wide study of translational regulation in two life cycle stages of the parasite (left). The second aim was the characterization of a ncRNA predicted from the ribosome profiling results (right). The third aim refers to the implementation of the CRISPR-Cas9 system in *T. brucei* with the purpose of mapping the stabilizing element in the CTR and editing UTRs to study translational regulation (bottom).

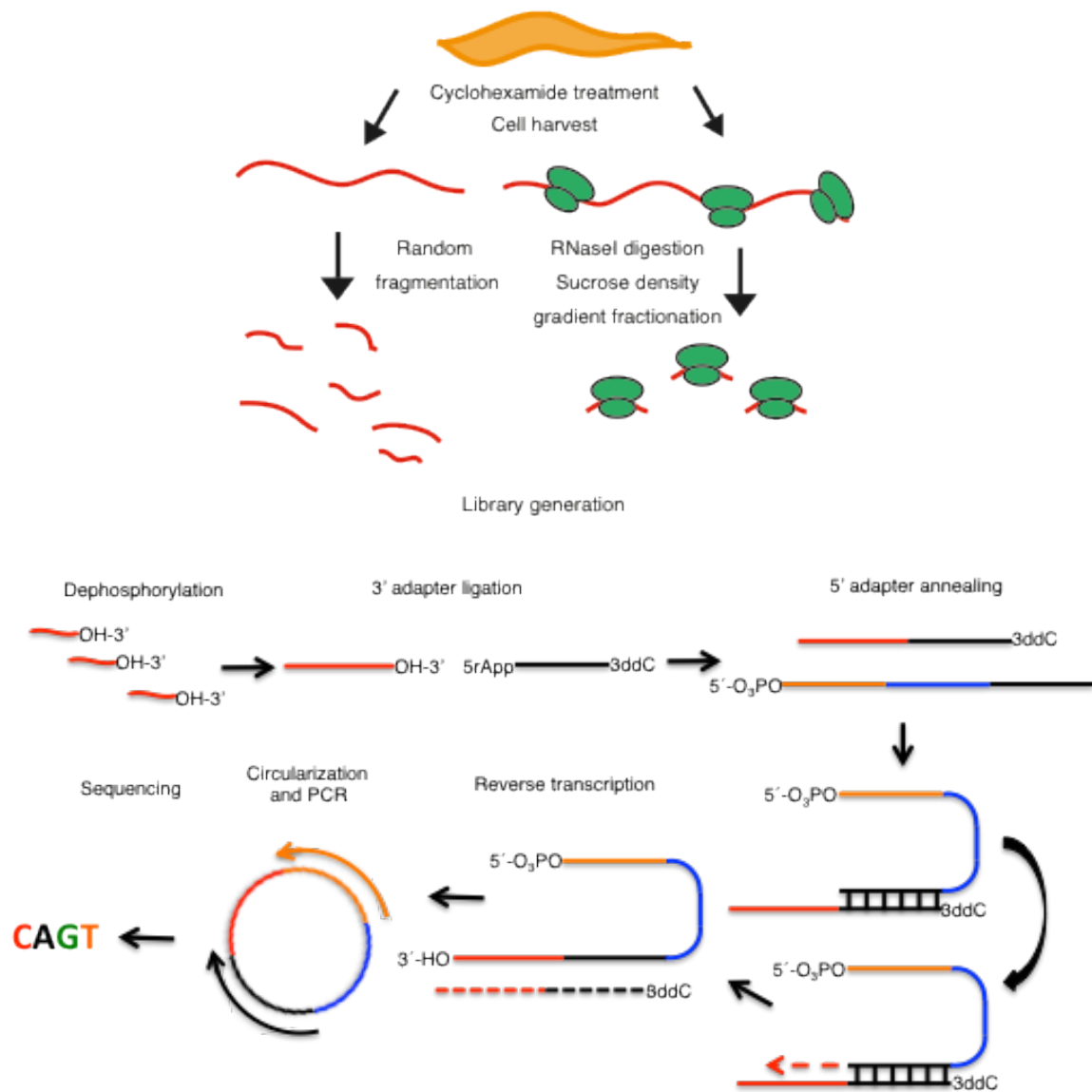


## **Chapter 2: Materials and Methods**

You can see the lists including the reagents, equipment and consumables used in this work at the end of the chapter.

### **Ribosome profiling**

The ribosome profiling (ribo-seq) protocol described here is based on the methodology published by Ingolia *et al.* (Ingolia et al., 2012). For an outline of the ribo-seq protocol used for *T. brucei* PF and BF see Figure 8.



**Figure 8.** Establishment of a ribo-seq protocol. Outline of the protocol used to convert ribosome footprints and fragmented mRNA into sequencing libraries. OH-3': 3'-hydroxyl group, 5rApp: Adenylation-5' (rApp), 3ddC: dideoxycytosine, 5'-O<sub>3</sub>PO: 5' phosphorylation. Adapted from Ingolia *et al.* 2009.

### ***T. brucei* culture and cell harvest**

Wild type BF of Lister 427 (MITat 1.2 clone 221) were cultured at 37 °C in HMI-11 up to a cell density of 1.5 x 10<sup>6</sup> cells/ml. Wild type PF of *T. brucei* strain Lister 427 were cultured at 27 °C in SDM-79 medium supplemented with 10% fetal bovine serum and hemin (7.5 mg/l) up to a cell density of 10<sup>7</sup> cells/ml.

Cycloheximide was added to a final concentration of 100 µg/ml and the cell cultures were gently shaken for 2 minutes (min) at room temperature.  $1 \times 10^9$  cells were collected by centrifugation at 3,000 x g, 4 °C for 5 min, washed with 1 ml of polysome lysis buffer (10 mM Tris-HCl pH 7.4, 300 mM KCl and 10 mM MgCl<sub>2</sub>) (Jensen *et al.*, 2005), transferred to a 1.5-ml microcentrifuge tube and pelleted again at 10,000 x g, 4 °C for 30 seconds (sec). To lyse the cells, 360 µl polysome lysis buffer, 40 µl of 10% n-octylglycoside and 20 units of TURBO DNaseI (Ambion) were added per  $1 \times 10^9$  cells and incubated for 30 min on ice. The lysate was centrifuged at 16,000 x g, 4 °C for 10 min, the supernatant transferred to a new microcentrifuge tube and the OD<sub>260</sub> was determined using a Nanodrop 2000. In this step, aliquots of the extract were frozen down in liquid nitrogen and stored at -80 °C for at least 6 months.

### **Nuclease digestion and monosome purification**

For both BF and PF, 200 µl aliquots of the lysate (OD<sub>260</sub> = 40) were digested with RNase I (Ambion) at room temperature (1,200 units) or on ice (1,600 units). After 1 hour, the digestions were stopped by adding 100 units of SUPERase•In RNase inhibitor (Ambion) to the RNase-treated samples. In parallel, 100 units of SUPERase•In RNase inhibitor were added to a 200 µl aliquot of lysate not containing RNase I (undigested control). Monosome purification were performed using sucrose gradients as described previously including some modifications (Ingolia *et al.*, 2009).

The samples were loaded onto sucrose gradients for monosome purification as soon as possible after stopping the digestions. The rotor SW40 was pre-chilled at 4 °C inside an Optima™ L80 XP ultracentrifuge (Beckman

Coulter) and the buckets in a fridge. Next, 10% and 50% (w/v) sucrose solutions were prepared in polysome gradient buffer (10 mM Tris-HCl [pH 7.4], 300 mM KCl, 10 mM MgCl<sub>2</sub> and 100 µg/ml cycloheximide) into 14 mm x 95 mm SW40 ultracentrifuge tubes. The centrifuge tubes were set inside a metallic support provided with the Gradient Station (Biocomp) and filled up with 10% sucrose solution (6-7 ml) to the line marked in the metallic tube. Using a syringe, the 50% sucrose solution was added from the bottom leaving empty 4 mm on the top of the tube. The rubber cap was gently inserted into the top of the tube in a way the hole present in the cap entered the last allowing air to escape. Once the sucrose solutions were added to all tubes, they were placed in the tube holder of the Gradient Station. The program "top 10%/ bottom 45%" (rotation at 82 °C, speed 19, for 1:25 min) was chosen in the Generation option of the Gradient Station (software version 2.07). Then, the samples were pipetted onto the sucrose density gradients and weighted on the precision balance adding additional 10% sucrose solution to the lighter gradients. The tubes were introduced into the SW40 pre-chilled buckets and the caps sealed with a screwdriver. The samples were centrifuged at 100,000 x g, 4 °C for 2.5 hours.

Once the samples were centrifuged, the absorbance of the collected material was measured Model EM-1 Econo™ UV Monitor (254 nm, 1.0 absorbance) to identify the 80S monosome peak. The undigested samples were analyzed first to ensure that they contained intact polysomes and to determine the approximate location of the 80S peak. Then, the digested samples were analyzed and the 80S monosome peak collected, immediately frozen down in liquid nitrogen and stored at -80 °C. The digested samples should contain larger

80S monosome peaks than the undigested sample and no polysome fraction.

The fractionation mode of operation on the Gradient Station was selected to collect the samples using the manual-triggered gradient option, setting the absorbance in 1.0 unit on the UV device and a speed of 0.2 units and distance/fraction of 80 mm in the Gradient Station. As soon as the 80S peak appeared on the screen, the sample was recovered in a 15 ml Falcon tube and immediately frozen down in liquid nitrogen and stored at -80 °C.

### **Total RNA extraction**

A fraction of the undigested sample was used to generate mRNA libraries for sequencing. Total RNA collected in the monosome fractions was purified using hot acid Phenol-Chloroform-Isoamyl alcohol (Ingolia, 2010). The monosome fraction samples were placed at 65 °C and SDS (30%) was added to a final concentration of 1% (w/v). Then, one volume of pre-warmed acid PCI (Phenol-Chloroform-Isoamyl alcohol v/v/v 25:24:1) was added to the samples and incubated at 65 °C for 5 min (vortexing frequently). Typically, I added 2-4 ml of PCI and 66.7-133 µl of 30% SDS to 2-4 ml of each monosome fraction sample. The samples were incubated on ice for 5 min and centrifuged at 16,000 x g, 4 °C for 5 min (4,800 x g, 4 °C for 15 min when using 15 ml Falcon tubes for the footprint samples). The upper aqueous phase was recovered, one volume of acid PCI was added again and the samples were then incubated for 5 min at room temperature (vortexing frequently). The tubes were centrifuged at 16,000 x g, 4 °C for 5 min. (4,800 x g, 4 °C for 15 min if using 15 ml Falcon tubes for the footprint samples). The upper aqueous phase was recovered for second time and

the RNA was precipitated in low-bind tubes after the addition of 1/10 volume of sodium acetate 3M (pH 5.5) (Ambion), 0.5  $\mu$ l GlycoBlue (Ambion) and 1.5 volume of isopropanol. The samples were stored at -20 °C for at least 30 min (or overnight) and then centrifuged at 16,000 x g, 4 °C for 30 min if the samples were in microcentrifuge tubes or at 4,800 x g, 4 °C for 60 min if the samples were in 15 Falcon tubes. The supernatants were discarded and the pellets washed with 800  $\mu$ l of 80% ethanol. The samples were centrifuged at 16,000 x g, 4 °C for 20 min and the supernatants discarded. The tubes were spun down for 30 sec to remove the remaining ethanol with a tip and then air-dried for 10 min. The RNA was resuspended in 20  $\mu$ l of 10 mM Tris-HCl (pH 8). The concentration was measured in a Nanodrop 2000 (the optimal 260/280 and 260/230 ratios should be over 2.0 for good quality RNA). Finally, the samples were stored at -80 °C.

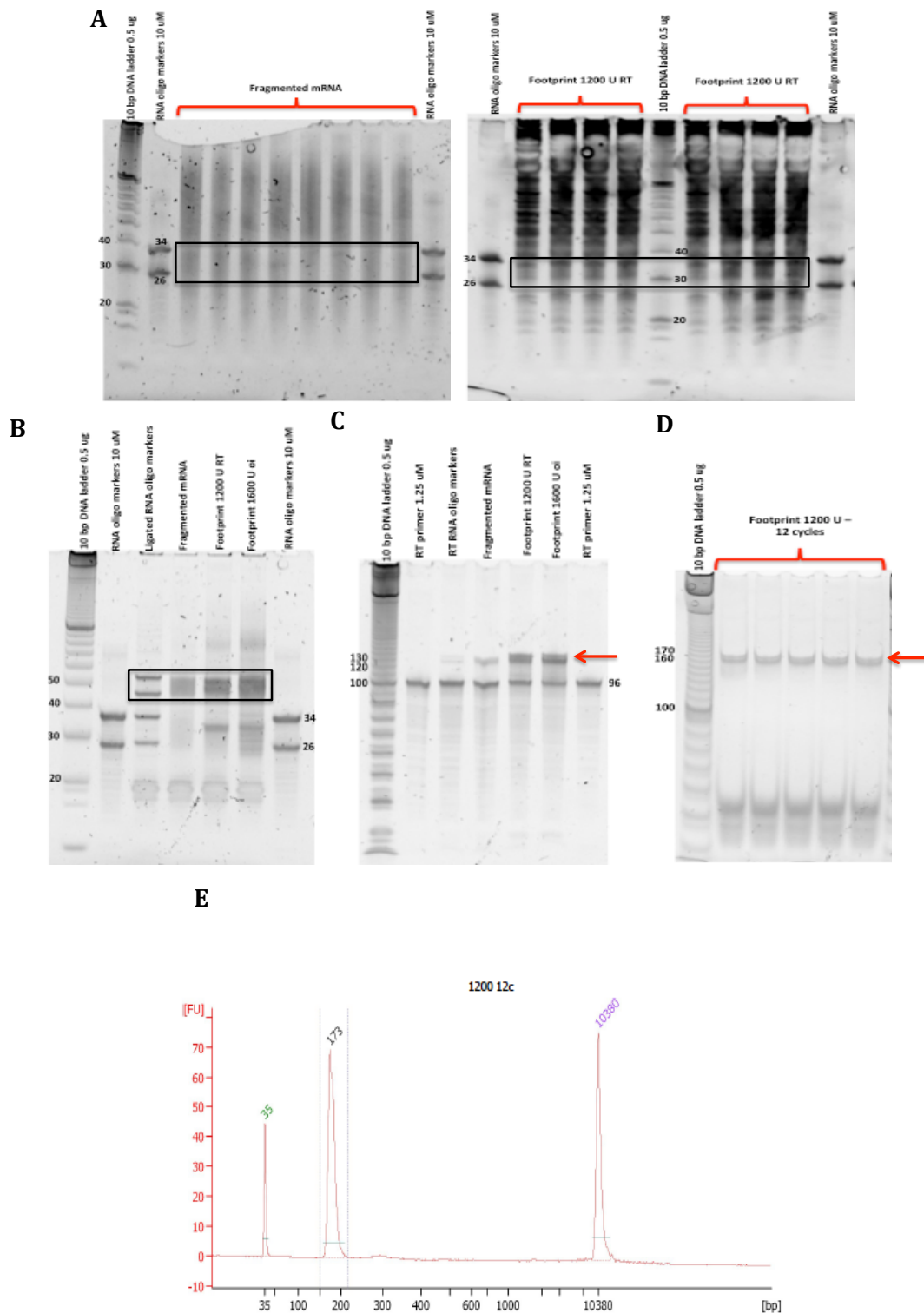
### **PolyA-enrichment and fragmentation of undigested RNA**

To generate mRNA libraries for sequencing, undigested RNA was polyA-enriched using a Dynabeads® mRNA Purification Kit (Ambion) according to the manufacturer's instructions. Between 2-5  $\mu$ g of polyA-enriched RNA was fragmented by incubation with an RNA Fragmentation Reagent (Ambion) at 70 °C for 30 min.

### **Fragmented RNA and footprint purification**

Both, ribosome footprints and fragmented mRNA (26–34 nt) were size-selected by electrophoresis using 15% (wt/vol) denaturing-PAGE gels and two custom-made (IDT) synthetic RNA markers NI-NI-19 and NI-NI-20 (See the section Oligonucleotides, primers and cell lines at the end of this chapter). Each

fragmented and footprint sample was adjusted to a final volume of 45  $\mu$ l with 10 mM Tris-HCl (pH 8). Next, 45  $\mu$ l of 2x Loading buffer was added to bring each sample to a final volume of 90  $\mu$ l. To avoid overloading and to ensure good resolution, the 90  $\mu$ l were loaded into 8 wells, 11.2  $\mu$ l per well. One  $\mu$ l of a 10-bp DNA ladder (1  $\mu$ g/ $\mu$ l) (Invitrogen) and 1  $\mu$ l of each synthetic RNA marker (10  $\mu$ M) were loaded into each gel as well. The samples were denatured for 90 sec at 80 °C and put immediately on ice. The gels were pre-run for 30 min, run for 65 min at 180 V and stained with SYBR Gold 1x for 3 min (in 1x TBE gel-running buffer). The region between 34-nt and 26-nt in the fragmented, footprint samples and in the RNA markers was excised. Each excised gel slice was placed in a clean 0.5 ml low-bind RNase-free tube (Figure 9a).



**Figure 9.** Representative gels from intermediate product purification in the library generation process. (A) Size selection of fragmented mRNA (left) and ribosome footprint fragments (right). Both samples were derived from PF and in the case of the footprint fragments they were digested with 1,200 U of RNaseI at room temperature (RT). In both cases the region between 34-nt and 26-nt was excised from the gels as showed by the black rectangle. (B) Purification of ligation



products. After the ligation of the linker, the ligation products had a length between ~ 45 and 55-nt. In the gel are shown the RNA oligo markers without the linker and with the linker and fragmented mRNA, and footprint fragments obtained upon digestion with 1,200 U and 1,600 U of RNaseI at room temperature or on ice respectively. The region excised from the gel is showed by the black rectangle. (C) Purification of reverse transcription products. The length of the resultant cDNA was around 120-130 nt. The red arrow indicates the gel band that was excised. This time instead of the RNA oligo markers, the reverse transcription primer (RT) was included in the gel as a guide. (D) Purification of polymerase chain reaction (PCR) products. The PCR products corresponding to 12 cycles of amplification of the footprint sample digested with 1,200 U of RNaseI at room temperature are shown in the gel. The red arrow indicates the ~173-nt product band that was excised from the gel. (E) BioAnalyzer profile of a sequencing library. A single 173-nt peak was observed for the footprint sample digested with 1,200 U of RNaseI at room temperature. The peaks at 35 and 10380-nt are the manufacturer's internal standard, which were included in all samples. The 10-bp DNA ladder was loaded in all the gels.

The RNA was extracted in the following way: the gel fragments were placed in a 0.5 ml low-bind RNase-free tube, piercing a hole into the bottom and forcing the gel into a 1.5 ml low-bind RNase-free tube by centrifugation (16,000 x g, room temperature for 2 min). The RNA was eluted by adding 0.5 ml of 0.3 M NaCl (Ambion) and placing the tubes at on rotating wheel at 4 °C, overnight. The resulting gel slurry was placed in a microcentrifuge tube spin filter and centrifuged at 16,900 x g for 2 min. Next, 1.5 µl of GlycoBlue and 500 µl of isopropanol were added, the samples were mixed by vortexing and placed at -20 °C for at least 30 min (or overnight). Subsequently, the samples were spun at 16,000 x g at 4 °C for 30 min, followed by the removal of the supernatant and the addition of 1.0 ml of 80% ethanol (mixed by vortexing). The samples were spun again at 16,000 x g, 4 °C for 20 min, the supernatant removed and followed by another brief spin to remove residual ethanol. The pellets were air-dried for 10

min. The size-selected RNA was resuspended in 10  $\mu$ l of 10 mM Tris (pH 8) and transferred to a clean low-bind RNase-free tube. RNA may be stored at  $-80^{\circ}\text{C}$ .

### **Preparation of RNA footprint and mRNA-sequencing libraries**

Sequencing libraries were prepared as described (Ingolia *et al.*, 2012) except for the omission of a rRNA depletion step during the footprint library preparation.

### **RNA dephosphorylation of the 3'-end**

Thirty three  $\mu$ l of RNase-free water were added to the fragmented and footprint samples (resuspended in 10  $\mu$ l 10 mM Tris-HCl) and then denatured for 90 sec at  $80^{\circ}\text{C}$ . The samples were equilibrated to  $37^{\circ}\text{C}$ , the reaction tabulated below was set-up and an incubation step was performed for 1 hour at  $37^{\circ}\text{C}$ . Thereafter, the enzyme was heat-inactivated for 10 min at  $70^{\circ}\text{C}$ .

<b>Component</b>	<b>Amount (<math>\mu</math>l)</b>	<b>Final concentration</b>
RNA sample	43.0	
10x T4 PNK buffer	5.0	1x
SUPERase In (20 U / $\mu$ l)	1.0	20 U
T4 PNK (10 U / $\mu$ l)	1.0	10 U

The RNA was precipitated by adding 39  $\mu$ l of water, 1  $\mu$ l of GlycoBlue and 10  $\mu$ l of 3 M sodium acetate (pH 5.5). The components were mixed and then 150  $\mu$ l of isopropanol were added to the tubes. Precipitation was carried out at  $-20^{\circ}\text{C}$  for at least 30 min (or overnight). The RNA was pelleted by centrifugation at 16,000 x g,  $4^{\circ}\text{C}$  for 30 min in a tabletop microcentrifuge carefully pipetting out all liquid from the tube

immediately afterwards. Next, 1 ml of 80% ethanol was added and the RNA was pelleted by centrifugation at 16,000 x g, 4 °C for 20 min in a tabletop microcentrifuge. The pellets were air-dried for 10 min. The dephosphorylated RNA was resuspended in 8.5 µl of 10 mM Tris-HCl (pH 8) and transferred it to a clean low-bind RNase-free tube. RNA may be stored at – 80 °C.

### **Ligation reaction**

First, 1.5 µl of preadenylylated linker (0.5 µg/µl) (See the section Oligonucleotides, primers and cell lines at the end of this chapter) was added to each samples, denatured for 90 sec at 80 °C and then cooled down to room temperature. Then, the following ligation reaction was set-up:

<b>Component</b>	<b>Amount (µl)</b>	<b>Final concentration</b>
RNA and linker	10.0	
10x T4 Rnl2 buffer	2.0	1x
PEG 8000 (50%, wt/vol)	6.0	15% (wt/vol)
SUPERase.In (20 U / µl)	1.0	20 U
T4 Rnl2(tr) (200 U / µl)	1.0	200 U

The samples were incubated for 2.5 hours at room temperature. Then, 338 µl of water, 40 µl of 3 M sodium acetate (pH 5.5), 1.5 µl of GlycoBlue and 500 µl of isopropanol were added to each reaction. Next, RNA was precipitated at – 20 °C for at least 30 min (or overnight). The RNA was pelleted by centrifugation at 16,000 x g, 4 °C for 30 min in a tabletop microcentrifuge, carefully pipetting out all liquid from the tubes. Next, 1 ml of 80% ethanol were added and the RNA was pelleted by

centrifugation at 16,000 x g, 4 °C for 20 min in a tabletop microcentrifuge. The samples were air-dried for 10 min. The ligated RNA was resuspended in 5 µl of 10 mM Tris-HCl (pH 8) and transferred to a clean low-bind RNase-free tube.

The ligation reactions were separated by polyacrylamide gel electrophoresis. A 15% (wt/vol) denaturing polyacrylamide gel was prepared following the procedure described above (Figure 9b). The ligation product bands, including the oligo marker ligation products were excised from the gel. The gel slices were placed in clean low-bind RNase-free tubes. The RNA from these samples was recovered as described above. The ligation products were resuspended in 10 µl of 10 mM Tris-HCl (pH 8) and transferred to a clean low-bind RNase-free tube. RNA may be stored at – 80 °C after this step.

### **Reverse transcription**

Two microliters of reverse transcription primer (See the section Oligonucleotides, primers and cell lines at the end of this chapter) at 1.25 µM were added to each sample, denaturing for 2 min at 80 °C in a thermal cycler and immediately placing the tubes on ice. Then, the reverse transcription reaction was set up in the following way:

<b>Component</b>	<b>Amount (µl)</b>	<b>Final concentration</b>
Ligation and primer	12.0	
5x First-strand buffer	4.0	1x
dNTPs (10mM)	1.0	0.5 mM
DTT (0.1M)	1.0	5 mM
SUPERase.In (20 U / µl)	1.0	20 U
SuperScript III (200 U / µl)	1.0	200 U

The reactions were incubated for 30 min at 48 °C in a thermal cycler. The RNA was hydrolyzed by adding 2.2 µl of 1N NaOH to each reaction and by incubating for 20 min at 98 °C. Next, 20 µl of 3 M sodium acetate (pH 5.5), 2 µl of GlycoBlue and 156 µl of water were added to each reverse-transcription reaction, followed by the addition of 300 µl of isopropanol. Precipitation was carried out at -20 °C for at least 30 min (or overnight). The cDNA was pelleted by centrifugation as mentioned above. The cDNA was resuspended in 5 µl of 10 mM Tris-HCl (pH 8) and transferred to a clean low-bind DNase-free tube.

The reverse-transcription products were separated from the unextended primer on an 8% (wt/vol) denaturant polyacrylamide gel. The gel was pre-run for 30 min and run for 40 min at 180V. Instead of the RNA oligo markers, 2 µl of reverse-transcription primer (1.25 µM) were loaded into the gel (Figure 9c). The reverse-transcription product bands were excised from the gel and placed in a clean low-bind DNase-free tube.

The DNA from the polyacrylamide gel was extracted by placing the gel fragments in 0.5 ml low-bind DNase-free tubes, piercing a hole into the bottom and forcing the gel slices into 1.5 ml low-bind DNase-free tubes by centrifugation (16,000 x g, room temperature for 2 min). To elute DNA, 0.4 ml of DNA gel extraction buffer (300 mM NaCl, 10 mM Tris-HCl [pH 8], 1 mM EDTA) was added and placed on a rotating wheel overnight at 4 °C. The gel slurry was placed on a microcentrifuge tube spin filter and centrifuged for 2 min at 16,000 x g. To precipitate the DNA,

1.5 µl of GlycoBlue and 500 µl of isopropanol were added and the tubes placed at -20 °C for at least 30 min (or overnight). The samples were spun at 16,000 x g, 4 °C for 30 min and the supernatant removed immediately. To wash the DNA, 1 ml of 80% ethanol was added and the tubes spun at 16,000 x g, 4 °C for 20 min. The supernatant was removed again and the resulting pellet air-dried for 10 min. The reverse-transcription products were resuspended in 16 µl of 10 mM Tris-HCl (pH 8) and transferred to a PCR tube. DNA may be stored at – 20 °C.

### **Circularization reaction**

The circularization reactions were prepared as tabulated below:

<b>Component</b>	<b>Amount (µl)</b>	<b>Final concentration</b>
First-strand cDNA	16.0	
10x CircLigase II buffer	2.0	1x
MnCl <sub>2</sub> (50mM)	1.0	2.5 mM
CircLigase II	1.0	100 U

The reactions were incubated for around 16 hours at 60 °C and then heat-inactivated for 10 min at 80 °C in a thermal cycler. Circularized DNA may be stored in the circularization reaction buffer indefinitely at – 20 °C.

Next, 2 µl of GlycoBlue, 10 µl of 3M sodium acetate (pH 5.5), 68 µl of water and 150 µl of isopropanol were added, carrying out the precipitation step at – 20 °C for at least 30 min (or overnight). The circularized DNA was precipitated as mentioned above. The circularized products were resuspended in 10 µl of 10 mM Tris-HCl (pH 8). DNA may be stored at – 20 °C.

## PCR amplification and barcode addition

The PCR reactions were set up as tabulated below:

Component	Amount ( $\mu$ l)	Final concentration
PCR grade Water	20.0	
2x KAPA HiFi Hot start Mix	25.0	1x
Forward library primer (10 $\mu$ M)	1.5	0.3 $\mu$ M
Reverse indexed primer (10 $\mu$ M)	1.5	0.3 $\mu$ M
Circularized DNA template	2.0	
Total	50.0	

The primers are listed in the section Oligonucleotides, primers and cell lines at the end of this chapter. The program used to amplify the DNA is tabulated below:

Cycle number	Denature	Anneal	Extend
1	95 °C – 5 min		
2-15	98 °C – 20 sec	65 °C – 15 sec	72 °C – 15 sec

The number of cycles will depend of the amount of the start material and the efficiency of each of the library preparation steps.

Once the PCR reactions were ready, 10  $\mu$ l of 6x nondenaturing loading dye were added to each tube. Each sample was split in three wells. A ladder sample was prepared with 1  $\mu$ l of 10-bp ladder, 3.4  $\mu$ l of 10 mM Tris-HCl (pH 8) and 1  $\mu$ l of 6x nondenaturing loading dye.

An 8% (wt/vol) polyacrylamide nondenaturing gels was prepared to separate the samples. The amplification reactions for the same sample were loaded in adjacent wells to facilitate direct comparison. The samples were separated by electrophoresis for 40 min at 180 V. The gel was stained for 3 min in 1x SYBR Gold in 1x TBE gel-running buffer. The DNA

was extracted from the polyacrylamide gel as mentioned above for the reverse transcription reaction (Figure 9d). The PCR products were resuspended in 12  $\mu$ l of 10 mM Tris-HCl (pH 8). DNA may be stored at  $-20$  °C. It is particularly important that the gel extractions remain at  $25$  °C or below to avoid the formation of reannealed partial duplexes. Such duplexes will complicate the quantification of the library.

Six sequencing libraries were generated in total: PF fragmented mRNA, PF footprint room temperature (1,200 units of RNase I), PF footprint on ice (1,600 units of RNase I), BF fragmented mRNA, BF footprint room temperature (1,200 units of RNase I) and BF footprint on ice (1,600 units of RNase I).

The six libraries were analyzed on a High sensitivity DNA chip (Agilent) (Figure 9e). The expected size band in the fragmented and the footprint libraries is between 163-175 nt. It is important that the samples sent for sequencing do not present a peak corresponding to the unextended primer.

Illumina Next-Generation Sequencing (NGS) was used for sequencing the libraries according to the manufacturer's protocol. See the Table 1 for information related about the details of the sequencing process.



**Table 1.** Relevant information about the sequencing process of the six ribosome profiling libraries generated in this study.

Sample	BF fragmented mRNA	BF footprint 1200 U	BF footprint 1600U	PF fragmented mRNA	PF footprint 1200U	PF footprint 1600U
Library name	NS025	NS026	NS027	NS028	NS029	NS030
Library construction protocol	TruSeq small RNA Illumina	TruSeq small RNA Illumina	TruSeq small RNA Illumina	TruSeq small RNA Illumina	TruSeq small RNA Illumina	TruSeq small RNA Illumina
Amplification (No. Cycles)	12	15	12	12	12	15
Bar code	CGACGT	GCAGCT	TACGAT	CGACGT	GCAGCT	TACGAT
Index type	4	5	6	4	5	6
DNA dissolved in	10mM Tris, pH 8	10mM Tris, pH 8	10mM Tris, pH 8	10mM Tris, pH 8	10mM Tris, pH 8	10mM Tris, pH 8
Concentration (ng/μl)	2	1.5	1.6	1	2	1.7
Determined by	Bioanalyser	Bioanalyser	Bioanalyser	Bioanalyser	Bioanalyser	Bioanalyser
Sample volume (μl)	15	15	15	10	10	10
Fragment size (range) bp	100-200	100-200	100-200	100-200	100-200	100-200
Requested No. of reads	30 million	30 million	30 million	40 million	40 million	40 million
Sequencing run conditions	100 bp single end	100 bp single end	100 bp single end	100 bp single end	100 bp single end	100 bp single end
Submission date	22/07/2013	22/07/2013	22/07/2013	12/08/2013	12/08/2013	12/08/2013
Experiment No.	RP E29	RP E29	RP E29	RP E34	RP E34	RP E34

Library name corresponds to the consecutive number used in Siegel lab. The field “Experiment No.” links to a detailed protocol for the generation of each library. RP: ribosome profiling.

## Data analysis

### Pre-processing and mapping of reads

Reads from all libraries were processed and mapped using the parameters reported in Ingolia *et al.* 2012. Adapter sequence (i.e. CTGTAGGCACCATCAAT) was trimmed from the 3'-end of reads using cutadapt (<http://code.google.com/p/cutadapt/>) and reads shorter than 20-nt after trimming were removed. Trimmed reads were mapped to the reference genome using bowtie-2 with default ‘local-sensitive’ mode (Langmead and Salzberg, 2012). Genome, as well as gene annotations, of strain 927 version 4.2 were downloaded from EuPathDB (Aurrecochea *et al.*, 2013) and used as the reference in all analyses.

## **Calculation of abundance, translational efficiency and read 5'-end periodicity**

A read is considered to map to a region when its midpoint falls within the annotated range. Abundance of a region in a library was defined as reads mapped per kilobase per million non-structural RNA reads (rpkm). Non-structural RNA refers to the reads mapped to the genome excluding annotated rRNA and tRNA regions. Translational efficiency of a region was defined as its ratio of abundance in the ribosome profiling library to that of the control RNA library of the same life cycle stage. To investigate the distinct 3-nt periodicity of the ribosome profiling libraries, 5'-ends of the mapped reads were piled up and the piled-up coverage of each position from each gene (using either start or stop codon as the reference point) were summed up in meta-gene analyses. Genes with less than 10 reads within the plotted region were excluded.

### **Definition of uORFs and novel CDSs**

A uORF was defined as any ORF 9-nt that was located within an annotated 5'-UTR and was on the same strand as the annotated ORF. Novel CDSs are defined as any translated ORF  $\geq$  30-nt located at least 20-nt away from any existing annotations. In cases in which multiple overlapping novel ORFs were found on different frames of the same strand, only the longest were retained. An ORF is defined as translated

when  $\geq 70\%$  of its CDS is being covered by the pooled ribosome profiling reads (i.e. four libraries) with  $\geq 2$  reads per nucleotide.

### **Calculation of ribosome release score**

Ribosome release scores (RRSs) of all ORFs, including annotated CDSs, putative CDSs and uORFs, were calculated as described with slight modifications (Guttman et al., 2013). Briefly, a pseudo 3'-UTR region (pUTR3) was first assigned to all ORFs. pUTR3 of an ORF was defined as the region between its stop codon and a downstream start codon in any of the three frames on the same strand. Reads mapped within the CDSs and pUTR3 per kilobase (i.e. CDSrpk and pUTR3rpk) in both ribosome profiling and RNA-sequencing control libraries were calculated for each ORF. The ratio of CDSrpk to pUTR3rpk in both ribosome profiling and RNA-sequencing control libraries (i.e. riboratio and RNAratio) were calculated. RRS of each ORF was then defined as the ratio of riboratio to RNAratio.

### **Identification of novel putative CDSs important for parasite's fitness**

The data of a published genome-wide RNA interference fitness-costs association study was reanalysed (Alsford et al., 2011) with the newly defined CDSs. Briefly, short read data were downloaded from European Nucleotide Archive under accession number ERP000431. Reads were then mapped to the genome using bowtie2 with default 'local-sensitive' mode. The number of reads mapping to the annotated and putative CDSs was counted. To identify those putative CDSs that have

significantly less reads mapped in the RNAi libraries, which potentially contribute to the parasite's fitness, differential read count analyses was performed on the pair-wise library comparisons (listed below) using three softwares, including DESeq1 (Anders and Huber, 2010), DESeq2 (<http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>) and EdgeR (Robinson et al., 2010), with default settings. CDSs that were identified to have significantly less reads in the RNAi libraries by all three algorithms ( $P < 0.05$  in each case) linear fold change  $\geq 5$  were considered to be potentially related to the loss of fitness. The pair-wise library comparisons include BF uninduced versus BF induced for 6 days, BF uninduced versus differentiated cells induced and BF uninduced versus PF induced.

### **Analysis of proteomics data**

All mass spectrometric raw data files from Butter *et al.* (Butter et al., 2013) were processed as described except that the data were searched additionally with the previously un-annotated CDSs. Briefly, MaxQuant 1.4.1.2 including the Andromeda search engine (Cox et al., 2011), (Cox and Mann, 2008) was used for processing raw data and database searching. The experimental design template used was as described before (Butter et al., 2013). The search was performed against a combination of three databases: a *T. brucei* database containing 19,103 annotated protein entries for strains 427 and 927 (version 5.0, Tb427 and Tb927, downloaded from <http://tritrypdb.org>), a list of uORFs and putative CDSs identified in this study and a database containing typical

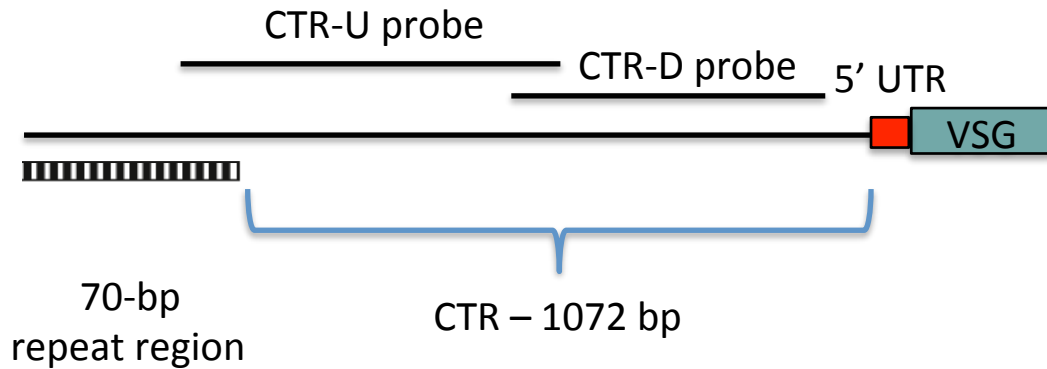
contaminants. Enzyme search specificity was Trypsin/P with three miscleavages for tryptic digests and LysC with two miscleavages for digests with lysyl endopeptidase. Carbamidomethylation on cysteines was set as fixed modification; methionine oxidation and protein N-terminal acetylation were considered as variable modifications. Mass accuracy tolerances (after recalibration) were 6 ppm for precursor ions, and 0.5 Da for CID MS/MS and 20 ppm for HCD MS/MS spectra, respectively. False discovery rate was fixed at 1% at peptide and protein level, with posterior error probability (PEP) as sorting criterion. Identifications were matched between runs within a 2 min window. Protein groups identified in the potential CDS database had at least one unique peptide and a PEP < 0.0005.

## **Characterization of a ncRNA transcribed from the CTR**

### **Northern blot**

#### **Synthesis of *VSG* and CTR probes by PCR**

Primer sequences for the amplification of the *VSG*, CTR-U and CTR-D probes are listed in the section Oligonucleotides, primers and cell lines at the end of this chapter. CTR-U stands for a probe that spans the first 712-nt of the CTR and 86-nt of the 70-bp repeat region immediately upstream. CTR-D stands for a probe that spans 514-nt overlapping 231-nt with CTR-U (Figure 10).



**Figure 10.** Schematic view of the CTR and the corresponding regions from where the CTR-U and CTR-D probes were synthesized.

The probes were synthesized using DreamTaq™ DNA polymerase (Thermo Scientific) according to the manufacturer's instructions and using genomic DNA (gDNA) from Lister 427 BF as a template. The PCR products were run on a 1% agarose gel and the bands corresponding to the desired size were excised and cleaned up with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). Thirty eight  $\mu\text{l}$  of the eluted samples were polyadenylated by using 5  $\mu\text{l}$  of 10x DreamTaq™ buffer, 5  $\mu\text{l}$  of dATP (2mM), 0.2  $\mu\text{l}$  of DreamTaq™ DNA polymerase (5 U/ $\mu\text{l}$ ) in a final volume of 50  $\mu\text{l}$ , and by incubating at 72 °C for 10 min in a thermocycler. Two  $\mu\text{l}$  of the polyadenylated products were cloned in the pCR™2.1-TOPO® vector (Invitrogen) following the manufacturer's instructions. Three  $\mu\text{l}$  of each TOPO cloning reaction were added into a vial of TOP10 chemically competent *E. coli* cells and transformed according to the manufacturer's instructions. Several white colonies were selected for each of the three inserts and the plasmids extracted using the NucleoSpin® Plasmid kit (Macherey-Nagel). The correct sequence for each clone was confirmed by Sanger sequencing.

## Probe radiolabelling using the DECAprime™ II Kit (Ambion)

### Preparation of template DNA

PCR products using the TOPO clones as templates (*VSG*, CTR-U and CTR-D) were diluted in water to a concentration of 2.5 ng/μl. Twenty-five ng of each linearized DNA template were mixed with 2.5 μl of 10x Decamer Solution and completed to a final volume of 14 μl with nuclease-free water. Template DNA was denatured in a heat block at 95 °C for 5 min. The denatured DNA/decamer mixtures were frozen down on dry ice. This step prevented self-annealing of the template DNA, allowing more efficient decamer binding and thus increasing product yield. The reaction mixtures were thawed-up, spun them down and finally placed on ice.

### Probe Synthesis Reaction

The following components were added to the denatured DNA/decamer mixture on ice:

Component	Volume (μl)
Denatured DNA/Decamer Mix	14
5x Reaction Buffer (-dATP)	5
[α- <sup>32</sup> P]dATP 10μCi/μl, 3,000 Ci/mmol	5
Exo-Klenow	1
Final volume	25

The content of the tubes were mixed by pipetting up and down and then incubated at 37 °C for 10 min. The reactions were stopped by adding 1 μl of 0.5 M EDTA. Illustra Microspin G25 columns (GE

healthcare) were used to remove the unincorporated nucleotides. The cutoff of these columns is  $\geq 10$  bases. The samples were run on a 5% denaturing PAGE gel at 180V for 40 min. The gel was removed from the glass plates and put over a filter paper. Then, it was covered in plastic wrap and two more filter papers were put above and below the gel. The gel was dried at 80 °C during 20 min in the Model 583 Gel Dryer (BioRad). The covered gel was exposed to Phosphor screen for ~15 min. The screen was scanned in a Thyphoon fla7000.

## **Isolation of total RNA**

### **Culture and harvest of BF Lister 427**

The wild type Lister427 strain that expresses VSG-221 in BF was grown at 37 °C in HMI-11 medium and harvested when a cell density of around  $1.5 \times 10^6$  cell/ml was reached (~150,000,000 cells). The cells were split in 3 50 ml-Falcon tubes and collected by centrifugation at 3,000 x g, 4 °C for 5 min. The medium was removed and the cells washed once in 1 ml of 1x TDS buffer per Falcon tube. The cells were spun down at 3,000 x g, 4 °C for 5 min. The supernatant was removed by pipetting it out of the tube.



## **RNA purification**

RNA purification was performed using the NucleoSpin RNA II kit (Macherey-Nagel) following the manufacturer's instructions. One column was used for every  $1 \times 10^8$  cells and the total RNA was recovered by eluting in 60  $\mu$ l of RNase-free water. The RNA concentration of each sample was quantified using a Nanodrop 2000 and immediately stored at  $-80^\circ\text{C}$ .

## **Preparation of agarose gels for RNA**

This protocol was supplied by Dr. Ana Eulalio and adapted when necessary. First, 1.4 g of agarose was dissolved in 90 ml of DEPC-treated water. Then, 10 ml 10x MOPS running buffer were added once the agarose cooled down to  $\sim 60^\circ\text{C}$ . Once the gel was solidified, it was installed in the chamber in the cold room. The chamber was filled up with 1x MOPS running buffer (10x MOPS: 0.2 M MOPS [pH 7], 20 mM sodium acetate and 10 mM EDTA [pH 8]).

## **Sample preparation**

Five and 30  $\mu$ l of RNA were used for the detection of the *VSG* and the CTR transcripts respectively. For each 6.75  $\mu$ l of RNA sample, 23.25  $\mu$ l of glyoxal reaction mixture were added. If the required amount of RNA was in a volume higher than 6.75  $\mu$ l then the volume of the glyoxal mixture was increased proportionally. The RNA samples were denatured for 10 min at  $74^\circ\text{C}$ , immediately

chilled on ice and spun down. Three  $\mu\text{l}$  of RNA loading buffer and 1  $\mu\text{l}$  of ethidium bromide were added to each sample.

Glyoxal reaction mixture:

<b>Reagent</b>	<b>Volume (<math>\mu\text{l}</math>)</b>
DMSO	60
Gyloxal	20
10x MOPS	12
Glycerol 80%	6
Water	2
Final volume	100

### **Electrophoresis**

The gel was pre-run for 10 min before loading the samples (70-90 V). Then, the samples were loaded into the wells and the gel was run for 5 hours (70-90 V). Once the electrophoresis was stopped, the gel was washed with DEPC-treated water for 5 min and an image of the gel was obtained to check the quality of the RNA. The gel was then washed again with DEPC-treated water for 5 min before the transfer.

### **RNA transfer to a nylon membrane**

The gel was equilibrated in 10x SSC for 30 min. One of the corners of the gel was cut out in order to align it with the top of the membrane during transfer and simplify orientation during the procedure. A piece of Whatman 3MM was put on a glass plate to form a bridge. It is important to make sure that the ends of the blotting paper drape over the edges of the plate. Two additional Whatman 3MM papers were put over the bridge

~ 1 mm larger than the gel in both dimensions. The filter papers were wet with 10x SSC buffer. The tray was filled with 10x SSC until the level of the liquid reached almost to the top of the support. When the blotting paper on the top of the support was thoroughly wet, all air bubbles were smoothed out with a glass rod. The nylon membrane (Amersham Hybond-XL) was ~ 1 mm larger than the gel in both dimensions. The membrane was always handled with gloves and forceps. The nylon membrane was floated on the surface of a dish of DEPC-treated water until was wet completely from beneath, and then it was immersed in 10x SSC for at least 5 min. A clean scalpel blade was used to cut out a corner from the membrane to match the corner cut out from the gel. The gel was carefully placed on the support in an inverted position so that was centered on the wet blotting paper. It was important to make sure that there were no air bubbles between the thick blotting paper and the gel. The top of the gel was wet with 10x SSC buffer. The wet nylon membrane was placed on top of the gel so that the cut corners are aligned. There were no air bubbles between the membrane and the gel. Two pieces of Whatmann 3MM papers (cut to exactly the same size of the gel) were wet in 10x SSC buffer and placed on top of the nylon membrane. All air bubbles were smoothed out with a glass rod or pipette. The gel was surrounded but not covered with Parafilm to prevent liquid from flowing directly from the reservoir to the paper towels placed on the top of the gel. A stack of paper towels (5-8 cm high) just smaller than the blotting papers was cut and placed on the blotting papers. A plastic tray was put on the top of the stack and weighed down with a 400 g weight. The upward transfer of RNA was allowed to

occur over night. The capillary transfer system was then dismantled. The positions of the slots on the membrane were marked with a pencil through the gel. The membrane was transferred to a glass tray containing 6x SSC at 23 °C. The tray was placed in a platform shaker and agitated very slowly for 5 min. The membrane was removed from the 6x SSC buffer and the excess fluid was allowed to drain away. The membrane was laid, RNA side upward, on a dry sheet of blotting paper for a few min. The gel was stained in ethidium bromide during one hour to check the efficiency of transfer. To cross-link the RNA, the membrane was placed on a piece of dry blotting paper and irradiated at 254 nm using the program 0.120 J (5 min).

### **Northern hybridization**

The membrane was cut in three fragments, corresponding to each of the probes (*VSG*, CTR-U and CTR-D). Each membrane was incubated for 2 hours at 65 °C in 50 ml of pre-hybridization solution (Roti®-Hybriquick) by using glass cylinders in an incubator with rotating wheels. Ten µl of the probes were resuspended in 100 µl of RNase-free water. The DNA double-stranded probes were denatured by heating for 10 min at 100 °C. The probes were immediately chilled on ice for two min. The denatured radiolabeled probes were added directly to the pre-hybridization solution. The incubation was continued for 12-16 hours at 65 °C. After hybridization, the membranes were transferred as quickly as possible to another cylinder containing 50 ml of 0.1x SSC, 0.1% SDS. The membranes were washed for 10 min at 60 °C. The washing step was

repeated two more times. The membranes were dried on blotting paper and wrapped in plastic foil. The membranes were introduced in a cassette with a phosphor imager screen for 48 hours. The screen was scanned in a Thyphoon fla7000.

### **Characterization of the CTR ends by circular RACE PCR**

The objective of this experiment was to detect the 5'- and 3'-ends of the ncRNA transcribed from the CTR in BF of the Lister 427 strain.

### **Isolation of total RNA**

The RNA was isolated using the NucleoSpin RNA II kit (Macherey-Nagel) following the manufacturer's instructions. The absence of gDNA in the samples was important for the downstream steps. For this reason an additional step of rDNase digestion was performed. Six  $\mu\text{l}$  of Reaction Buffer for rDNase and 0.6  $\mu\text{l}$  of rDNase were added to 60  $\mu\text{l}$  eluted RNA. The tube was gently swirled in order to mix the solution and spun down to collect the solution at the bottom of the tube. The sample was incubate at 37 °C for 10 min. The RNA was precipitated by adding 0.1 volume of 3 M sodium acetate (pH 5.5) and 2.5 volumes of absolute ethanol to one volume of sample. After mixing thoroughly, the sample was incubated at -20 °C for 30 min. The tube was centrifuged at 16,000 x g, 4 °C for 30 min and the supernatant discarded. The RNA was washed with 500  $\mu\text{l}$  of 70% ethanol. The tube was centrifuged at 16,000 x g, 4 °C for 15 min and the supernatant discarded. The RNA pellet was dried at room temperature for 10 min and then resuspended in 60  $\mu\text{l}$  of RNase-free water. A PCR using

DreamTaq™ DNA polymerase (Thermo Scientific) was performed using the isolated RNA as a template to check for the presence of gDNA.

### **RACE using circularized RNA**

This protocol was adapted from: McGrath, PT. 2012. "Characterizing cDNA Ends by Circular RACE". *Methods in Molecular Biology*, vol. 772 (Wisotzkey and Newfeld, 2012). The total RNA used was free of genomic DNA.

### **Decapping the RNA**

To perform the Tobacco Acid Pyrophosphatase reaction (+TAP), 5 µg of total RNA were combined in an RNase-free microfuge tube with 2 µl of 10x TAP Reaction buffer, 0.5 µl of RNaseOUT Recombinant RNase Inhibitor (Invitrogen), 10 U of TAP (10 U/µl) (Epicentre) and enough nuclease-free water to bring the total reaction volume to 20 µl. Spurious PCR products in downstream steps can result from degraded or incompletely processed transcripts, which in any case are not capped. Therefore, as a control the same reagents were combined with 5 µg of total RNA but excluding TAP from the mixture. Both tubes were incubated at 37 °C for 2 hours.

In addition, total RNA was treated with RNA 5'-Polyphosphatase instead of TAP. RNA 5'-Polyphosphatase is an enzyme that sequentially removes the  $\gamma$  and  $\beta$  phosphates from 5'-triphosphorylated RNA (primary RNA transcripts) and 5'-

diphosphorylated RNA. RNA 5'-Polyphosphatase has no activity on RNA with a 5'-cap (e.g., 5'- m7GpppN-----OH 3') or a 5'-monophosphorylated (5'-pN-----OH 3') end. Therefore, 5 µg of total RNA were added in one tube together with 2 µl of 10x RNA 5'-Polyphosphatase Reaction Buffer, 0.5 µl of RNaseOUT Recombinant RNase Inhibitor and 1 µl of RNA 5'-Polyphosphatase (20U/µl) (Epicentre). Nuclease-free water was added to bring the total reaction volume to 20 µl. Another reaction not including the enzyme was prepared. Both tubes were incubated at 37 °C for 30 min.

### **Circularizing the RNA**

The decapped and Polyphosphatase-treated RNA were used as substrates in the next reaction. Eleven point five µl of enzymatically-treated or not treated RNA were combined with 2 µl of 10x Reaction buffer, 1 µl ATP (1 mM), 4 µl PEG 8000 (50%), 0.5 µl RNaseOUT Recombinant RNase Inhibitor and 1 µl of T4 RNA ligase (NEB 10 U/µl) to a final volume of 20 µl. The four reactions were incubated in a thermocycler at 16 °C overnight. The next day, the T4 RNA ligase was inactivated by incubating the reaction at 65 °C for 15 min.

### **Reverse transcription of the circularized RNA into cDNA**

A gene specific primer (420\_CTR\_R3) was used in this step (See the section Oligonucleotides, primers and cell lines at the end

of this chapter). In two nuclease-free PCR tubes, 10  $\mu$ l of circularized RNA from the enzymatically-treated or not treated reactions were combined with 2  $\mu$ l of 420\_CTR\_R3 primer (1  $\mu$ M) and 1  $\mu$ l of dNTP mix (10 mM). The mixture was heated to 65  $^{\circ}$ C for 5 min and then incubated on ice for at least 1 min. Then, 4  $\mu$ l of 5x First-Strand Buffer, 1  $\mu$ l DTT (0.1 M), 1  $\mu$ l RNaseOUT Recombinant RNase Inhibitor, and 1  $\mu$ l of SuperScript III RT (200 U/ $\mu$ l) (Invitrogen) were added to each sample. Then, the tubes were incubated at 50  $^{\circ}$ C for 45 min, followed by another incubation at 55  $^{\circ}$ C for 45 min. The reaction was inactivated by heating at 70  $^{\circ}$ C for 15 min. To remove the complementary RNA, 2 units of RNase H (Ambion) were added to each tube by incubating at 37  $^{\circ}$ C for 20 min.

### PCR reaction using the Kapa HiFi Ready Mix

The following components were added to a 0.2 ml micro-centrifuge tube on ice:

Reagent	Volume ( $\mu$ l)
Sterile water	8.4
Kapa HiFi Ready Mix	10.0
Forward primer (10 $\mu$ M)*	0.6
Reversed primer (10 $\mu$ M)*	0.6
cDNA	0.4
Final volume	20.0

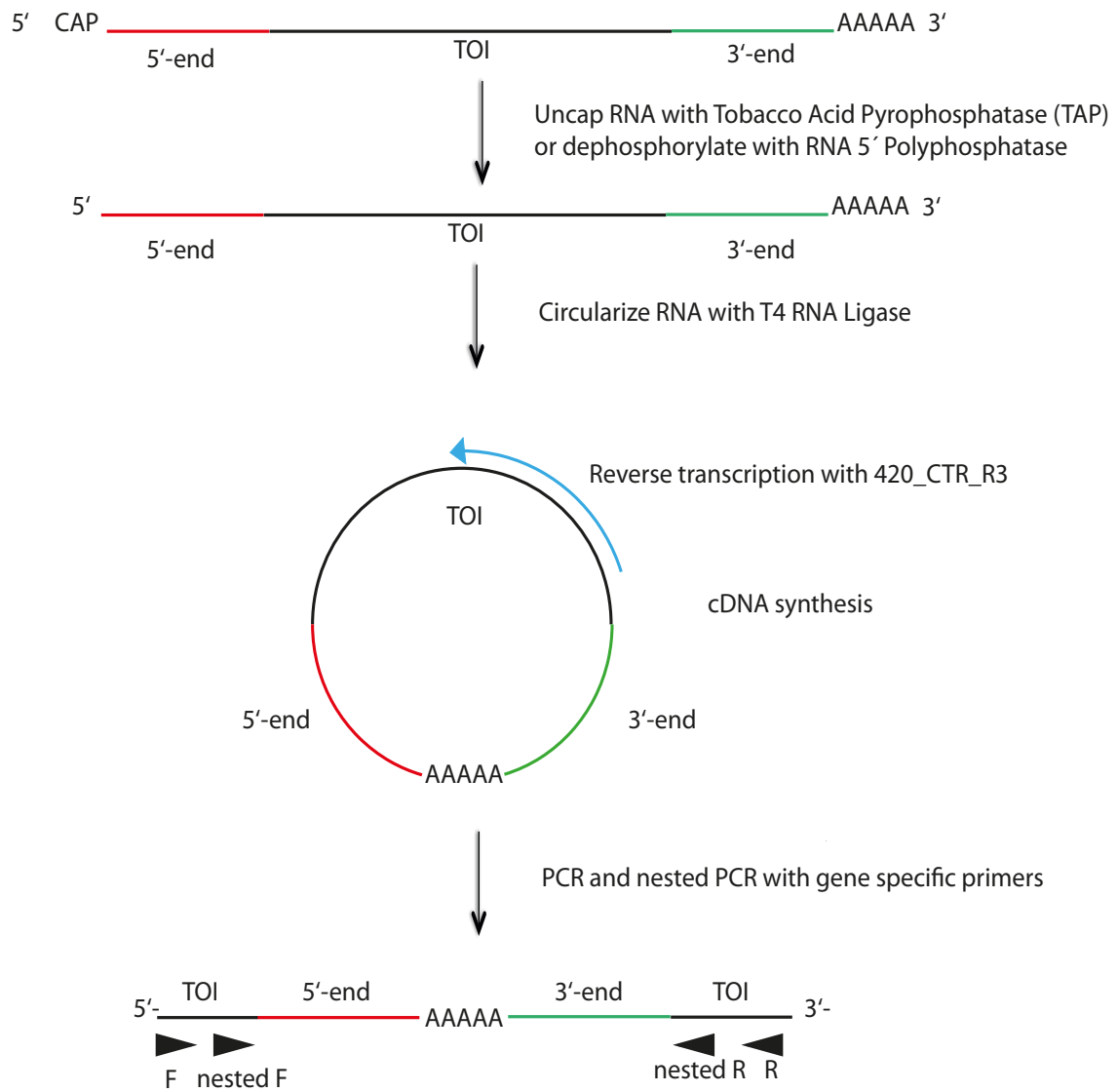
(\*) Primer pairs used to amplify the CTR cDNA (See the section Oligonucleotides primers and cell lines at the end of this chapter)



Primers 1009\_CTRFD and 1011\_CTRRD only amplify in case of obtaining cDNA from circularized RNA (See the section Oligonucleotides, primers and cell lines at the end of this chapter). The amplicons obtained with this pair of primers should contain the 5'- and 3'-ends of the CTR transcript. Primers 617\_3race\_F2 and 718\_CTR\_R3\_nested only amplify cDNA obtained from the original linear RNA (See the section Oligonucleotides, primers and cell lines at the end of this chapter). The amplicons obtained with this pair of primers serve as a control for the reverse transcription reaction and the presence of the CTR transcript but do not contain the 5'- and 3'-ends of the CTR transcript (Figure 11).

### **Nested PCR using the Kapa HiFi Ready Mix**

Since the CTR transcript is found at low levels in wild type cells, it was necessary to perform a nested PCR reaction to improve the specificity of the PCR products. To perform a nested PCR, the Kapa HiFi Ready Mix system was used as showed immediately above (1010\_CTRFDn and 1012\_CTRRDn to amplify the circularized template and 618\_3race\_F2\_n and 718\_CTR\_R3\_n to amplify the linear template). The band present in the enzymatically-treated samples and absent in the non-enzymatically treated was excised from the gel and sent for Sanger sequencing.



**Figure 11.** Schematic illustration of a circular RACE PCR. TOI stands for Transcript Of Interest. The blue curved arrow indicates the direction of reverse transcription. Black arrowheads correspond to primers used to amplify the 5'- and 3'-ends of the TOI. Figure adapted from McGrath (Wisotzkey and Newfeld, 2012).

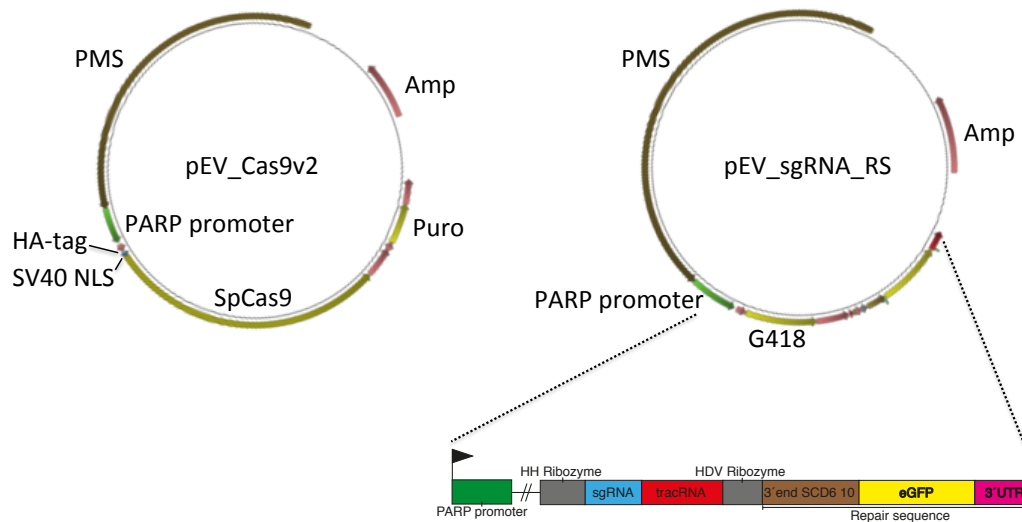
## CRISPR-Cas9 system

### Tagging the 3'-end of the SCD6 CDS with the eGFP CDS

The generation of episome pEV\_Luc is described in Patnaik *et al.* (Patnaik et al., 1994). In turn, the pEV\_Luc episome is a modified version of the original episome pT13-11 described in 1993 (Patnaik et al., 1993).

In order to generate the episome pEV\_Cas9v2\_puro, the G418 resistant cassette was removed from pEV\_Luc and replaced by the puromycin resistant cassette; keeping the SAS, PAS, 5'- and 3'-UTRs of the PARP gene. Additionally, the luciferase CDS and a vector sequence upstream of the start codon of luciferase were removed for the introduction of the Cas9v2 insert (left panel).

The episome pEV\_sgRNA\_SCD6\_GFP was generated by removing the luciferase CDS and a vector sequence upstream of the luciferase start codon. The G418 resistant cassette was kept and downstream was inserted the sgRNA flanked by the hammerhead (HHR) and the hepatitis delta virus (HDVR) ribozymes and the SCD6\_GFP repair sequence. This construct was inserted as a gBlock ordered from IDT (Figure 12, right panel).



**Figure 12.** Episomes used for tagging the SCD6 CDS with the eGFP sequence using the CRISPR-Cas9 system. (Left panel) The Cas9 endonuclease bearing the SV40 nuclear localization signal (NLS) and a HA-tag was expressed in the episome pEV\_Cas9v2, which contains a puromycin resistant cassette. (Right panel) The episome pEV\_sgRNA\_RS contained the sgRNA-expression cassette flanked by the HH and HDV ribozymes and the repair sequence (RS) for tagging the SCD6 CDS with GFP. pEV\_sgRNA\_RS contains a G418 resistant cassette. The transcription is driven in both episomes from the *T. brucei* PARP polymerase I promoter.

The PCR reactions for amplification of inserts were performed using Phusion High-Fidelity DNA Polymerase as described by the manufacturer (Thermo Scientific). The reactions were run in agarose gels and the corresponding bands were excised from the gel and cleaned up using the Nucleospin gel and PCR clean up kit (Macherey-Nagel). The concentration was measured using a Nanodrop 2000. The primers used for the PCR reactions are listed in the section Oligonucleotides, primers and cell lines at the end of this chapter.

The In-Fusion® HD Cloning Kit (Clontech) was used for cloning the PCR products in the digested episome following the manufacturer instructions. The reactions were incubated at 50 °C for 15 min and Stellar™ Competent Cells were immediately transformed. Five colonies were picked for each event of transformation, added to liquid LB medium plus ampicillin (final concentration 50 µg/ml) and incubated at 180 rpm, 37 °C overnight. The plasmids were extracted using the Nucleospin plasmid kit (Macherey-Nagel) and the concentration was measured using a Nanodrop 2000. The correct insertion of the fragments was confirmed digesting the plasmids with the appropriate restriction enzymes and the positive clones sequenced.

### **Removal of the luciferase CDS and a vector sequence upstream of the luciferase start codon**

The first step for the generation of the CRISPR-Cas9 episomes was the removal of the luciferase CDS and a vector sequence upstream of the luciferase start codon. The vector sequence contained an ATG codon 43-nt upstream of the luciferase start codon. It has been reported that the presence of this “uORF” can decrease the translation efficiency of the main coding sequence located downstream (Siegel et al., 2005). For this reason this short vector sequence was deleted. Therefore, a PCR was performed using the primers 951\_Epi\_F1 and 980\_Epi\_novector1 to keep the polylinker region and the SAS and 5'-UTR of the PARP gene that were originally present upstream of the luciferase CDS. A second PCR was performed with the primers 955\_Epi\_F3 and 956\_Epi\_R3 to keep the PAS and 3'-UTR of the actin gene present downstream of the luciferase CDS.

The sequence of the primers used for the PCR reactions are listed in the section Oligonucleotides, primers and cell lines at the end of this chapter. In turn, the episome pEV\_Luc was digested with KpnI and SacI and an In-Fusion reaction was performed for the cloning of the 2 PCR products (951\_Epi\_F1/980\_Epi\_novector1 and 955\_Epi\_F3/956\_Epi\_R3). The resulting plasmid was named pEV\_noLuc.

### **Insertion of the puromycin resistant cassette instead of the G418 resistant cassette for the generation of pEV\_puro**

One PCR reaction was performed using the primers 1086\_Epi\_Puro\_F2 and 1087\_Epi\_Puro\_R2 to amplify the puromycin resistant cassette. Because the puromycin resistant cassette has a high GC content, the KAPA HiFi HotStart ReadyMix was used adding DMSO to a final concentration of 1.25%. pEV\_noLuc was digested with SacI and BamHI and an In-Fusion reaction was performed for cloning the 1086\_Epi\_Puro\_F2/1087\_Epi\_Puro\_R2 PCR product. The resulting plasmid was named pEV\_puro.

The next step involved the cloning of the first version of SpCas9 in the region where the luciferase CDS was present in pEV. This version of the SpCas9 contains the nuclear localization signal (NLS) of the histone 2B (H2B) immediately upstream of the Cas9 start codon.

### **Generation of pEV\_Cas9v1\_puro**

One PCR reaction was performed to keep the SAS and 5'-UTR of

the PARP gene using the primers 1162\_Epi\_Puro\_F1 and 1085\_Epi\_Puro\_R1, a second one to amplify the H2B NLS and the SpCas9 CDS using the primers 1180\_pEV\_Cas9\_puro\_F1 and 1181\_pEV\_Cas9\_puro\_R1 and a third one to amplify part of the actin PAS using the primers 1182\_pEV\_cas9\_puro\_F2 and 956\_Epi\_R3. pEV\_puro was digested with KpnI and SacI and an In-Fusion reaction was performed including the 3 PCR reactions (1162\_Epi\_Puro\_F1/1085\_Epi\_Puro\_R1, 1180\_pEV\_Cas9\_puro\_F1/1181\_pEV\_Cas9\_puro\_R1 and 1182\_pEV\_cas9\_puro\_F2/956\_Epi\_R3) and the digested backbone. The resulting plasmid was called pEV\_Cas9v1\_puro. Nevertheless after performing an IF, the nuclear localization of this Cas9 version was low. For this reason a second version of Cas9 was included in pEV\_puro.

### **Generation of pEV\_Cas9v2\_puro**

This second version of Cas9 contains a HA-Tag, the NLS found in pLEW13 that is a shorter version of the SV40 NLS and the SpCas9 CDS (see the left panel in Figure 12). pEV\_Cas9v1\_puro and pCW20 (that contains the Cas9v2 construct and was developed by MSc. Carolin Wedel) were digested with SbfI and AflIII and ligated using the T4 DNA Ligase system (Thermo Scientific). The construct was named pEV\_Cas9v2\_puro.

### **Generation of pEV\_SCD6\_GFP\_gBlock**

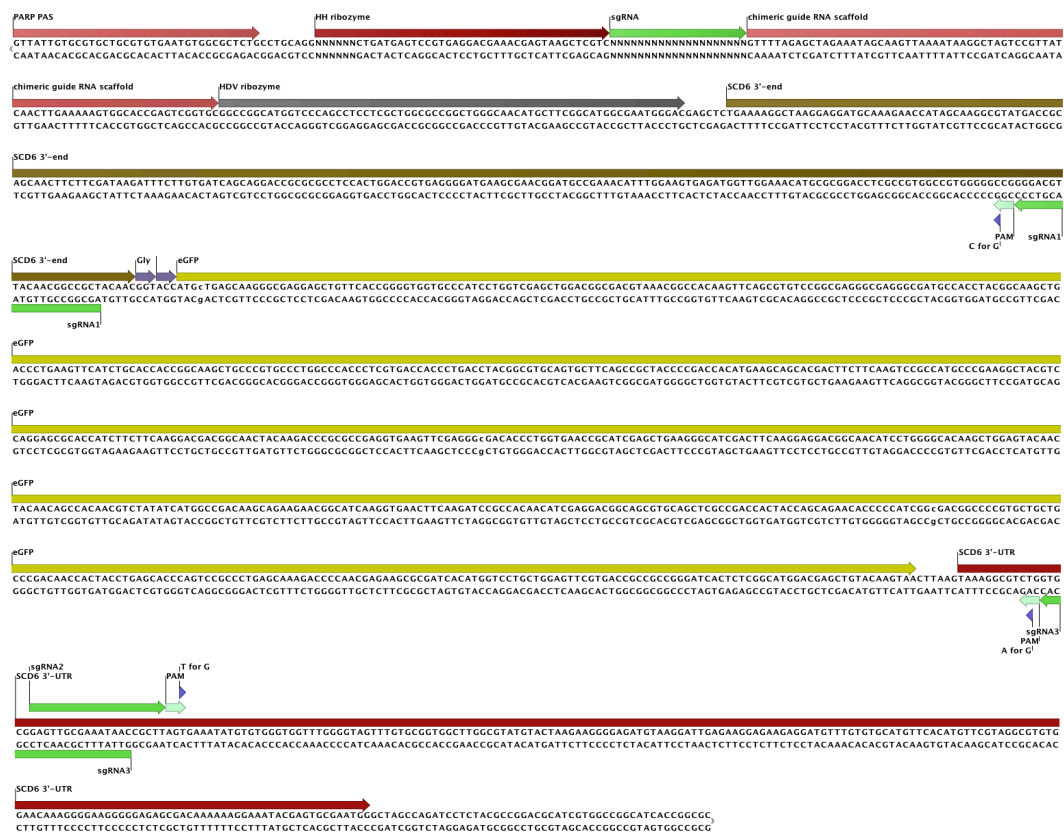
A SbfI restriction site was introduced in pEV\_noLuc by primer site-directed mutagenesis in order to insert the repair sequence. Two PCR

reactions were performed: one upstream and other downstream of the new SbfI restriction site. The primers 1170\_SDM1 and 1171\_SDM2 were used for the first reaction and the primers 1172\_SDM3 and 1173\_SDM4 for the second one. After running a 1% agarose gel, the bands were excised from the gel and cleaned-up. Then, a PCR reaction was performed using the outer primers 1170\_SDM1 and 1173\_SDM4 and the products from reactions 1 (1170\_SDM1/1171\_SDM2) and 2 (1172\_SDM3/1173\_SDM4) as templates. The resultant PCR product and the plasmid pEV\_noLuc were cleaned-up and digested with BamHI and NheI. A ligation reaction was performed using a 5:1 molar ratio of the digested PCR product and the backbone respectively. The ligation was performed by using 5 U/ $\mu$ l of T4 DNA Ligase (Thermo Scientific) following the manufacturer instructions. Once it was confirmed that one of the resulting clones included the new SbfI restriction site, the plasmid (pEV\_noLuc\_SbfI) was digested with SbfI and NheI and an In-Fusion reaction was performed for cloning the repair sequencing containing the 3'-end of the SCD6 CDS and the eGFP CDS (SCD6\_GFP). This repair sequence (1,466-bp) was cloned using two gBlocks named gBlocks 12 and gBlocks 13. The resulting construct was named pEV\_SCD6\_GFP\_gBlock. Digesting with SbfI and NheI releases the original repair sequence and a new repair sequence for other editing purposes can be cloned into this episome.



## Insertion of the HH ribozyme, sgRNAs protospacers, the chimeric guide RNA scaffold and the HDV ribozyme sequences in pEV\_SCD6\_gBlock

Three sgRNA protospacers were evaluated for the tagging of eGFP CDS at the 3'-end of the SCD6 CDS. For this reason three gBlocks were ordered from IDT. The protospacers were designed using the <http://spot.colorado.edu/~slin/cas9.html> resource (Figure 13).



**Figure 13.** Repair sequence for the insertion of the eGFP sequence at the 3'-end of the SCD6 CDS and the three sgRNAs evaluated. The first 6-nt in the sense strand of the HH ribozyme sequence correspond to the reverse complement sequence for the first 6-nt in the sense strand of the protospacer. The targeted sequences for the three sgRNAs evaluated are located either in the SCD6 3'-end (sgRNA1) or in the SCD6 3'-UTR (sgRNA2 and sgRNA3). The PAM sequence (5'-NGG) present in the target is depicted in pale green. Nevertheless, one G in each PAM sequence was substituted in order to avoid Cas9-induced DSBs in the repair sequence.

pEV\_SCD6\_GFP\_gBlock was digested with SbfI and SacI and three separate In-Fusion reactions were performed including the backbone and the gBlocks. The resulting plasmids were named pEV\_sgRNA1\_SCD6\_GFP, pEV\_sgRNA2\_SCD6\_GFP and pEV\_sgRNA3\_SCD6\_GFP.

### **Sequential transfection of wild type PF cells with pEV\_Cas9v2\_puro and pEV\_sgRNA1\_SCD6\_GFP, pEV\_sgRNA2\_SCD6\_GFP or pEV\_sgRNA3\_SCD6\_GFP**

SDM-79 medium containing 20% fetal calf serum (FCS) was used for the transfection of PF. Twenty million wild type PF were harvested by centrifugation at 1,400 x g, room temperature for 10 min. The supernatant was carefully discarded by pipetting it out, keeping around 1 ml of medium inside the tube. The residual medium was transferred to a 1.5 ml microcentrifuge tube and centrifuged at 1,400 x g, room temperature for 10 min. Then the medium was completely removed and the pellet was resuspended in 400 µl of cold transfection buffer. Ten µg of pEV\_Cas9v2\_puro were added to a cold cuvette (VWR 2 mm, 400 µl) and afterwards the cells resuspended in buffer. In another cuvette, no DNA was added (mock treatment). The cells were transfected using the protocol X-014 (Nucleofector™2b). The cells were immediately transferred to 10 ml of SDM79 medium enriched with 20% FCS and then split in two flasks each one with 5 ml of medium. The cells were incubated at 27 °C. The next day, puromycin was added to the flasks at a final concentration of 1 µg/ml. PF transfected with pEV\_Cas9v2\_puro survived the drug selection for at least three weeks. Then the cells were transfected with the episomes pEV\_sgRNA1\_SCD6\_GFP, pEV\_sgRNA2\_SCD6\_GFP and pEV\_sgRNA3\_SCD6\_GFP following the same procedure described above. The next day G418 was added to

a final concentration of 15 µg/ml and the cells were kept under drug pressure for 10 days. Then the cultures were serially diluted to 1:1 x 10<sup>1</sup>, 1:1 x 10<sup>2</sup>, 1:1 x 10<sup>3</sup>, 1:1 x 10<sup>4</sup>, 1:1 x 10<sup>5</sup> and 1:1 x 10<sup>6</sup> in 96-well plates adding 200 µl of culture per well. The medium contained 20% FCS, G418 and puromycin. The cells were incubated at 27 °C until a yellowish coloration was observed in the wells. In order to obtain clonal populations, cells from the more diluted plates were transferred to 24-well plates adding 300 µl of SDM-79 (20% FCS) and drugs. The following days half volume of SDM-79 (20% FCS) were added to the wells until the culture was transferred to T25 flasks. The cells were grown continually under drug pressure.

#### **Co-transfection of wild type PF cells with pEV\_Cas9v2\_puro and pEV\_sgRNA1\_SCD6\_10\_GFP**

The same protocol described above was followed with the exception that 10 µg of both episomes were added in only one event of transfection. The next day both drugs were added and the drug pressure was removed after five days.

#### **Generation of a SCD6-GFP fusion protein using the pMOTag3G and 4G tagging system**

The pMOTag3G and 4G series (Oberholzer et al., 2006) allow the introduction of a gene coding for GFP after the 3'-end of a CDS, in this case SCD6. The endogenous 3'-UTR of SCD6 is replaced by the α- and β-tubulin intergenic region followed by the G418 (pMOTag3G) or the hygromycin (pMOTag4G) resistant cassette and the SCD6 3'-UTR.

One PCR was performed using the primers 1337\_pMOT\_SCD\_F1 and 1338\_pMOT\_SCD\_R1 to amplify the 3'-end of the SCD6 CDS and another PCR using the primers 1339\_pMOT\_SCD\_F2 and 1340\_pMOT\_SCD\_R2 to amplify the 3'-UTR of the SCD6 gene. pMOTag3G and pMOTag4G and the PCR product 1337\_pMOT\_SCD\_F1/1338\_pMOT\_SCD\_R1 were digested with ApaI and XhoI for cloning the 3'-end of SCD6. The digestion products were ligated using the T4 DNA ligase system (Thermo Scientific). Then, the constructs pMOTag3G\_SCD\_3end, pMOTag4G\_SCD\_3end and the PCR reaction 1339\_pMOT\_SCD\_F2/1340\_pMOT\_SCD\_R2 for the cloning of the 3'-UTR of the SCD6 gene were digested with BamHI and SacI before ligation. The positive clones were named pMOTag3G\_SCD6 and pMOTag4G\_SCD6. Both constructs were digested with ApaI and XhoI and the resulting fragment was precipitated by the addition of 1/10 sodium acetate 3M (pH 5.5) and 2.5 volumes of absolute ethanol at -20 °C overnight. The next day, the DNA was precipitated at 16,000 x g, 4 °C for 30 min. The supernatant was discarded and the DNA washed with 500 µl of 70% ethanol at 16,000 x g, 4 °C for 20 min. The pellet was air-dried for 10 min and resuspended in 20 µl of sterile water.

Wild type PF were stably transfected following the protocol described above and adding G418 or hygromycin respectively.

### **Evaluation of the nuclear localization of Cas9v1 and Cas9v2 by immunofluorescence (IF)**

Between 10-20 million cells were spun down at 700 x g, room temperature for 10 min. The supernatant was discarded except 1 ml, and the

cells were transferred to a 1.5 ml microcentrifuge tube. To fix the cells, 54  $\mu$ l of 37% formaldehyde were added to the samples and incubated at room temperature for 5 min. The cells were centrifuged at 4,000 x g, room temperature for 5 min. The supernatant was discarded and the cells resuspended in 1 ml 1x PBS by vortexing. The washing step was repeated 3 times. After the last washing step the cells were resuspended in 100  $\mu$ l of 1x PBS. Coverslips were placed on Parafilm and the cells were pipetted onto the coverslips. The cells were settled down by gravity for 30 min. The coverslips were washed 5 times with 1x PBS and then the cells were permeabilized for 5 min in 100  $\mu$ l of 0.2% NP40 in PBS. The cells were washed 3 times with 1x PBS. Then the coverslips were blocked 2 times in 100  $\mu$ l PBG (0,2% cold water fish gelatin, 0,5% BSA) for 10 min. The primary antibodies used were Mouse mAb for CRISPR-Cas9 (Abcam, 191468, 1 mg/ ml) diluted 1:2000 in PBG and Rabbit  $\alpha$ H2A.Z #3 (0.256 mg/ ml) diluted 1:2000 in PBG. Fifty  $\mu$ l of the antibody dilution were pipetted on Parafilm and the coverslips were put upside down on the drop. The cells were incubated in the dark for 1.5 hours. The coverslips were turned side up and washed 5 times with PBG during 5 min. The secondary antibodies used were Anti-mouse Alexa Fluor-488 donkey (2 mg/ml) (Thermo Scientific) diluted 1:500 and Anti-rabbit Alexa 594 IgG chicken (2 mg/ml) (Thermo Scientific) diluted 1:500 in PBG. One hundred  $\mu$ l were used per coverslip, which were incubated in the dark for 1.5 hours. The coverslips were washed 5 times with 1x PBS for 5 min. Five  $\mu$ l of Vectashield 4,6-diamidino-2-phenylindole (DAPI) were pipetted onto the slides. The coverslips were rinsed briefly in water and placed (cells downwards) onto the microscope slides. The

slides were sealed with nail polish and incubated at room temperature in the dark. For long term, the slides were stored at 4 °C.

### **Evaluation of genome editing and allelic efficiency by PCR**

The Phusion Human Specimen Direct PCR kit (Finnzymes) was used for the extraction of gDNA, following the instructions provided by the manufacturer. The gDNA was extracted from  $1 \times 10^6$  cells and was resuspended in 20  $\mu$ l of the dilution buffer included in the kit. The primers used for the amplification of the SCD6\_eGFP sequence are listed in the section Oligonucleotides, primers and cell lines at the end of this chapter. The primer 1327 (Forward) anneals 753-bp upstream of the 5'-end of the repair sequence and the primer 1328 (Reverse) anneals 269-bp downstream of the 3'-end of the repair sequence. Therefore, these primers cannot anneal in the episome sequence, guaranteeing the amplification of the SCD6\_eGFP sequence that has been inserted into the genomic DNA of the parasite. The PCR reactions were performed by adding 2.0  $\mu$ l of 10x Dream Taq Buffer, 0.4  $\mu$ l of dNTPs (10 mM), 0.4  $\mu$ l of forward primer (10  $\mu$ M), 0.4  $\mu$ l of reverse primer (10  $\mu$ M), 0.1  $\mu$ l of Dream Taq™ DNA Polymerase (5 U/ $\mu$ l), 100 ng of gDNA, adjusting the final volume to 20  $\mu$ l with sterile water. The program used include the following steps: 95 °C for 3 min, 95 °C for 30 sec, 62 °C for 30 sec, 72 °C for 2:30 min in 30 cycles, and a final extension step at 72 °C for 10 min. The band for the endogenous allele corresponds to 1,466-bp and for SCD6\_GFP to 2,186-bp.

## **Evaluation of GFP positive cells using flow cytometry (MACSquant analyzer 10)**

One ml of cell culture was transferred to a 15 ml conical tube. For GFP fluorescence, a laser excitation of 488 nm coupled with 530/30 and 505LP sequential filter was used. The photomultiplier voltage was set at 388 V. Fifty thousand events were measured for each sample.

## **Evaluation of GFP positive cells by microscopy**

Transfected and wild type cells were fixed following a similar procedure as for the IF described above with the exception of the antibody incubation and washing steps.

### **Fluorescence microscopy**

Microscopy images were taken on a Leica DMI6000B microscope equipped with a Leica CRT6000 illumination system (Leica) and a color camera Leica DFC630FX. The images were taken with a HCX PL APO oil immersion objective with 100x1.47 magnification or with a HCX PL FLUOTAR dry objective with 40x0.6 magnification. Linear image processing was done using Leica Application Suite Advance Fluorescence Software. The DAPI signal was detected using an excitation filter 359 nm and emission filter 461 nm (excitation filter BP 360/40 and suppression filter BP 470/40). The eGFP fluorescence signal was detected using an excitation filter 489 nm and an emission filter 508 nm (excitation filter BP 480/40 and suppression filter BP 527/30). The Alexa 594 signal was

detected using an excitation filter 558 nm and an emission filter 582 nm (excitation filter BP 546/12 and suppression filter BP 600/40). Excitation times were 46, 182 and 234 msec, respectively. Transmitted light images were taken with 11 msec of excitation time.

### **Confocal microscopy**

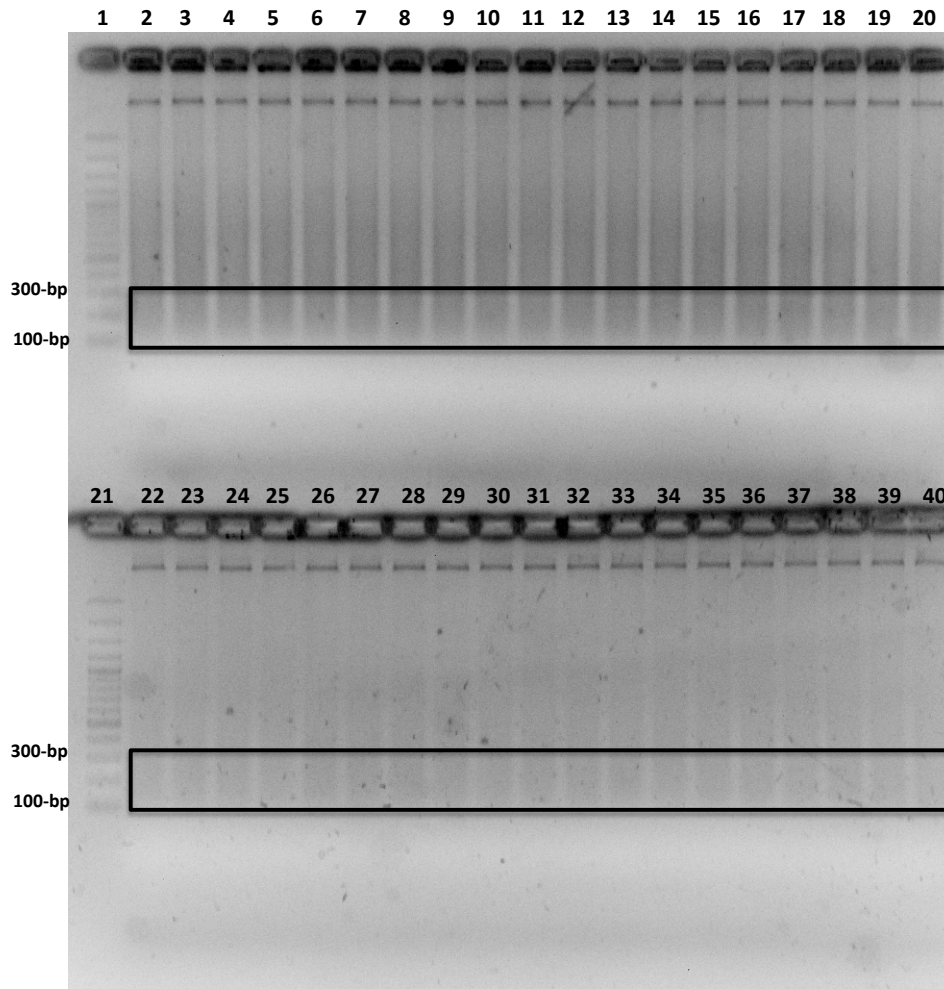
Bright field and fluorescence images were acquired using a TCS SP5 II confocal microscope (Leica). The images were taken with a HCX PL APO oil immersion objective with 100x1.47 magnification or a HCX PL APO oil immersion objective with 40x1.25-0.75 magnification. The hardware settings included: Argon laser power at 25% and 496 nm laser intensity at 15%. Bright field images were collected using the PMT-1 Trans scan channel at 512 V with a gain offset of -0.15%. Fluorescent images were collected using the HyD-2 channel with a gain of 10 and an emission bandwidth of 500 nm for excitation and 550 nm for emission (excitation filter BP 470/40 and suppression filter BP 525/20). The acquisition mode included a xyz scan mode, with z-stacks in the z-wide mode from 4 to 8  $\mu\text{m}$ .

### **Whole genome sequencing**

The DNeasy Blood and Tissue Kit (Qiagen®) was used to prepare the genomic DNA from wild type and sequentially transfected PF cell lines. Low-power sonication in a Bioruptor Plus (Diagenode®) performing 10 min cycles (7x), with 30 sec ON pulses followed by 30 sec OFF; was used to fragment gDNA (6  $\mu\text{g}$ ) in 100  $\mu\text{l}$  of TE buffer (10 mM



Tris-HCl, 1 mM EDTA, pH 8.0). Both sonicated samples were run on a 2% agarose gel and fragments between 100-300 bp were excised (Figure 14) and cleaned up using the Gel and PCR Clean Up Kit (Macherey & Nagel).



**Figure 14.** Separation of sonicated gDNA in an agarose gel for library preparation. Separation of sonicated gDNA from wild type (top) and sequentially transfected PF (bottom) in a 2% agarose gel. Lines 1 and 21: 100-bp plus DNA ladder (Thermo Scientific). Lines 2-20: sonicated gDNA from wild type PF. Lines 22-40: sonicated gDNA from sequentially transfected PF. The black box indicates the region between 100 and 300-bp excised from the gel to prepare the libraries.

The samples were eluted 40 µl of elution buffer and the concentration measured in a Qubit Fluorometer (Life Technologies®) using a Qubit dsDNA HS Assay Kit. The resultant DNA fragments (40 ng) were used to prepare indexed DNA libraries for next-generation sequencing using the NextSeq 500/550 v2 Kit (Illumina®) following the manufacturer's instructions. After purification, fragment sizes and library quality were estimated in a Bioanalyzer 2100 (Agilent®) and the DNA concentration was determined in a Qubit Fluorometer (Life Technologies®). Then, 100-bp reads were sequenced in a NextSeq500 sequencer (Illumina®) in the Core Unit Sys Med at the ZIMF/IMIB University of Würzburg. Refer to the Table 2 for information related about the details of the sequencing process.

**Table 2.** Relevant information about the sequencing process of the two libraries generated to evaluate the on and off-target effects of the CRISPR-Cas9 system.

<b>Sample</b>	<b>PF wild type</b>	<b>PF SCD6_GFP</b>
Library name	NS168	NS170
Library construction protocol	Illumina ChIP-seq	Illumina ChIP-seq
Kit	TruSeq® v1/v2/LT Sample Prep Kits	TruSeq® v1/v2/LT Sample Prep Kits
Amplification (No. Cycles)	9	12
Bar code	TTAGGC	ACAGTG
Index type	3	5
DNA dissolved in	10 mM Tris, 0.1 mM EDTA pH 8.0	10 mM Tris, 0.1 mM EDTA pH 8.0
Concentration (ng/µl)	3.7	5.84
Determined by	Qubit	Qubit
Sample volume (µl)	10	10
Fragment size (range) bp	150-350	150-350
Requested No. of reads	5,000,000	5,000,000
Reads obtained	4,915,282	6,423,321
Sequencing run conditions	2x 75 bp paired end	2x 75 bp paired end
Submission date	21/10/2015	21/10/2015
Experiment No.	CC E18	CC E18

Library name corresponds to the consecutive number used in Siegel lab. The field "Experiment No." links to a detailed protocol for the generation of each library. CC: CRISPR-Cas9.

## Off-target analysis

The VarScan2 software (Koboldt et al., 2012) was used to detect indels and indel/off-target matches occurring in wild type and sequentially transfected PF cell lines after comparing with the *T. brucei* genome version 26 downloaded from TriTrypDB database (retrieved on 23.11.2015). The command mpileup2indel was used to call indels using the following default options: --min-coverage 8 (minimum read depth at a position to make a call), --min-reads2 2 (minimum supporting reads at a position to call variants), --min-avg-qual 15 (minimum base quality at a position to count a read), --min-var-freq 0.01 (minimum variant allele frequency threshold), --min-freq-for-hom 0.75 (minimum frequency to call homozygote), --strand-filter1 (ignore variants with >90% support on one strand), --output-vcf (if set to 1, outputs in VCF format), --variants 0 (report only variant [SNP/indel] positions [mpileup2cns only]). The only option changed was --p-value (p-value threshold for calling variants) from 0.01 to 0.05 as was reported in Ghorbal *et al.* (Ghorbal et al., 2014). 0.05 was the highest p-value that minimized the number of false negatives (maximized sensitivity).

The Protospacer Workbench software suite (MacPherson and Scherf, 2015) was used to find putative off-targets up to 10 nucleotides different from the original seed, 'AGCGGCCGTTGTAACGTCCC' corresponding for the sgRNA1 targeting SCD6. The software scored each off-target according to the algorithm Cas-OFFinder (Bae et al., 2014). Cas-OFFinder is a fast and versatile algorithm that is not limited by the number of mismatches, allows variations in PAM sequences recognized by Cas9 and reports the majority of off-targets.

## **Generation of a HSVTK-negative selection system in cells containing the episome pEV\_Cas9v2\_puro**

The objective of this experiment was to determine the stability of the episome once the drug pressure was released. The pEV\_Cas9v2\_puro episome that already contains a positive selection marker (puromycin resistant cassette) was modified introducing a negative selection marker (Herpes simplex virus thymidine kinase: HSVTK) (Scahill et al., 2008). Wild type parasites are insensitive to the nucleoside analog ganciclovir, which is not the case for parasites expressing HSVTK. The idea with this approach was to initially kill the cells containing the episome and after puromycin addition to kill the remaining cells sensitive to the drug. One PCR was performed to amplify the PAS from actin and the SAS of PARP present upstream of the puromycin resistant cassette using primers 1459\_TK\_puro\_F1 and 1460\_TK\_puro\_R1. A second PCR was performed to amplify the puromycin-HSVTK cassette using the primers 1086\_Epi\_Puro\_F2 and 1461\_TK\_puro\_R2. The KAPA HiFi HotStart ReadyMix was used for the amplification of the puromycin-HSVTK cassette. pEV\_Cas9v2\_puro was digested with BamHI and the 2 PCR reactions (1459\_TK\_puro\_F1/1460\_TK\_puro\_R1 and 1086\_Epi\_Puro\_F2/1461\_TK\_puro\_R2) were cloned in the vector performing one In-Fusion reaction. Once one clone was evaluated for the correct insertion of the cassette (by digestion and sequencing), wild type PF cells were transfected following the protocol described above.

## **Evaluation of episome integration in the genome of the parasite by Southern blot**

The Southern blot protocol described here is based in the Hybond-XL instruction manual (Amersham Biosciences) and Sambrook and Russell (Sambrook, 2001).

### **Synthesis of probes by PCR**

Probes were designed from the SpCas9 CDS (nucleotides 1,136 to 1,715) and the lactamase gene (nucleotides 72 to 612).  $\beta$ -tubulin was used as a positive control for genomic DNA (nucleotides 967 to 1,088). The probes were synthesized by performing 3 PCRs using the primers 1529\_probe\_Cas9\_F and 1530\_probe\_Cas9\_R for the Cas9 probe, the primers 1531\_probe\_AmpR\_F and 1532\_probe\_AmpR\_R for the lactamase probe and the primers 376\_Tub\_F\_qPCR and 267\_  $\beta$ \_tubulin\_R for the  $\beta$ -tubulin probe.

PCR reactions were performed using the Phusion High-Fidelity DNA Polymerase as described by the manufacturer (Thermo Scientific). The PCR products were run on a 1% agarose gel and the corresponding bands excised from the gel and cleaned-up. The probes were labeled using the DECAprime II Procedure (DECAprime™ II Kit Ambion) as mentioned in the section Characterization of the CTR transcript.

## **Genomic DNA extraction and digestion**

Genomic DNA was extracted using DNAzol® reagent (Thermo Scientific). The material was obtained from PF cells sequentially transfected with pEV\_Cas9v2\_puro and pEV\_SCD6\_GFP\_sgRNA1 that were selected for 10 days and the drug pressure removed for one month before collection.

Thirty µg of each gDNA were digested with 200 U of KpnI. The reactions were digested 37 °C, while shaking at 80 rpm, overnight. The next day 100 U of KpnI were added and the reaction was incubated for 4 additional hours to complete the digestion. The DNA was precipitated by adding 1/10 sodium acetate 3M (pH 5.5) and 2.5 volumes of absolute ethanol and then by incubating on ice for 1 hour. The DNA was centrifuged at 16,900 x g, 4 °C for 30 min. The supernatant was discarded and the DNA washed with 1 ml of 70% ethanol. The samples were again centrifuged at 16,900 x g, 4 °C for 15 min and then dried for 10 min under the hood. Finally the DNA was resuspended in 50 µl of water. Thirty µg of pEV\_Cas9v2\_puro were digested with 40 units of KpnI and afterwards cleaned-up. The digestion of the episome had two purposes: as a hybridization control for Cas9 and lactamase probes and to load different concentrations in order to perform a semi-quantitative Southern blot.

## **Electrophoresis**

Digested gDNA from wild type and sequentially transfected parasites corresponding to 150 million of cells were loaded in a 0.7% agarose gel in triplicate and run at 30 V overnight. Amounts of digested pEV\_Cas9v2 equivalent to  $150 \times 10^4$ ,  $150 \times 10^5$  and  $150 \times 10^6$  molecules were loaded as well. It is assumed that one cell contains on average one of each episome (Patnaik et al., 1993).

## **Preparation of the gel for transfer**

The gel was submerged in denaturation buffer (1.5 M NaCl and 0.5 M NaOH) and incubated for 30 min with gentle agitation and then washed for 10 min. The neutralization process was done placing the gel in sufficient neutralization buffer (1.5 M NaCl and 0.5 M Trizma base, pH 7.5) and incubated for 30 min with gentle agitation. The gel was washed again for 10 min.

## **Transfer of the DNA to a nylon membrane**

The gel was washed in 20x SSC for 10 min. The procedure to transfer the DNA was basically the same indicated in the northern blot section (see above). Upward transfer of DNA occurred overnight. The next day, the capillary transfer system was dismantled. The positions of the slots on the membrane were marked with a pencil through the gel. The membrane was transferred to a glass tray containing 300 ml of 5x SSC at 23 °C. The tray was placed in a platform shaker and agitated for 5 min.

The membrane was removed from the 5x SSC buffer and excess fluid was allowed to drain away. The membrane was laid down, DNA side upward, on a dry sheet of blotting paper for a few minutes. The gel was stained in ethidium bromide during one hour to check the efficiency of transfer. To cross-link by UV irradiation, the membrane was placed on a piece of dry blotting paper and irradiated at 254 nm using the program 0.120 (1 min).

### **Southern hybridization**

The membrane containing the target DNA was floated on the surface of a tray containing 5x SSC until it became thoroughly wetted from beneath for 2 min. The membrane was cut in three parts for each probe: Cas9, lactamase and  $\beta$ -tubulin. The resulting small membranes were incubated for two hours at 68 °C in 50 ml of 1x prehybridization solution (5x SSC, 5x Denhardt's solution, 0.5% (w/v) SDS, 10  $\mu$ g/ml herring sperm) using glass cylinders in an incubator with rotating wheels. Then, 12.5  $\mu$ l of each probe were resuspended in 100  $\mu$ l of water. The probes were denatured by heating for 10 min at 100 °C. Then they were chilled immediately on ice for two min. The denatured radiolabeled probes were added directly to the prehybridization solution. The incubation was continued for 12-16 hours at 68 °C. After hybridization, the membranes were transferred as quickly as possible to another cylinder containing 50 ml of 2x SSC, 0.5% SDS. The membranes were washed for 5 min at room temperature. After 5 min, the first rinse solution was poured off into a radioactivity disposal container and 50 ml of 2x SSC, 0.1% SDS were added to the cylinders. The membranes were



incubated for 15 min at room temperature. The rinse solution was replaced with 50 ml of fresh 0.1x SSC, 0.1% SDS. The membranes were incubated for 2 hours at 65°C. Then, they were briefly washed with 0.1x SSC at room temperature. The membranes were dried on blotting paper and wrapped in plastic foil. Then they were introduced in a cassette with a phosphor imager screen for 1 hour. The screen was scanned in a Thyphoon fla7000.

## Oligonucleotides, primers and cell lines

### Ribosome profiling

Oligonucleotide RNA markers (Sigma):

NI-NI-19: 5' -AUGUACACGGAGUCGAGCUCAACCCGCAACGCGA-(Phos)-3'

NI-NI-20: 5'-AUGUACACGGAGUCGACCCAACGCGA-(Phos)-3'

Preadenylylated and 3'-blocked linker (IDT):

1/5rApp/CTGTAGGCACCATCAAT/3ddC/

The designation 5rApp indicates preadenylation and 3ddC indicates that the linker is blocked at the 3'-end.

Reverse transcription primer (IDT):

5'-(Phos)-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGC-  
(SpC18)-CACTCA-(SpC18)-TTCAGACGTGTGCTCTTCCGATCTATTGATGGTGCCTACAG-3'

The designation (Phos) indicates 5' phosphorylation and -(SpC18)- indicates a hexa-ethyleneglycol spacer.

PCR primers (IDT):

Forward library primer: 5'-AATGATACGGCGACCACCGAGATCTACAC-3'

RNaseq Rev # 1 PCR primer:

5'-CAAGCAGAAGACGGCATAACGAGATAGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

RNaseq Rev # 2 PCR primer:

5'-CAAGCAGAAGACGGCATAACGAGAT ACTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

RNaseq Rev # 3 PCR primer:

5'-CAAGCAGAAGACGGCATAACGAGAT ATGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

RNaseq Rev # 4 PCR primer:

5'-CAAGCAGAAGACGGCATAACGAGAT ACGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

RNAseq Rev # 5 PCR primer:  
 5'-CAAGCAGAAGACGGCATAACGAGAT *AGCTGC*  
 GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'  
 RNAseq Rev # 6 PCR primer:  
 5'-CAAGCAGAAGACGGCATAACGAGAT *ATCGTA*  
 GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3'

The six-nucleotide barcode index for each reverse primer is indicated in italics.

## Northern blot

List of primers used to synthesize probes for VSG and CTR:

Probe	Primer name	Primer sequence	Ta (°C)	Band size (bp)
VSG	NB_VSG_F	5'-CGAGGAGCTAGACGACCAAC-3'	64	577
VSG	NB_VSG_R	5'-CATAGTGACCGCTGCAGAAA-3'	64	
CTR-U	CTR2 F1	5'-TACTAAGAGCAGTAATAAGTA -3'	46.3	805
CTR-U	CTR R3	5'-TTCCGCTGCCTTCATTTT-3'	63.7	
CTR-D	CTR3 F1	5'-GCTGATGATTGCACCACGTCT-3'	67.8	514
CTR-D	CTR R4	5'-GCGTTGCTCCTATATACT-3'	52.4	

Ta: Temperature of annealing.

## Circular RACE PCR

List of primers used to amplify cDNA from circular or linear RNA

Template	Primer name	Primer sequence	Ta (°C)	Observation
-	420_CTR_R3	5'-TTCCGCTGCCTTCATTTT-3'	-	Reverse transcription primer
Circular	1009_CTRFD	5'-CAAGCACAATTTTCATCCTCCT-3'	57	First PCR
Circular	1011_CTRRD	5'-TTGTGCACACAACCATAACG-3'	57	First PCR
Linear	617_3race_F2	5'-GAGTATGGCTGTTCTGCAGATTC-3'	60	First PCR
Linear	718_CTR_R3_n	5'-AATCAGAAAAACCTGAGCGC-3'	60	First PCR
Circular	1010_CTRFDn	5'-AAGAATTCGGAGACGAAGAGC-3'	57	Nested PCR
Circular	1012_CTRRDn	5'-TGAATCTGCAGAACAGCCATA-3'	57	Nested PCR
Linear	618_3race_F2_n	5'-GTTCTGCAGATTCAAACACTAGCG-3'	60	Nested PCR
Linear	718_CTR_R3_n	5'-AATCAGAAAAACCTGAGCGC-3'	60	Nested PCR

## CRISPR-Cas9 system

### List of plasmids:

Plasmid	Observations
pT13-11	Original episome for PF
pEV_Luc	Includes a polylinker
pEV_noLucUTR	Removal CDS, SAS, PAS and UTRs of luciferase and a vector sequence upstream
pEV_puro	Exchange of G418 for puromycin resistant cassette
pEV_Cas9v1_puro	Express Cas9v1
pEV_Cas9v2_puro	Express Cas9v2
pEV_SCD6_GFP_gBlock	Includes a gBlock containing the respiar sequence for SCD6_GFP
pEV_sgRNA1_SCD6_GFP	Includes the sgRNA1 flanked by HH and HDV ribozymes
pEV_sgRNA2_SCD6_GFP	Includes the sgRNA2 flanked by HH and HDV ribozymes
pEV_sgRNA3_SCD6_GFP	Includes the sgRNA3 flanked by HH and HDV ribozymes
pMOTag3G_SCD6	GFP tag in the C-terminus of SCD6 - G418 resistance cassette
pMOTag4G_SCD6	GFP tag in the C-terminus of SCD6 -Hygromycin resistance cassette
pEV_Cas9v2_puro_HSVTK	For positive (puromycin) and negative (HSVTK) selection

### Cell lines generated:

Cell lines	Observations
PF_pEV_Cas9v2_puro	PF cells expressing Cas9v2
PF_pEV_Cas9v2_puro/pEV_sgRNA1_SCD6_GFP	PF cells expressing Cas9v2 and SCD6 tagged with GFP using the CRISPR-Cas9 system
PF_pMOTag3G_SCD6	PF cells expressing SCD6 tagged with GFP using the pMOTag system

List of primers:

Purpose	Primer name	Primer sequence	Template	Observation
Deletion of short vector sequence	951_Epi_F1 980_Epi_novector1	5'-TTTGTGGCGGGTACCGGGCACAGCAAGCT-3' 5'-TTATTTTATGGCAGCAACGACACCT-3'	pEV_Luc	Keeps the polylinker region and the PARP SAS and 5'-UTR
Keeps the PARP 3'-UTR and SAS	955_Epi_F3 956_Epi_R3	5'-GAGCTGTACAAGTAAAATGTAACCTGTATTTCAGCGA-3' 5'-CGGTGTTAGGATCCGAGCTCGCAATTCCTTTTACT-3'	Template: pEV_Luc	
Insertion of the puromycin resistant cassette	1086_Epi_Puro_F2 1087_Epi_Puro_R2	5'-CACAAAGCTTTTAAACATGACCGAGTACAAGCCC-3' 5'-TTGATCCCGGATCCGGATCAGGCACCGGGCTTGGC-3'	pNS110 puro	
Amplifies the H2B NLS and SpCas9	1180_pEV_Cas9_puro_F1 1181_pEV_Cas9_puro_R1	5'-GCTGCCATAAAAATAACCTGACAGGATGGCCACTCCTAAGAGCAC-3' 5'-TTAGTCGCCCTCCAGCTG-3'	pPa_H2BNLSCas9	
Keeps the SAS and 5'-UTR of the PARP gene	1162_Epi_Puro_F1 1085_Epi_Puro_R1	5'-CTAAAAGGAATTCGAGC-3' 5'-GTTAACAGCTTTGTAAT-3'	pEV	
Amplifies part of the actin PAS	1182_pEV_cas9_puro_F2 956_Epi_R3	5'-CTGGGAGGGACTAACTTAAAGATGTAACCTGTATTTCAGGCGATG-3' 5'-CGGTGTTAGGATCCGAGCTCGCAATTCCTTTTACT-3'	pEV_Luc_puro	
Upstream reaction	1170_SDM1 1171_SDM2	5'-TCCGGATCCGGATCAAGGG-3' 5'-AGCTAGCTTTGCATGCCCTGAGGCAAGCCACATT-3'	pEV_noLucUTR	
Downstream reaction	1172_SDM3 1173_SDM4	5'-AATGTGGGCTCTGCCCTGAGGCAATGCAAGCTAGCT-3' 5'-CTGGCTAGGGATGACCTGCTG-3'	pEV_noLucUTR	
Cloning the 3'-end of SCD6	1337_pMOT_SCD_F1 1338_pMOT_SCD_R1	5'-ACGGGGCCACAAGCCCACTCCACA-3' 5'-CAAGCTCGAGTTGTAGCCGGCTTGTGA-3'	gDNA	
Cloning the 3'-UTR of SCD6	1339_pMOT_SCD_F2 1340_pMOT_SCD_R2	5'-CAAGGATCCATAAGGGGCTCGGGTGGG-3' 5'-TTGGAGCTCTCTCTTACTTCTCATCCGCTCA-3'	gDNA	
Amplifies the PAS from actin and the SAS of PARP	1459_TK_puro_F1 1460_TK_puro_R1	5'-ATTGAGCTCGGATCCCTA-3' 5'-GTTAACAGCTTTGTGAATTTACTT-3'	pEV_Cas9v2_puro	
Amplifies the puromycin-HSVTK cassette	1086_Epi_Puro_F2 1461_TK_puro_R2	5'-CACAAAGCTTTTAAACATGACCGAGTACAAGCCC-3' 5'-TTGCATCCGGATCCGGATCAGTACGCTCCCCCAT-3'	pyrFEKO_puro	
Evaluation of genome editing and allelic efficiency	1327_SCD_gDNA_F 1328_SCD_gDNA_R	5'-GAGGAGACCTGGGTGTTGAA-3' 5'-TCCTCATCCCTCATCTTTG-3'	gDNA	gDNA from cells transfected with the 2 episomes
Evaluation of genome editing and allelic efficiency	1529_probe_Cas9_F 1530_probe_Cas9_R	5'-GCTACGCGGGCTAGATTGAC-3' 5'-AGGTCCAAGATGGCTTTT-3'	pEV_Cas9v2_puro	Length: 580
Cas9 probe	1531_probe_AmpR_F 1532_probe_AmpR_R	5'-CCGAAAGCTGGTGAAG-3' 5'-TTGTTCCGGGAAGCTAGAG-3'	pEV_Cas9v2_puro	Length: 541
Lactamase probe	376_Tub_F_qPCR 267_β_tubulin_R	5'-ACTGGGCAAGGGCCACTAC-3' 5'-ATCTTCACTCCCTGGGTGG-3'	gDNA	Length: 790

## Reagents, equipment and consumables

### Ribosome profiling

Reagents:

Reagent	Manufacturer	Catalogue number
0.5 M EDTA pH 8.0	ThermoFisher Scientific	AM9260G
10 bp DNA ladder	ThermoFisher Scientific	10821015
1M MgCl <sub>2</sub>	ThermoFisher Scientific	AM9530G
2-propanol 99.9%	Roth	T910.1
2M KCl	ThermoFisher Scientific	AM9640G
2X KAPA HiFi Hot start Mix	KAPA BIOSystems	KM2605
3M Sodium acetate pH 5.5	ThermoFisher Scientific	AM9740
5M NaCl	ThermoFisher Scientific	AM9760G
CircLigase II	Epicentre	CL9025K
Cycloheximide	Applichem	A0879.0005
D(+)-Saccharose	Roth	9097.1
DEPC	Roth	K028.2
Dynabeads® mRNA Purification Kit	ThermoFisher Scientific	61006
Ethanol 99.8% p.a	Roth	9065.4
GlycoBlue	ThermoFisher Scientific	AM9515
n-octylglycoside	Sigma	O8001-1G
NaOH 99% p.a	Roth	6771.1
RNA Fragmentation Reagent	ThermoFisher Scientific	AM8740
RNase I (Ambion)	ThermoFisher Scientific	AM2295
RNase-free water	ThermoFisher Scientific	AM9939
Roti-Aqua-Phenol-Chloroform-Isoamyl alcohol	Roth	X985.1
Roti-Stock 20% SDS	Roth	1057.1
Rotiphorese Gel 40 (19:1)	Roth	3030.1
SUPERase•In RNase inhibitor	ThermoFisher Scientific	AM2696
SuperScript III reverse transcriptase	ThermoFisher Scientific	18080044
SYBR Gold nucleic acid gel stain	ThermoFisher Scientific	S11494
T4 polynucleotide kinase	NEB	M0201S
T4 RNA Ligase 2, truncated	NEB	M0242L
Tris-HCl pH 7.5	ThermoFisher Scientific	15567027
Tris-HCl pH 8.0	ThermoFisher Scientific	15568025
TURBO DNaseI	ThermoFisher Scientific	AM2238
Urea 99.5%	Roth	3941.2

## Equipment:

Equipment	Manufacturer	Catalogue number
Bioanalyzer 2100	Agilent Technologies	G2940CA
C1000 Touch <sup>a</sup> Thermal Cycler	Biorad	1851148
Gradient Station	Biocomp	153-001
Model EM-1 Econo™ UV Monitor	Biorad	1-800-424-6723
Nanodrop 2000	ThermoFisher Scientific	ND-2000
Optima™ L80 XP ultracentrifuge	Beckman Coulter	392049
XCell SureLock® Mini-Cell	ThermoFisher Scientific	EI0001

## Consumables:

Item	Manufacturer	Catalogue number
Costar® Spin-x® centrifuge tube filter	Sigma	CLS8162-96EA
Empty Gel Cassettes, mini, 1.0 mm	ThermoFisher Scientific	NC2010
Eppendorf® LoBind microcentrifuge tubes 0.5 ml	Eppendorf	0030 108.035
Eppendorf® LoBind microcentrifuge tubes 1.5 ml	Eppendorf	0030 108.051
High sensitivity DNA chip	Agilent Technologies	5067-4626
Open-top polyclear centrifuge tubes 14x95 mm	Seton	7031

## Characterization of the CTR

### Reagents:

Reagent	Manufacturer	Catalogue number
Agarose		
2X KAPA HiFi Hot start Mix	KAPA BIOSystems	KM2605
5'-Polyphosphatase	Epicentre	RP8092H
$\alpha$ P <sup>32</sup> -dATP	Hartmann Analytic	-
DECAprime™ II Kit	ThermoFisher Scientific	AM1455M
DMSO 99.5%	Sigma	D4540
DreamTaq™ DNA polymerase	ThermoFisher Scientific	EP0703
Glycerol Ultrapure	ThermoFisher Scientific	15514-011
Glyoxal	Sigma	50649-25ML
NucleoSpin Gel and PCR Clean-up kit	Macherey-Nagel	740609.250
NucleoSpin RNA II kit	Macherey-Nagel	740955.50
NucleoSpin® Plasmid kit	Macherey-Nagel	740588.250
pCR™2.1-TOPO® vector	ThermoFisher Scientific	45-0641
Pre-hybridization solution (Roti®-Hybriquick)	Roth	A981.1
RNaseOUT Recombinant Rnase Inhibitor	ThermoFisher Scientific	10777019
T4 RNA Ligase 1 (ssRNA Ligase)	NEB	M0204S
Tobacco Acid Pyrophosphatase (TAP)	Epicentre	187250
UltraPure agarose	ThermoFisher Scientific	16500-500

## Equipment:

<b>Equipment</b>	<b>Manufacturer</b>	<b>Catalogue number</b>
C1000 Touch <sup>a</sup> Thermal Cycler	Biorad	1851148
Model 583 Gel Dryer	Biorad	1651745
Nanodrop 2000	ThermoFisher Scientific	ND-2000
Thyphoon fla7000	GE healthcare	28-9558-09

## Consumables:

<b>Item</b>	<b>Manufacturer</b>	<b>Catalogue number</b>
Illustra Microspin G25 columns	GE healthcare	27-5325-01
Nylon membrane (Amersham Hybond-XL)	GE healthcare	RPN203S

## CRISPR-Cas9 system

### Reagents:

<b>Reagent</b>	<b>Manufacturer</b>	<b>Catalogue number</b>
AflII	NEB	R0520S
Anti-mouse Alexa Fluor-488 donkey	ThermoFisher Scientific	A21202
Anti-rabbit Alexa 594 IgG chicken	ThermoFisher Scientific	A-21442
Apal	NEB	R0114L
BamHI	NEB	R3136L
DNAzol	ThermoFisher Scientific	10503027
Blood & cell culture DANN MIDI kit	Qiagen	13343
Formaldehyde	Sigma	F8775
Ganciclovir	Sigma	G2536
Herring sperm	Promega	D1811
In-Fusion <sup>®</sup> HD Cloning Kit	Clontech	638918
KpnI-HF	NEB	R3142L
Mouse mAb for CRISPR-Cas9	Abcam	191468
NextSeq 500/550 v2 Kit	Illumina	FC-404-2005
NheI	NEB	R3131S
NP40 (IGEPAL <sup>®</sup> CA-630)	Sigma	I8896-50ML
Phusion High-Fidelity DNA Polymerase	ThermoFisher Scientific	F-530L
Phusion Human Specimen Direct PCR kit	ThermoFisher Scientific	F-150BID
RNase H E. coli	ThermoFisher Scientific	AM2292
SacI-HF	NEB	R3156S
SbfI-HF	NEB	R3642L
SDS pellets	Roth	CN30.3
T4 DNA Ligase system	ThermoFisher Scientific	EL0011
Vectashield DAPI	Biozol	VEC-H-1200
XhoI	NEB	R0146S



Equipment:

<b>Equipment</b>	<b>Manufacturer</b>	<b>Catalogue number</b>
Bioanalyzer 2100	Agilent Technologies	G2940CA
Bioruptor Plus	Diagenode	B01020001
C1000 Touch <sup>a</sup> Thermal Cycler	Biorad	1851148
Leica DMI6000B microscope	Leica	Leica DMI6000B
MACSquant Analyzer 10	Miltenyi Biotec	130-096-343
Model 583 Gel Dryer	Biorad	1651745
NextSeq500 sequencer	Illumina	SY-415-1001
Nucleofector™2b	Lonza	AAB-1001
Qubit 2.0 Fluorometer	ThermoFisher Scientific	Q32866
TCS SP5 II confocal microscope	Leica	Leica TCS SP5 MP
Thyphoon fla7000	GE healthcare	28-9558-09

Consumables:

<b>Item</b>	<b>Manufacturer</b>	<b>Catalogue number</b>
Elektroporationsküvetten Cuvettes Plus™, BTX	VWR	732-0021
High sensitivity DNA chip	Agilent Technologies	5067-4626
Illustra Microspin G25 columns	GE healthcare	27-5325-01
Nylon membrane (Amersham Hybond-XL)	GE healthcare	RPN203S

# **Chapter 3. Implementation of ribosome profiling in two life cycle stages of *T. brucei* for the study of translational regulation**

## **Introduction**

Gene expression regulation in eukaryotic cells can occur at the level of epigenetic control, transcriptional initiation, elongation, RNA processing and stability, protein translation and protein turnover (Lelli et al., 2012; Moore, 2005). Apparently, the impact of each one of these factors is different between organisms and even between cell types in multicellular organisms. The importance of gene expression regulation at the level of translation has become evident in several eukaryotic organisms such as *P. falciparum* asexual (Caro et al., 2014) and sexual blood stages (Mair et al., 2006), *Arabidopsis* (Juntawong et al., 2014) and mammalian cells (de Klerk et al., 2015; Ingolia et al., 2012; Jang et al., 2015; Michel and Baranov, 2013; Schwanhäusser et al., 2011). In addition, other studies have shown the weak correlation between mRNA and protein levels in various organisms (de Godoy et al., 2008; Lahav et al., 2011; Lu et al., 2007; McNicoll et al., 2006; Nagaraj et al., 2014).

Three features suggest that there is no regulation at the level of transcription initiation for RNA polymerase II transcribed genes in *T. brucei*: the absence of promoter sequence motifs, an open chromatin structure at TSSs and the organization of genes in PTUs (Siegel et al., 2009; Wright et al., 2010). The apparent lack of regulation at this step makes *T. brucei* a good model organism to

study post-transcriptional control of gene expression. Post-transcriptional gene regulation is important for the adaptation of the parasite to the different environments it is exposed during its life cycle. For example, in the mammalian bloodstream the temperature is 37 °C and the production of energy depends on glycolysis; but in the tsetse midgut the temperature is 27 °C and the metabolism is based in oxidative phosphorylation (Clayton and Michels, 1996; Evans and Brown, 1972; Fairlamb and Opperdoes, 1986; Opperdoes, 1987). Before the implementation of ribosome profiling in *T. brucei*, the knowledge about post-transcriptional regulation was limited to the genome-wide analysis of mRNA stability (Manful et al., 2011), identification of sequence motifs capable of modulating RNA stability or translational efficiency in a life cycle-specific manner (Furger et al., 1997; Hehl et al., 1994; Helm et al., 2009; Hotz et al., 1997; Mayho et al., 2006; Walrad et al., 2009) and the finding that a bias in codon usage can affect translational efficiency (Horn, 2008).

A polysome profiling approach was the only precedent to quantify translation in *T. brucei* (Brecht and Parsons, 1998; Capewell et al., 2013) but unfortunately this approach lacks single ribosome resolution in the larger polysomal fractions (Arava et al., 2003). After the introduction in 2009 of the higher performance approach termed ribosome profiling, a new opportunity for the study of translational regulation in a more accurate way became available. For this reason, I was interested in implementing this powerful technology in an organism that relies in post-transcriptional gene regulation control. Besides, the *T. brucei* genome codes for an unusually larger number of proteins involved in translation initiation control, including two to three homologues of poly(A)

binding proteins as well as four isoforms of the translation initiation factor (eIF4E) and five of eIF4G (Dhalia et al., 2005; Gaudenzi et al., 2005; Yoffe et al., 2004, 2006; Zinoviev and Shapira, 2012). Moreover, RNA-seq analyses have revealed widespread alternative trans-splicing, resulting in transcripts with different 5'-UTR lengths (Kolev et al., 2010; Nilsson et al., 2010; Siegel et al., 2010) that may contain regulatory elements influencing translational efficiency (Siegel et al., 2011) such as uORFs.

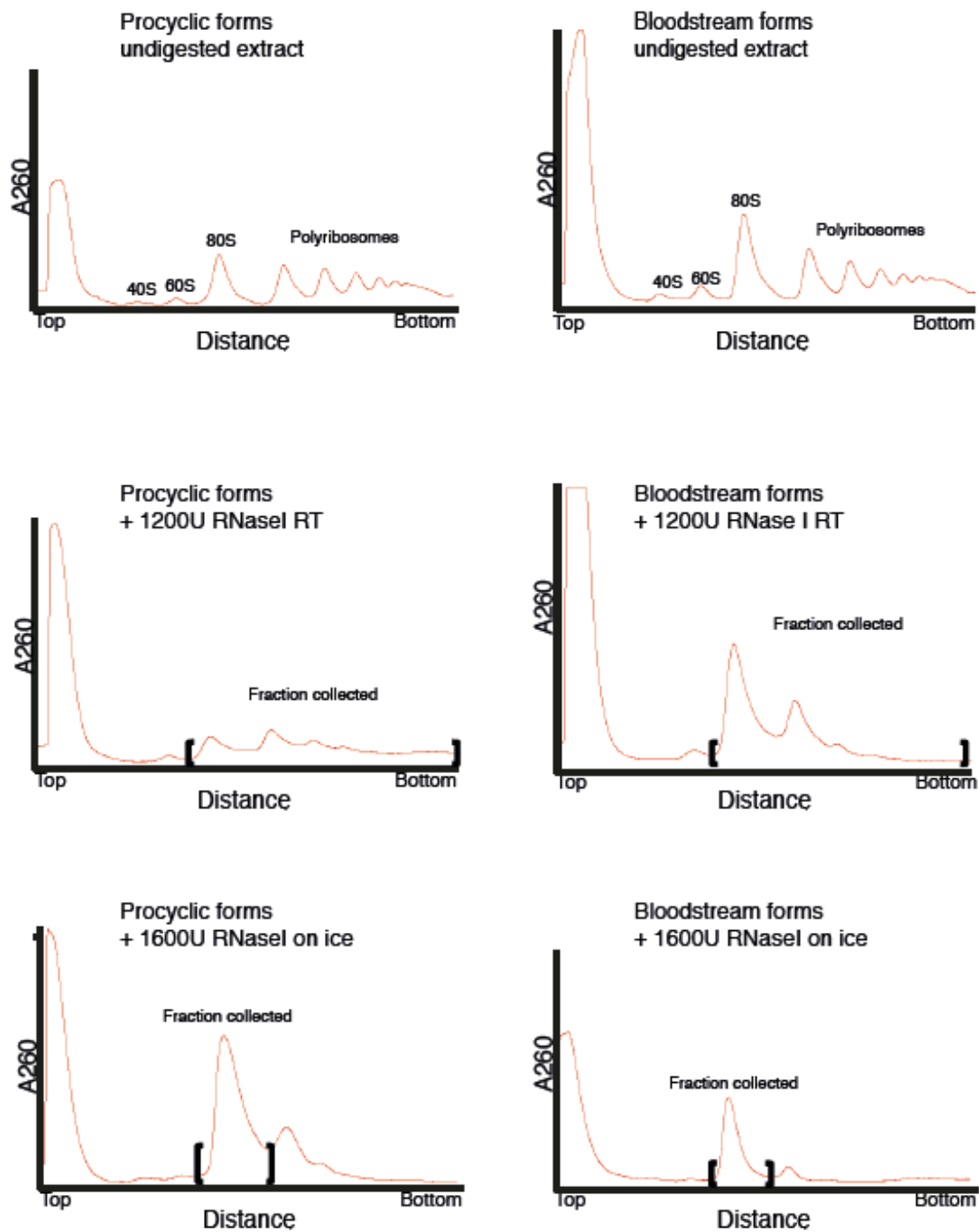
In addition of quantifying translation, ribosome profiling provides position-specific information regarding the location of ribosomes on mRNA transcripts that can be used to accurately map the correct CDSs and to identify novel CDSs that have been missed in previous genome annotations (Michel and Baranov, 2013).

Here I describe in detail the protocol used to generate ribosome-profiling libraries in two life-cycle stages of the parasite. Our analysis confirmed the importance of translational regulation in *T. brucei*. For example we found a 100-fold range in translational efficiency among genes and life cycle-specific differences in translational efficiency for a large number of genes. In addition, the ribosome profiling data enabled us to identify previously un-annotated CDSs, incorrectly annotated translation initiation sites and the potential regulatory importance of uORFs (Vasquez et al., 2014).

## Results

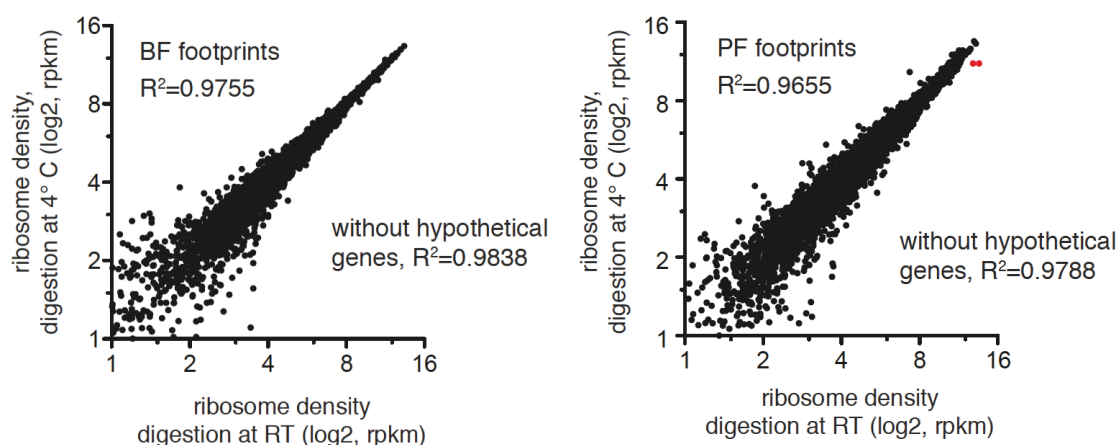
### **Ribosome footprints indicate ribosome position at sub-codon resolution**

I based my work in two protocols previously published to implement ribosome profiling in yeast and mammalian cells (Ingolia et al., 2009, 2012). In brief, I sequenced ribosome footprints and fragmented mRNA from two trypanosome stages adapted to *in vitro* culture, the PF of the tsetse fly midgut and the BF of the mammalian bloodstream. The cells were treated for 2 min with cycloheximide to arrest translating ribosomes and immediately harvested because it has been shown that this antibiotic stabilizes RNA transcripts in *T. brucei* (Webb, 2005). To find optimal digestion conditions, I generated two different ribosome footprint libraries by using two different RNaseI concentrations (1,200 and 1,600 units) and temperatures (4 °C and room temperature). Next, I used sucrose gradients to recover specifically the monosome fraction (Figure 15). The spread pattern observed in the PF sample digested with 1,200 units of RNaseI at room temperature looked unusual but the results obtained after analyze the corresponding library were similar to the sample treated with 1,600 units of RNaseI at 4 °C.



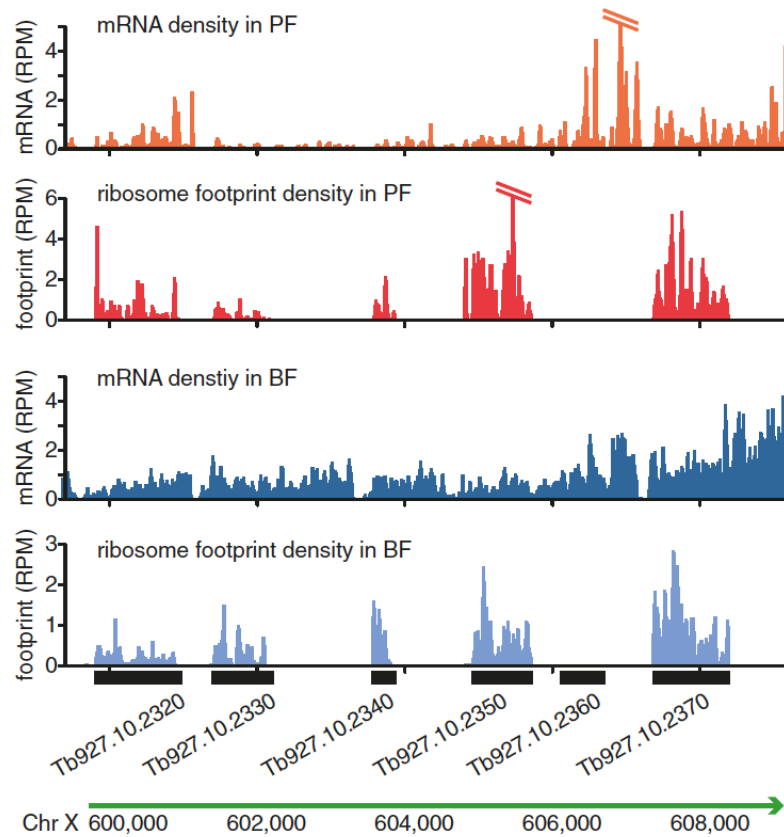
**Figure 15.** Sucrose density gradient profiles of undigested and RNaseI-digested samples from PF and BF. In the undigested samples, the peaks from left to right correspond to free RNA, the 40S ribosome subunit, the 60S ribosome subunit, the 80S or monosome fraction and the polysome fraction. In the nuclease treated samples, the brackets indicate the fraction collected and that corresponds to the 80S peak.

Although the libraries were produced under different conditions (units of RNaseI and temperature of incubation), they presented high reproducibility, which highlights the power of the ribosome profiling approach in spite of the numerous steps necessary to generate the libraries (Figure 16).



**Figure 16.** Pair-wise comparison of ribosome footprint density from BF (left panel) and PF (right panel) samples digested at different temperatures of incubation (room temperature and 4 °C). Density was measured as rpkm to normalize for differences in gene length and library size. In the right panel the genes marked in red are hypothetical genes presenting similarity with rRNA genes.

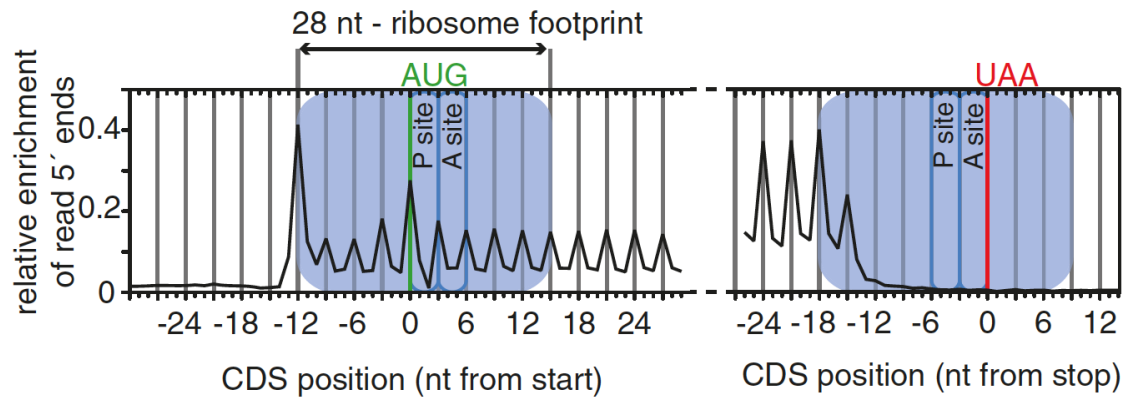
In this study, RNA was obtained from the Lister 427 strain and sequence reads were aligned to the more complete genome of the 927 strain. Even though, RNA-seq analysis indicates widespread transcription, alignment of ribosome profiling reads revealed that the nuclease-protected reads predominantly mapped along CDSs with very few reads aligning to intergenic regions (Figure 17).



**Figure 17.** Ribosome footprints are enriched specifically across CDSs. mRNA densities and footprint densities are shown as rpm to normalize for differences in library size. The region shown corresponds to a section of chromosome X. Black boxes represent CDSs and the corresponding gene ID number is shown below. The green horizontal line represents the direction of transcription.

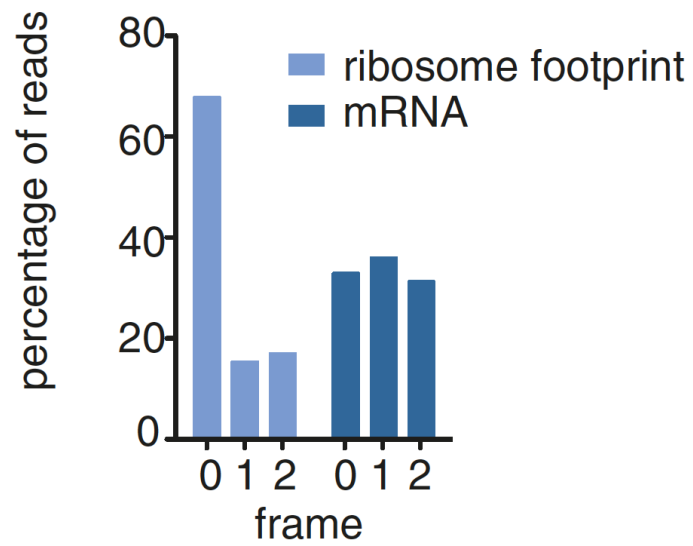
This result confirmed that the nuclease-resistant RNA fragments obtained corresponded to ribosome footprints and not to other RNA fragments protected by RNA-binding proteins. Furthermore, the 5'-nucleotide of the ribosome profiling reads aligned from 12-nt upstream of the ATG codon to ~18-nt upstream of the stop codon (Figure 18).



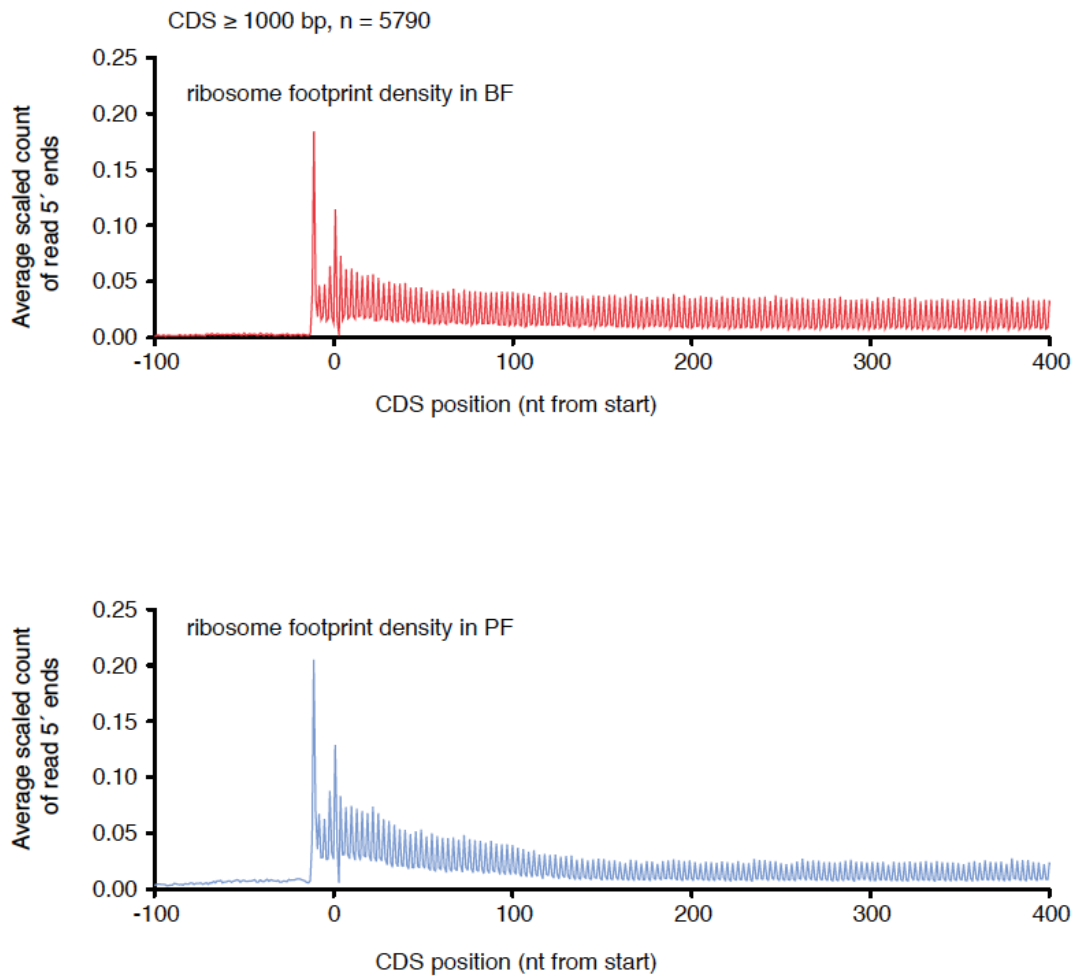


**Figure 18.** Alignment of the 5'-nucleotide from ribosome footprint reads that align close to translation start or translation termination sites. Blue boxes indicate the approximate size of the ribosome footprint.

Taking into account that the footprint length is approximately 28–30 nt, the ribosome protects mRNA from 12-nt upstream of the start codon to 9–11 nt downstream of the stop codon. In contrast, RNA-seq reads aligned along the CDSs and UTRs, as expected (Figure 17). Additional evidence supporting that the nuclease-resistant RNA fragments were real footprints came from the alignment of the 5'- nucleotide reads. The pattern observed revealed a distinct 3-nt periodicity, with 68% of BF ribosome profiling reads starting at the first nucleotide of a codon which is not expected for the mRNA sequence reads (Figures 18, 19 and 20).

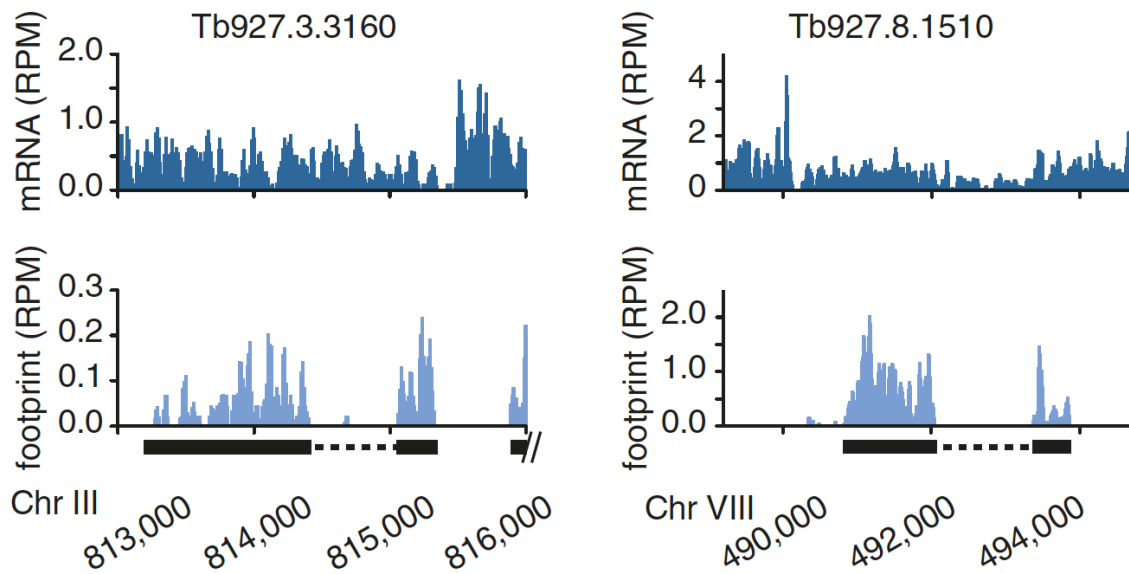


**Figure 19.** Ribosome footprints reveal coding sequences at sub-codon resolution. Percentage of reads according to their position relative to reading frame in BF.



**Figure 20.** Meta-gene analysis of ribosome density. Alignment of the 5'-nucleotide from ribosome footprint reads that map close to translation start or translation termination sites. To calculate average scaled count of read 5'-ends, the peak count within each gene was scaled to 1. The scaled counts at each position of all genes were summed up and then divided by the number of genes.

Furthermore, in *T. brucei* only 2 genes have been reported to have introns (poly(A) polymerase and a DNA/RNA helicase) (Berriman et al., 2005; Mair et al., 2000) and we did not find evidence those introns were involved in translation (Figure 21).



**Figure 21.** RNA-seq and ribosome profiles for the two intron-containing genes poly(A) polymerase (Tb927.3.3160) and a DNA/RNA helicase (Tb927.8.1510). Exons are represented as black boxes and introns as a dashed line.

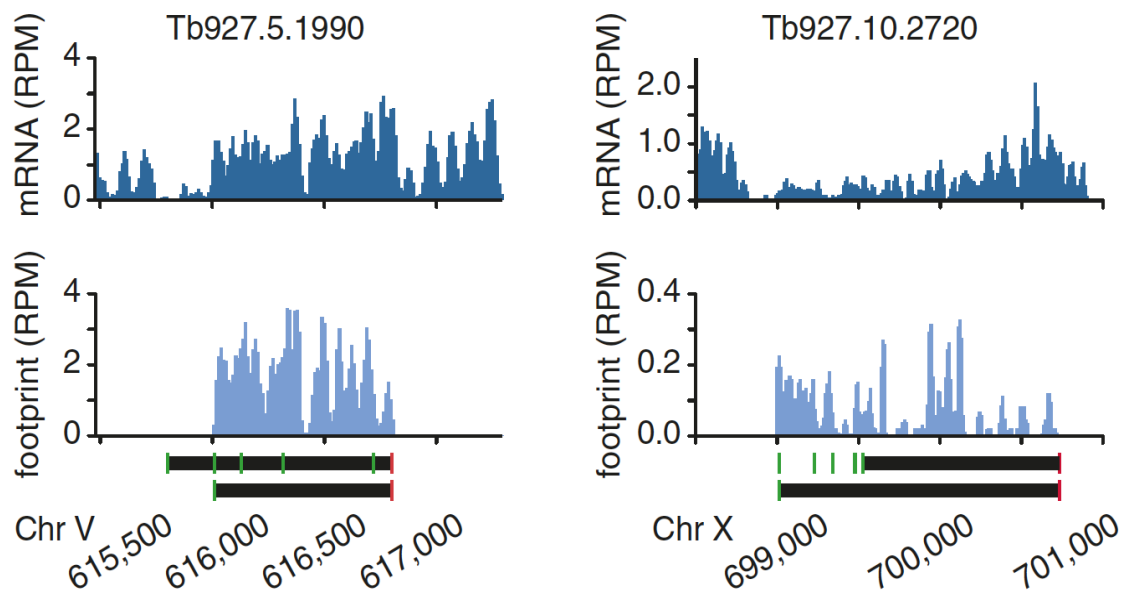
The enrichment of nuclease-protected reads across CDSs, the absence of them across introns and the observed 3-nt periodicity are all characteristic of ribosome-protected RNA footprints (Ingolia et al., 2009). These footprints can be used as a proxy to measure protein synthesis and to identify new CDSs. Moreover, some previously annotated CDSs did not show evidence of translation (for an example see the gene Tb927.10.2360 in Figure 17) because they may have been either incorrectly annotated or not translated in the two *in vitro* life cycle stages examined here.

### Unambiguous identification of translation initiation sites

Ribosome profiling allows the accurate identification of translation initiation sites that was not always possible in the past using the RNA-seq available data. Indeed, SASs were sometimes mapped within or upstream of annotated CDSs (Kolev et al., 2010; Nilsson et al., 2010; Siegel et al., 2010). These

findings mean that for some genes the true CDS was shorter than the annotated CDS while for others, translation initiation occurred upstream of the annotated translation initiation site. With the information generated in my thesis, it is now possible to identify mis-annotations and to determine the most probable translation initiation sites for those CDSs.

Tb927.10.2720 is an example of a gene that presents translation upstream and in frame of the annotated CDS suggesting that the true CDS is longer than the one previously annotated (Figure 22, right panel). On the other hand, for Tb927.5.1990 no translation was observed between the first and the second AUG (green bars) indicating that the upstream region is not involved in translation (Figure 22, left panel).



**Figure 22.** Ribosome profiling results of two likely mis-annotated CDSs. Black boxes indicate annotated CDSs. Green lines represent ATG codons and red lines represent stop codons. (right panel) Tb927.10.2720 represents an example of a gene that is probably longer than the annotated one. (left panel) Tb927.5.1990 is an example of a gene for which translation begins downstream of the initially reported start site.

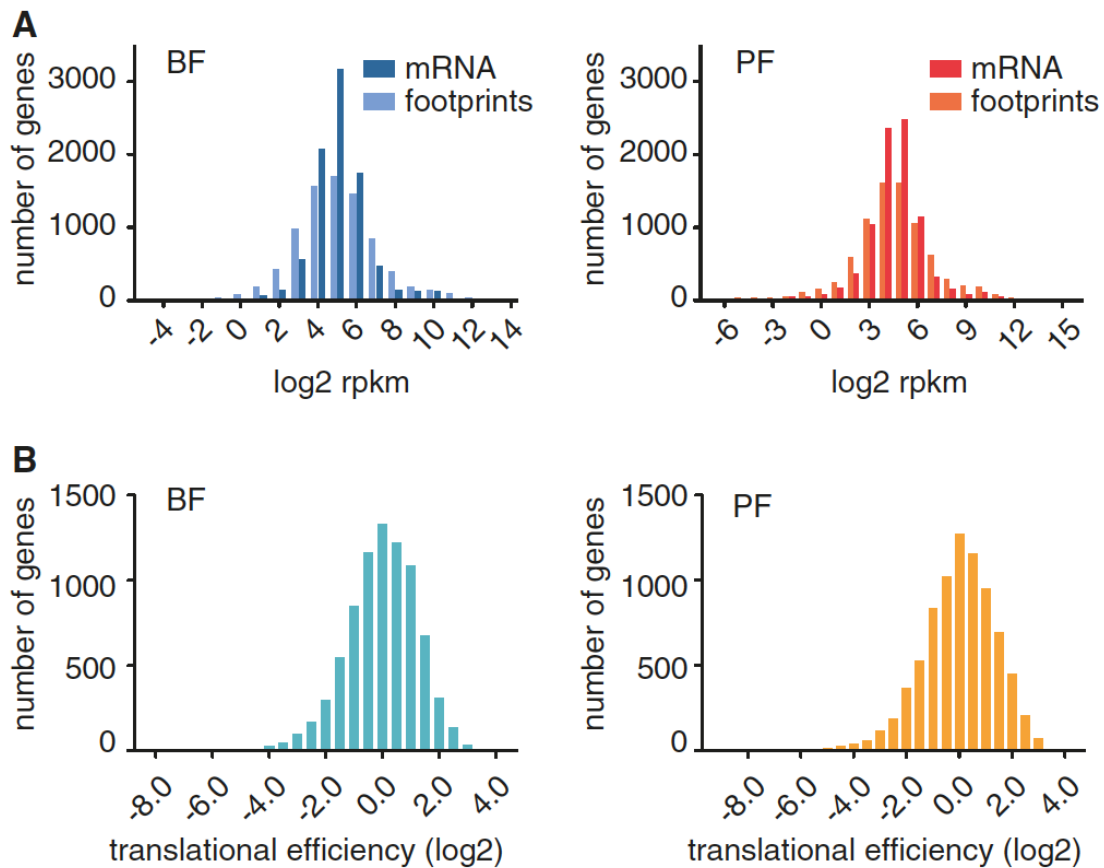
### Evaluation of translational efficiency by ribosome profiling

We used the ribosome profiling dataset to evaluate on a genome-wide scale the degree of regulation occurring at the level of translation in *T. brucei*. This study complemented a previous one measuring the impact of RNA stability in the control of gene expression (Manful et al., 2011).

It is possible to estimate the rate of protein synthesis by measuring the density of ribosome footprints on its transcript (measured as rpkm). Of 77 million PF and 44 million BF ribosome footprint sequence reads, 28% (PF) and 14% (BF) were aligned to annotated CDSs, while 70% (PF) and 85% (BF) mapped to known structural ncRNA. The high percentage of structural ncRNA corresponds mainly to rRNA present in the monosome fraction as has been

shown previously (Ingolia et al., 2009). The lower percentage of footprint reads in the BF sample compared with the PF sample can be explained by the fact the libraries were generated from the wild-type Lister 427 strain but mapped to the annotated genome of the TREU 927 strain. These two strains differ for example in the sequence of the *VSG* genes and therefore we were not able to align footprints from these genes in this analysis.

A gene was considered translated when we were able to map  $\geq 10$  ribosome footprint reads on it. That was the case for 8,072 genes, corresponding to 82% of previously annotated CDSs. The rate of translation differed among proteins for the two life cycle stages analyzed. In the PF, a 11,448-fold difference in ribosome density was observed (3,548-fold difference in mRNA levels) between the 1% most highly and the 1% most weakly translated proteins. In BF, a 1,623-fold difference in ribosome density was observed (262-fold difference in mRNA levels) (Figure 23A). These numbers indicate an important variability in the rate of translation for different genes in both life cycle stages analyzed.



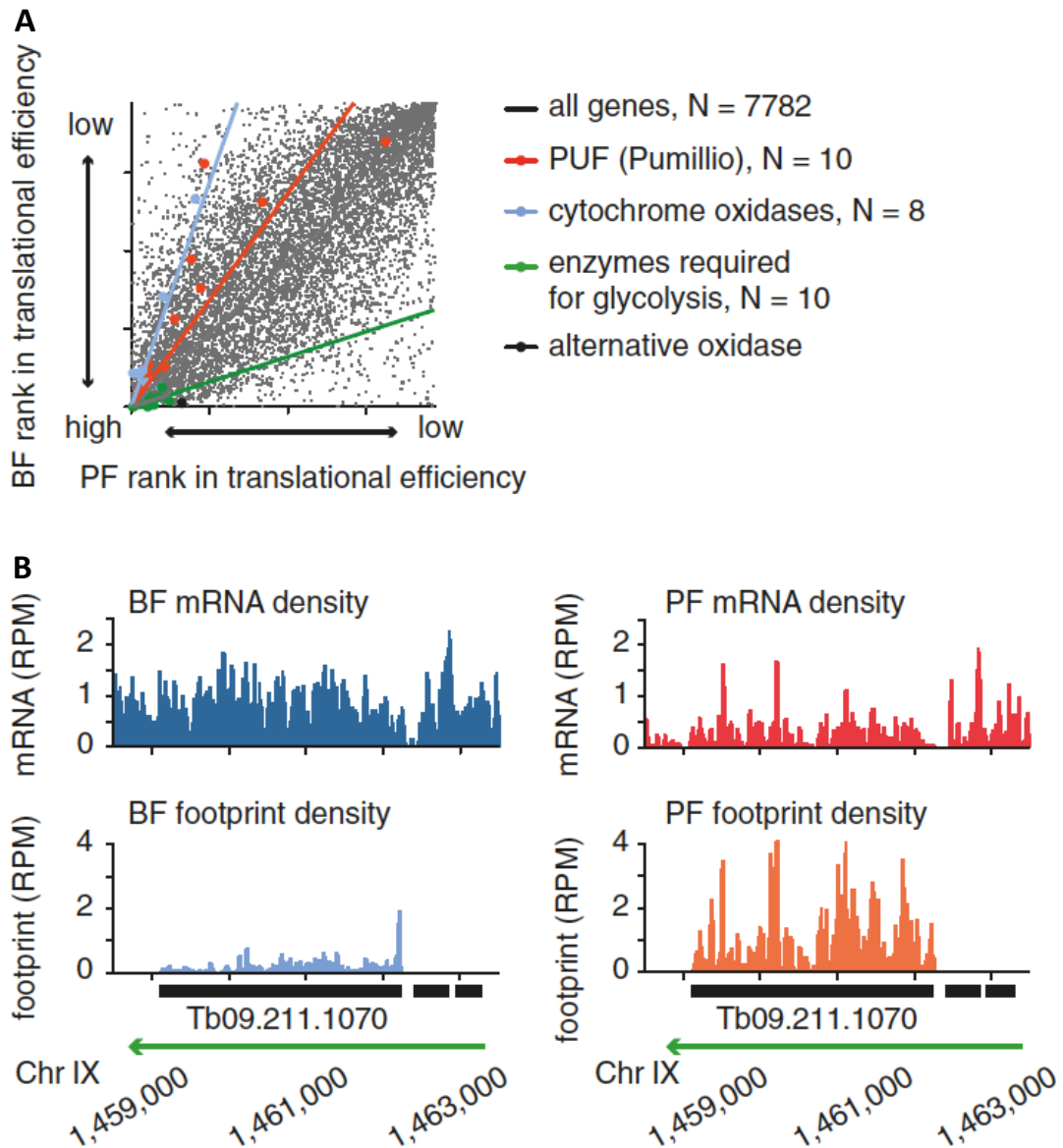
**Figure 23.** Translational regulation in PF and BF. (A) Histograms of mRNA abundance and ribosome density (rate of protein synthesis) for BF (left panel) and PF parasites (right panel). (B) Histogram of translational efficiency (ratio of ribosome footprint density to mRNA abundance) for BF (left panel) and PF cells (right panel).

Translational efficiency is defined as the amount of proteins produced per transcript and was calculated upon normalizing differences in mRNA abundance for individual genes (dividing the ribosome footprint density by the levels of mRNA abundance). Whenever determining the translational efficiency, mRNA and footprint reads mapping to the first 40-nt of a CDS were excluded to avoid possible artifacts from ribosome stalling at translation initiation sites (Ingolia et al., 2009). This is important to be mentioned because there was a minor increase in read density towards the 5'-end of CDSs in the PF samples (Figure 20). We observed a 117- and a 64-fold range in translational efficiency between the 1%



most efficiently and the 1% least efficiently translated proteins in PF and BF respectively (Figure 23B). No correlation between RNA abundance and translational efficiency was observed ( $R^2= 0.03$ ), which implies that translational efficiency is regulated independently of RNA stability. In summary, these results indicate that differences in translational efficiencies can significantly contribute to the regulation of gene expression.

Next, we compared the translational efficiency between the two life cycle stages evaluated in this work. It is possible to compare the rate of protein synthesis when the speed of translation is assumed to remain constant, which has been observed in mouse embryonic stem cells (Dana and Tuller, 2012; Ingolia et al., 2011). Nevertheless, it is not valid to assume this is true in the two life cycle stages because they live at different temperatures. For this reason, we ranked the genes in each stage based on their relative translational efficiency and compared their rank between the two life cycle stages. A pair-wise comparison points to a positive correlation between translational efficiency in the two life cycle stages analyzed (Figure 24A, Pearson's correlation coefficient = 0.7428,  $P < 0.0001$ , 95% CI 0.7327–0.7526). However, some genes presented distinct life cycle specific differences. For example, 58 genes found in the lowest 25% group of translational efficiency in the PF were among the top 25% most efficiently translated proteins in the BF (Table 3 and for an example see Figure 24B).



**Figure 24.** Life cycle regulation of translational efficiency. (A) Pairwise comparisons of translational efficiency in PF and BF. CDSs were ranked based on translational efficiency (1=highest translational efficiency, 7,782=lowest translational efficiency) in the two life cycle stages analyzed. A correlation of  $R=0.7428$  was obtained after comparing the ranks in translational efficiencies between both samples. Color code for gene families with developmentally regulated translational efficiencies: Pumillio genes (red), cytochromes oxidase (blue) genes required for glycolysis (green) and the alternative oxidase (Tb927.10.7090, black). (B) Ribosome profile of the Tb09.211.1070 gene with developmentally regulated translational efficiency. The gene presents a higher translational efficiency in PF. Green arrow indicates the direction of transcription.

**Table 3.** Developmentally regulated translation in BF and PF

Gene ID	Description	BF rank	PF rank	Change in rank (PF-BF)
<b>Translation up-regulated in BF</b>				
Tb927.7.3250	Expression site-associated gene 6 (ESAG6) protein, putative	1372	7405	6033
Tb927.4.3980	Chaperone protein DNAJ, putative	1301	7248	5947
Tb927.10.4780	GPI inositol deacylase precursor (GPIdeAc)	390	5994	5604
Tb927.8.8140	Small GTP-binding rab protein, putative	1217	6804	5587
Tb927.6.3480	RNA-binding protein, putative (DRBD5)	1856	7262	5406
Tb11.01.4701	Membrane-bound acid phosphatase 1 precursor (MBAP1)	689	6065	5376
Tb927.1.4650	Cyclin-like F-box protein (CFB2)	10	5289	5279
Tb927.3.5660	UDP-Gal or UDP-GlcNAc-dependent glycosyltransferase, putative	797	6041	5244
Tb927.2.6000	Glycosylphosphatidylinositol-specific phospholipase C (GPI-PLC)	1634	6793	5159
Tb927.4.5310	Serine/threonine-protein kinase a, putative	1656	6776	5120
<b>Translation up-regulated in PF</b>				
Tb927.5.440	trans-Sialidase, putative	6785	418	-6367
Tb927.7.6850	trans-Sialidase (TbTS)	7513	1445	-6068
Tb927.7.7470	Receptor-type adenylate cyclase GRESAG 4, putative	7001	1268	-5733
Tb927.1.2120	Calpain, putative	6258	739	-5519
Tb927.4.360	1,2-Dihydroxy-3-keto-5-methylthiopentene dioxygenase, putative	6587	1089	-5498
Tb09.160.5550	Calpain-like cysteine peptidase, putative	6936	1868	-5068
Tb927.8.7690	Amino acid transporter (pseudogene), putative	6693	1731	-4962
Tb927.8.1610	MSP-B, putative	7199	2272	-4927
Tb11.01.6650	Serine/threonine-protein kinase, putative	5280	363	-4917
Tb927.7.7110	Leucine-rich repeat protein (LRRP), putative	6501	1601	-4900
<b>Genes for which translational up-regulation in PF was previously shown</b>				
Tb927.10.280	Cytochrome oxidase subunit VI (COXVI)	2820	810	-2010
Tb09.160.1820	Cytochrome oxidase subunit V (COXV)	935	291	-644
Tb927.10.14000	Aconitase (ACO)	657	31	-626
Tb927.6.510	GPEET2 procyclin precursor	3848	365	-3483
Tb927.10.10260	EP1 procyclin (EP1)	4682	34	-4648
Tb927.10.10220	Procyclin-associated gene 2 (PAG2) protein (PAG2)	6903	5375	-1528
Tb927.5.330	Receptor-type adenylate cyclase GRESAG 4, putative	3779	809	-2970

List of genes with the highest change in translational efficiency between PF and BF. The list includes only those genes to which a putative function has been assigned.

BF and PF differ not only in the protein that coats their surface but also in the metabolic pathways used to generate energy. Whereas BF uses exclusively the glycolytic pathway, PF depends on oxidative phosphorylation (Millerioux et al., 2013). For these reasons, we performed an analysis of the enzymes involved in energy metabolism and the surface proteins to validate the differential translational efficiency results. Aconitase, the cytochrome oxidases V and VI and procyclins presented an increased translational efficiency in the PF (Table 3), which confirm previous observations (Furger et al., 1997; Mayho et al., 2006; Saas et al., 2000) and validate our results. In addition, certain groups of proteins involved in glucose metabolism showed translational efficiencies that corresponded to the changes reported for the glucose metabolism (Figure 24A).

Other results were unexpected, such as the efficient translation of some members of the PUF (Pumilio) proteins in PF (Figure 24A). This family of RNA binding proteins is able to regulate mRNA stability and translation and therefore control gene expression (Archer et al., 2009; Miao et al., 2010; Wharton and Aggarwal, 2006). Additionally, some isoforms of proteins involved in the control of translation initiation such as poly(A) binding proteins, eIF4Es and eIF4G were more efficiently translated in the PF than in the BF.

According to the results presented above, it is valid to suggest that there are considerable differences in translational efficiency between PF and BF and that they may contribute to the control of gene expression in *T. brucei*.

#### **uORFs may regulate translational efficiency**

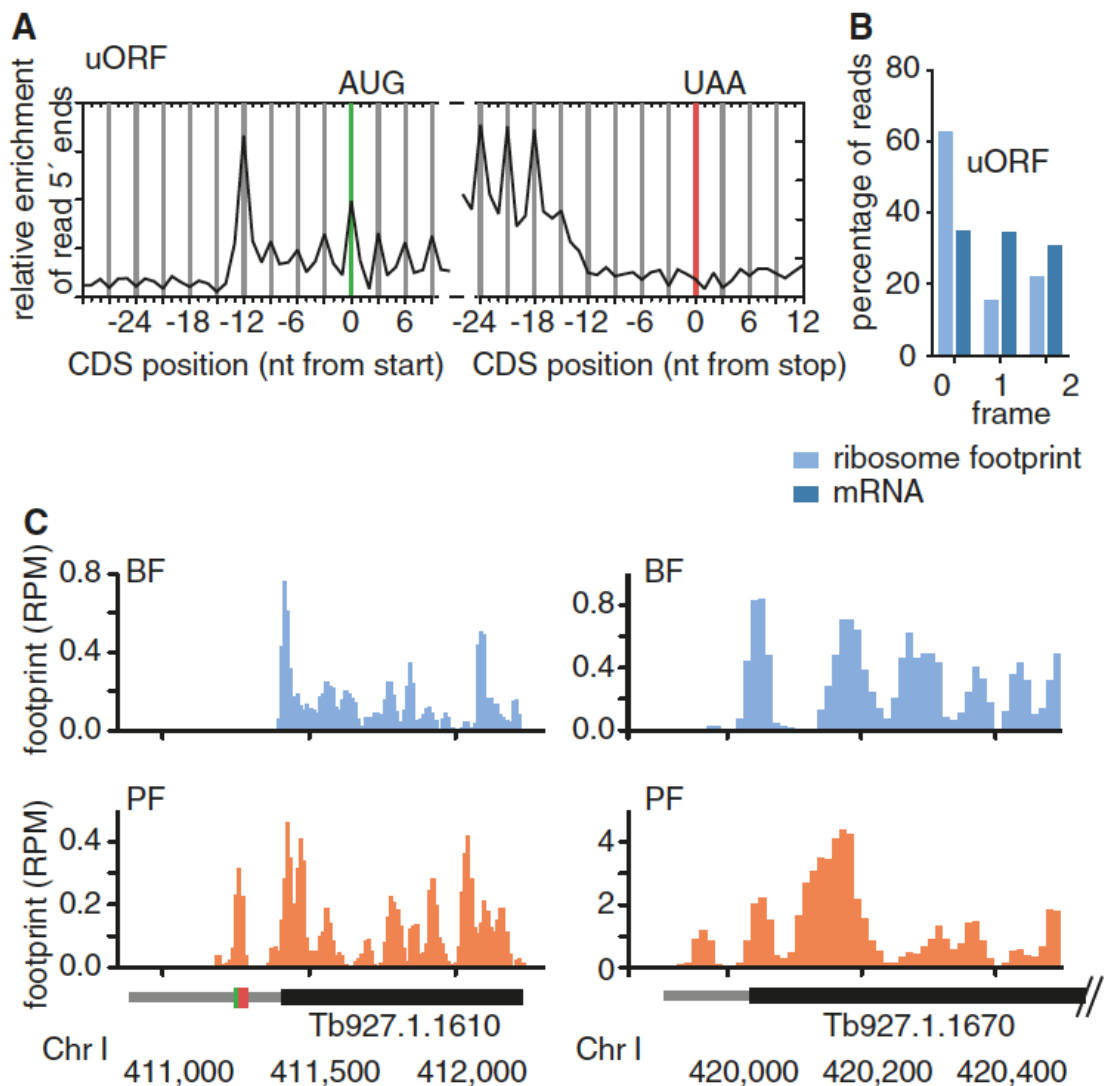
The main mechanisms described for translational regulation include those that control translation initiation, elongation, termination, uORFs, codon usage and start codon context (Jackson et al., 2010; Somers et al., 2013). Sequence motifs in the 3'-UTR affecting the translational efficiency have been characterized only for a few genes in *T. brucei*, such as those coding for procyclins (Hotz et al., 1997). However, for the majority of genes these motifs have not been described as well as the other proposed mechanisms of translational regulation. As mentioned above, uORFs can regulate gene expression in other eukaryotes (Wethmar et al., 2014). For example, the analysis of mRNA and protein levels from more than 10,000 mammalian genes revealed a strong correlation between the presence of an uORF and a significant reduction in expression from the downstream CDS. Furthermore, it has been estimated

using dual-luciferase reporter systems that the presence of a uORF results in a decrease in gene expression of 30–80% with only a minor reduction of mRNA levels (Calvo et al., 2009). In *T. brucei* it has been shown that the removal of a uATG leads to a 7-fold increase in protein levels of a luciferase reporter gene (Siegel et al., 2005). In our study, the terms ORF and uORF referred to regions of DNA sequences beginning with an ATG and ending with a termination codon that may or may not encode for proteins. In turn, the term CDS refers to ORFs that encode proteins. In addition, here we used the definition for uORFs employed by Hood *et al.* (Hood et al., 2009): sequences consisting of 9-nt, containing an upstream start codon (uATG), at least an additional sense codon and a termination codon that may be located downstream of the start codon of the main CDS.

By the time our dataset was analyzed, 5'-UTRs had only been assigned for 4909 genes of the TREU927 genome (version 4.2). Among those 4,909 genes, we were able to identify a total of 8,310 uORFs and 1,092 (22%) 5'-UTRs containing at least one uORF. The proportion of genes containing uORFs in *T. brucei* is lower compared to the numbers found for example in mammals, where 40–50% of genes contain uORF (Iacono et al., 2005; Matsui et al., 2007).

To determine the degree of translation from uORFs, the footprint density across uORFs was calculated, which allowed the identification of 1,834 uORFs (22%) with  $\geq 2x$  read-coverage and  $\geq 70\%$  of the ORF covered. The average length of these uORFs was 22 amino acids. The mapping of ribosome profiling reads across these uORFs revealed a distinct 3-nt periodicity, with 63% of BF ribosome profiling reads starting at the first nucleotide of a codon (Figure 25A

and B). This percentage is slightly lower compared to the one found for CDSs (68%) but still suggest the involvement of uORFs in translation.



**Figure 25.** Characteristic 3-nt periodicity of ribosome profiling footprints in uORF. **(A)** 5'-nucleotides from ribosome footprint reads aligned close to translation start or translation termination sites of uORFs. **(B)** Percentage of reads according to their position relative to reading frame. **(C)** Ribosome profiles of two genes with uORF (left panel) and without uORF (right panel) in BF and PF. 5'-UTRs are represented as narrow grey boxes, AUG-codons as green lines and termination codon as red lines.

Then, we were interested in studying the regulatory potential of uORFs. For this reason we compared the ribosome densities between transcripts

containing at least one uORF (N = 1,092) and those without uORFs (N = 3,817). In the BF, median ribosome density across the CDSs of transcripts with uORFs was 30.31 rpkm compared with 44.86 rpkm for genes without uORFs (Table 4). In the PF, median ribosome densities were 20.01 rpkm (genes with uORF) and 36.50 rpkm (genes without uORF). These results suggest that in both BF and PF, a higher ribosome density for genes without uORFs was observed than for genes with uORFs ( $P < 0.0001$ ). The mRNA level for genes without uORFs compared with genes with uORFs was higher ( $P < 0.0001$ ) but the differences in mRNA levels were lower than the differences in ribosome density (Table 4). These results indicate that the median translational efficiency is higher for genes not preceded by an uORF, which suggests that the presence of uORFs may decrease translation of downstream CDSs.

**Table 4.** Translational efficiency of genes preceded or not by uORF

	Transcripts with uORF (N = 3817)	Transcripts without uORF (N = 1092)	Mann-Whitney test
Bloodstream form			
Ribosome density (median rpkm)	30.31	44.66	$P < 0.0001$
mRNA levels (median rpkm)	30.83	36.13	$P < 0.0001$
Translational efficiency (ribosome density mRNA levels)	1.00	1.27	$P < 0.0001$
Procyclic form			
Ribosome density (median rpkm)	20.01	36.50	$P < 0.0001$
mRNA levels (median rpkm)	24.04	30.38	$P < 0.0001$
Translational efficiency (ribosome density mRNA levels)	0.87	1.26	$P < 0.0001$

Number of genes for which annotated 5'-UTR are available, N=4,909. Reads mapping within the first 40-nt of a CDS were not considered when calculating translational efficiency.

Remarkably, while median ribosome density for CDSs was higher in the BF than in the PF (Table 4), the opposite was observed for 5'-UTRs ( $P < 0.0001$ ). For 5'-UTRs in PF the ribosome density was higher (median rpkm: 7.73) than in the BF (median rpkm: 2.89) (See an example in figure Figure 25C). Additionally, 5'-UTRs with uORF (4.08 rpkm) in BF have a higher ribosome density than 5'-

UTRs without uORFs (2.54 rpkm;  $P < 0.0001$ ). The opposite was found in the PF, where 5' UTRs without uORF (8.34 rpkm) have a higher ribosome density than 5' UTRs with uORF (5.01 rpkm;  $P < 0.0001$ ).

Non-canonical or non-AUG uORFs were found in this analysis as well (Figure 25C, right panel). These kind of uORFs are common in yeast and mice embryonic stem cells (mESCs) and an increase in 5'-UTR translation was observed in yeast on starvation and a decrease in 5'-UTR translation was observed on differentiation of pluripotent mESCs into embryoid bodies (Ingolia et al., 2009, 2011). Taking into account the results shown above, it is interesting to speculate that a similar mechanism can be operating in the two life cycle stages analyzed.

The scope of this part of my thesis was not to prove or validate the translation of uORFs or their involvement in the regulation of the translational efficiency of the main CDS. But the existence of multiple 5'-UTR isoforms resulting from differential trans-splicing in several genes (Helm et al., 2008; Nilsson et al., 2010; Siegel et al., 2010) raises the possibility that the presence of uORFs may represent a regulatory mechanism to modulate translational efficiency in the parasite.

### **Ribosome footprints allow the identification of previously un-annotated CDSs**

The importance of small proteins (<200 amino acids) in plant, animal and even *T. brucei* physiology has become evident recently (Camby, 2006; Ericson et al., 2014; Fletcher, 1999). The data analysis presented here showed that the



great majority of ribosome footprints aligned to annotated features, but 101,197 reads (0.23%) in BF and 1,116,030 reads (1.44%) in the PF did not. These reads might correspond un-annotated UTRs or CDSs. Given that ribosome profiling captures the average ribosome positions at nucleotide resolution, it is possible to address if these reads reveal un-annotated CDSs. Previously, an RNA-seq-based analysis detected 1,011 un-annotated transcripts in the PF with the potential to encode one or more peptides ( $\geq 25$  amino acids) (Kolev et al., 2010). The authors confirmed translation for 19 of the 1,011 transcripts after comparing their results with the available proteomics dataset (Kolev et al., 2010).

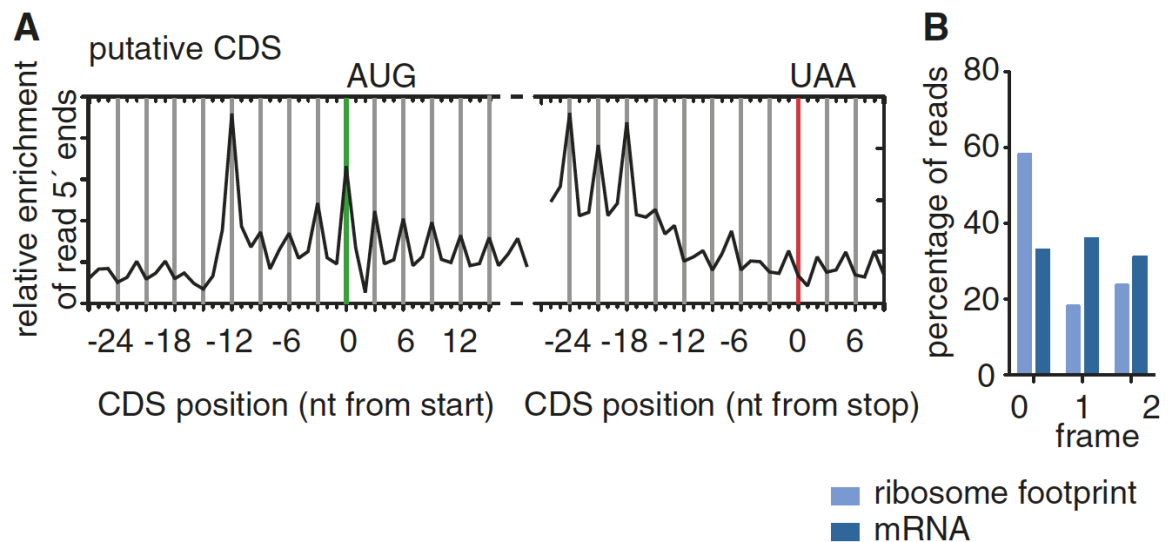
In order to perform a similar approach with the dataset generated, we determined the ribosome footprint density for all potential CDSs at least 10 amino acids in length from both life cycle stages. In this analysis we only considered ORFs located at least 20-nt away from annotated features to avoid overlapping with annotated CDSs. With this approach we were able to identify 2,021 candidate CDSs with  $\geq 2x$  read-coverage and  $\geq 70\%$  of the ORF covered. The putative CDSs ranged in length from 10 to 378 amino acids with an average of 30 amino acids. The average ribosome footprint read-coverage was  $\geq 2x$  for 797 of the 1011 previously identified transcripts.

Next, we compared the results with a published proteomic dataset (Butter et al., 2013) to search for protein products and we were able to find hits for 24 of the 2,021 candidate CDSs (average size 117 amino acids, 13 kDa) although 4 of these identified peptides also matched annotated CDSs. In addition, it was possible to identify protein products for 31 uORFs. The proteomics experimental set-up applied by Butter *et al.* may explain the low number of hits. For example,

proteins smaller than 5–10 kDa typically run out of the 1D-SDS-PAGE gels used to separate them, which preclude the possibility to detect small peptides. In our analysis 67 (3%) of the candidate CDSs were  $\geq 90$  amino acid ( $\geq 10$  kDa) and for 28% of those large candidate CDSs we were able to identify peptides. Additionally, the lower sensitivity of mass spectrometric analyses compared with DNA sequencing-based techniques is well known (Altelaar et al., 2013). It is also possible that some of the reads obtained in our analysis co-sediment with the monosome fraction but are not involved in translation as has been shown in mice (Ingolia et al., 2011).

We use the so-called Ribosome Release Score (RRS) to evaluate the coding potential of a transcript and to separate lncRNA from small coding genes (Guttman et al., 2013). The RRS is based on the fact that translating ribosomes are released once they reach a stop codon. This release results in an abrupt decrease in ribosome occupancy between CDS-regions and the subsequent 3'-UTR. In consequence, the RRS was defined as the ratio of footprint reads in the putative CDS to footprint reads in the corresponding 3'-UTR divided by the ratio of RNA reads in the putative CDS to RNA reads in the corresponding 3'-UTR.

The averaged ribosome footprint data indicated a clear 3-nt periodicity for sequence reads occurring near translation initiation sites but a less well-defined drop across translation termination sites (Figure 26A and B), which may indicate a lack of productive translation.



**Figure 26.** Ribosome footprints allows the identification of un-annotated CDSs. (A) Alignment of the 5'-nucleotides from ribosome footprint reads that map close to translation start or translation termination sites of putative CDSs. (B) Percentage of reads according to their position relative to reading frame.

Candidate CDSs were defined as ORFs with a minimum length of 10 amino acids and average ribosome footprint coverage of 2x over at least 70% of the ORF. The 3'-UTR was defined as the region beginning immediately downstream of the ORF and ending at the first subsequent start codon in any reading frame (Guttman et al., 2013). RRS can be calculated for an ORF if at least one entire footprint read and one RNA read maps to the putative CDS and the corresponding 3'-UTR. Because these criteria were not fulfilled for several 3'-UTRs, the RRS was determined for only 1,445 of genes annotated in *T. brucei*. For these 1,445 genes, the median RRS was 83.9 compared with an RRS of 0.19 for the subset of genes annotated as hypothetical unlikely (N = 90). Then, we determined the RRS for the un-annotated candidate CDSs identified in this study and that fulfilled the criteria for RRS calculation (N = 265). The median RRS of these candidate CDSs was 79.91, 3.99 lower than the median RRS of the

annotated genes (83.9). Thirteen putative CDSs had an RRS higher than 83.9 and 98 had an RRS larger than the intron-containing poly(A) polymerase (RRS = 9.63). We were able to calculate the RRS only for one of the candidate CDSs with an identified protein product (RRS = 57.44).

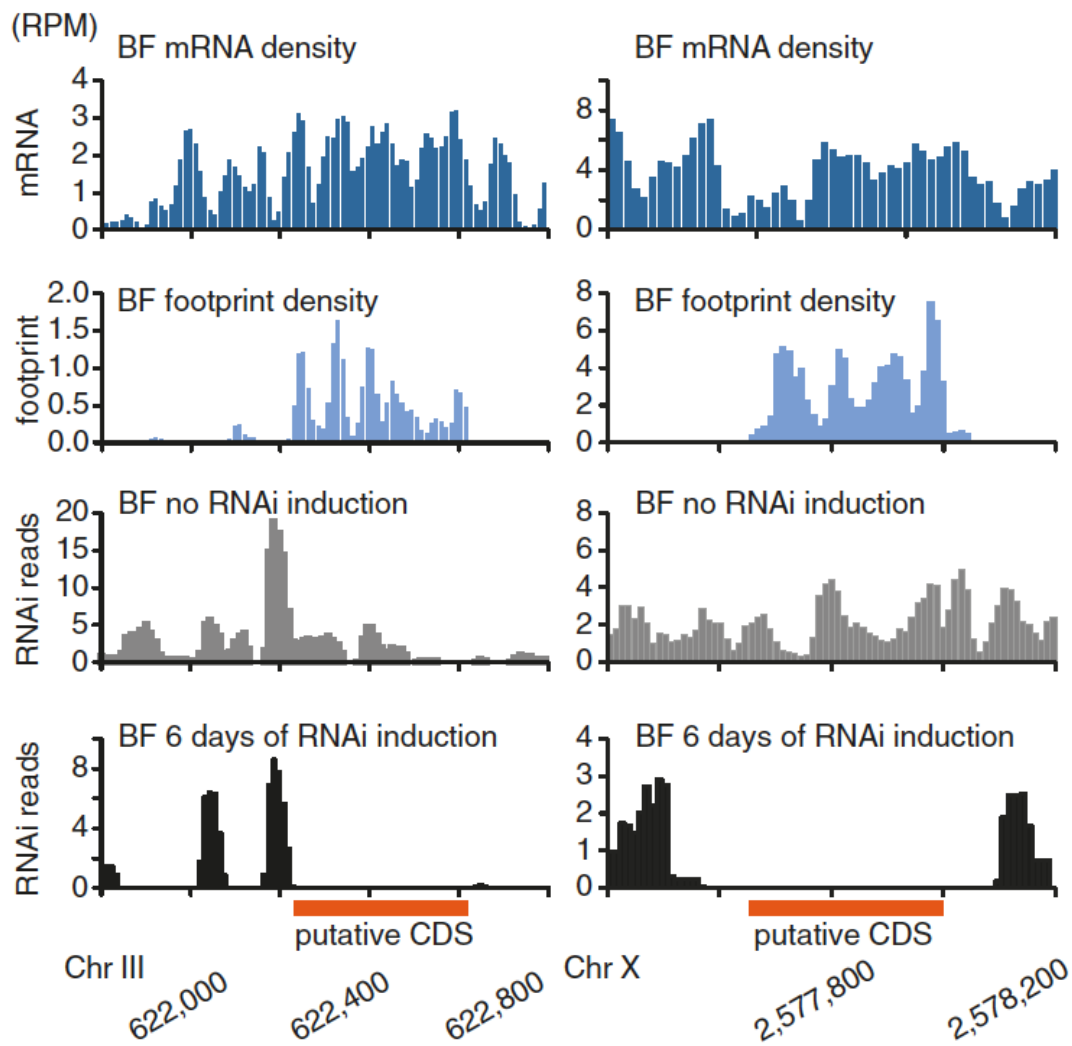
Finally, of all the candidates with an identified protein product or an RRS above 10, only one could be linked to a function using InterProScan. The candidate was located within the histone H2B gene array and contained an H2B signature motif.

Taken together the results from the proteomic and RRS analyses suggest that at least a few previously un-annotated CDSs can be translated in *T. brucei*.

### **Ribosome profiling newly identified CDSs may be important for parasite fitness**

ORFs with a low RRS and without mass spectrometric evidence may not be translated into functional proteins but may play regulatory roles such as ncRNAs. Therefore, the biological significance for parasite survival of the new putative CDSs and ORFs was determined using an approach different than RRS and mass spectrometric evidence. A previously published high-throughput phenotyping approach, termed RNA interference target sequencing (RIT-seq), measured the fitness–cost associations using RNAi (Alsford et al., 2011). RIT-seq is based on the transfection of the parasites with RNAi libraries and the comparison of the recovery of RNAi targets from populations before and after RNAi induction. The RNAi libraries were generated using genomic DNA and the effect on parasite fitness was only determined for annotated genes. The RIT-seq

data revealed a significant loss in fitness upon RNAi induction for 2,724 CDSs in the BF, 1,972 CDSs in the PF and 2,677 CDSs in differentiated parasites. Then, we used this same dataset to compare it with our ribosome profiling results. We found a significant loss of fitness upon RNAi induction for 214 candidate CDSs in the BF (6 days after RNAi induction), 16 CDSs in the PF and 227 CDSs upon differentiation (See an example in Figure 27). This analysis suggests that the previously un-annotated CDSs found with the ribosome profiling approach, may have an important biological role for the survival of the parasite specially the form that is infective to mammals.



**Figure 27.** Ribosome profiling and RIT-seq profiles of two previously un-annotated putative CDSs in BF. From top to bottom, the first panel corresponds to RNA-seq data, the second one to ribosome profiling footprints and the third and fourth ones to RIT-seq before and 6 days upon RNAi induction respectively. The orange box corresponds to the putative CDS. Observe the clear absence of reads upon RNAi induction in the bottom panel.

## Discussion

Eukaryotic gene expression is regulated at multiple levels and translation of mRNA into proteins has been shown to be one of the most important (Moore, 2005; Schwanhäusser et al., 2011). In this chapter of my thesis I report the first genome-wide analysis of translational regulation and strand-specific analysis of

RNA transcript levels for *T. brucei*, for a eukaryotic pathogen and for an organism without transcriptional control.

The ribosome profiling approach used here allowed us to determine simultaneously the translome and transcriptome of BF and PF and to evaluate the translational efficiency of individual transcripts. The results obtained suggested this step is important in gene expression control as has been shown in other eukaryotes. The study reveals large differences in translational efficiency among transcripts in the same life cycle stage and for the same transcript in different stages. Translational efficiency varied over two orders of magnitude, which is similar to the value measured for RNA stability (Manful et al., 2009) and emphasizes its importance for gene regulation. However, no correlation between RNA abundance and translational efficiency was found, which suggests that translational efficiency is regulated independently of RNA stability. This range of variation in translational efficiency is similar to what has been observed in yeast and approximately 10-fold higher than what has been described for mice (Ingolia et al., 2009, 2011).

In mouse embryonic stem cells the rate of translation has been found to be similar between different transcripts, the kinetics of elongation independent of length and protein abundance and the speed of translation independent of codon usage (Dana and Tuller, 2012; Ingolia et al., 2011). These measurements suggest that the speed of translation is approximately constant and for this reason it is valid to use ribosome density to estimate the rate of protein synthesis. Nevertheless, in *T. brucei* the speed of translation may be different between the life cycle stages that live at 37° C and 27° C as is the case for BF and

PF respectively. In addition, our current protocol does not permit the measurement of the absolute mRNA and footprint abundance, which precludes the possibility of comparing translation efficiencies for individual transcripts. One way to obtain more accurate numbers is to include synthetic RNAs as spike-in controls, covering a concentration range optimal to derive standard curves for quantifying the abundance of transcripts and footprints (Jiang et al., 2011). Therefore, we did not compare translational efficiency directly but instead we determined the 'rank' in translational efficiency for all genes in both BF and PF. The logic behind this approach is based in the assumption that a difference in rank indicates life cycle-dependent translational control. We observed a general positive correlation in translational efficiency between the two life cycle stages (Pearson  $r = 0.7428$ ), but for a subset of genes there was developmental regulation of translation. The observed regulation agrees well with the information available for some proteins such as cytochrome oxidases and procyclins in the PF and enzymes important for glycolysis and VSGs in the BF. These results are consistent with differences in energy metabolism and the protein surface coat of the different life cycles stages. Interestingly, another subset of genes includes proteins previously unknown to be developmentally regulated like members of the Pumilio family and translation initiation factor isoforms.

In other organisms, one of the contributors of translational regulation is the presence of a uORF that generally correlates with reduced gene expression (Wethmar et al., 2014). However, for some genes, like the transcription factor GCN4 in yeast and the activating transcription factor ATF4 in mammals, this



correlation is reversed on stress (Hinnebusch, 1993; Vattem and Wek, 2004). Thus, uORFs can exert positive and negative effects on translation of downstream CDSs. In this study, the median translational efficiency was higher for genes without uORF than for genes with uORF, which suggests that uORFs may be important contributors of translational regulation. The ribosome profiling data indicate as well an increase in translation from 5'-UTRs and a more ubiquitous translation initiation from non-AUG sites in the PF compared with the BF. A similar increase in the translation from 5'-UTRs has been observed in yeast upon starvation (Ingolia et al., 2009). One of the most widespread stress-induced mechanisms to regulate the initiation of translation requires the inactivation of the translation initiation factor eIF2 $\alpha$  by phosphorylation at a conserved serine (Dever et al., 1992; Kong and Lasko, 2012). This initiation factor has been involved as well in determining the stringency in initiator codon selection (Huang et al., 1997). Therefore, it would be interesting to determine whether eIF2 $\alpha$  phosphorylation at the corresponding serine plays a role in the selection of initiator codons and in the regulation of translation in *T. brucei*.

We combined ribosome profiling, proteomics and genome-wide RIT-seq data to identify novel genes. Comparison with the proteomics dataset allowed the identification of protein products for 20 relatively large candidate CDSs. Comparison with the RIT-seq dataset enabled the identification of >200 putative CDSs that appear to be essential for parasite fitness. Furthermore, RIT-seq data suggested that candidate CDSs are 13-fold less likely to be essential for viability in the PF than in the BF or during differentiation of cells from BF to PF. These findings point to a possible role of small peptides in the adaptation of

trypanosomes for the survival in the mammalian host. Therefore, our data suggest the presence of a set of small proteins and that may be important for the fitness of the parasite. These small proteins have not been intensively explored, but their further characterization will have to include a more targeted analyses.

The results obtained in this part of my work may be useful for the *T. brucei* scientific community and may open as well several interesting possibilities for future research. First of all, they will allow to unambiguously identify the start codon of a given CDS in *in vitro* cultured PF and BF. Second, the set of new putative CDSs identified here will help to complement the current annotation of the *T. brucei* genome. Third, ribosome profiling allows the identification of putative ncRNAs that may play important roles in the physiology of the parasite, such as the CTR transcript I have characterized and which is the topic of the next chapter of my thesis.

However, the most relevant finding of this part of my thesis is the importance of translational regulation between life cycle stages and between genes. The importance of this step in gene expression regulation was expected given its relevance in other eukaryotic organisms and the apparent lack of regulation as the level of transcription initiation in the parasite. However, few reports had addressed this issue before we published the first genome-wide analysis of translational regulation in *T. brucei*. Ribosome profiling libraries can be generated to study other life-cycle stages or PF and BF under some kind of stress for example parasites obtained from infected tsetse flies or mice. Furthermore, I think it may also be interesting to study the mechanism(s) of translational regulation in the parasite. For example, UTRs and uORF play a role

at this level of regulation in other eukaryotes and with the implementation of editing tools like the CRISPR-Cas9 system in *T. brucei* it is now possible to precisely modify these regions and evaluate their impact in the translational efficiency of the main CDS.

Taken together, all the findings mentioned in this chapter of my thesis highlight the power of ribosome profiling to provide a new picture of the translational landscape in the two life cycle stages analyzed and paving the way for new studies that complement what we already know about gene regulation at the level of RNA stability and trans-splicing efficiency.

## **Chapter 4. Identification and characterization of a ncRNA transcribed from the CTR in the active BES of the Lister 427 strain**

### **Introduction**

In the previous chapter I described the identification of incorrectly annotated translation initiation sites and of previously un-annotated CDSs. Twenty four of these putative new CDSs were validated using a published proteomic dataset. Additionally, RIT-seq allowed us to identify around 200 new ORFs with potential relevance for the parasite fitness. However, transcriptome analyses in human cells have revealed genome-wide transcription of RNA mostly considered to be non coding (International Human Genome Sequencing Consortium, 2004) and different transcriptome analyses have shown that an important fraction corresponds to ncRNAs (Berretta and Morillon, 2009). At present, it is known that many ncRNAs have a biological function and intensive research has led to the identification of regulatory mechanisms controlled by these transcripts. One challenge for the identification of ncRNAs is that many of them contain ORFs making it difficult to unambiguously classify the transcript as non-protein-coding (Spicuglia et al., 2013). In addition, lncRNAs containing ORFs longer than 100 amino acids have been found associated with ribosomes but not necessarily in translation (Guttman et al., 2013; Ingolia et al., 2009). These two characteristics complicate the characterization and annotation of CDSs and ncRNAs, emphasizing the importance to analyze transcripts individually and to develop new approaches for their correct annotation.

Ribosome profiling is one of the approaches that can help to elucidate the translational potential of a given ORF. Basically, the presence of footprint reads for a specific ORF would suggest its translation into protein. According to our ribosome profiling analysis, we observed transcription for many annotated CDSs but not ribosome footprints. This may be explained in two ways: they may correspond to CDSs that are either translated in other life cycle stages not analyzed in our study or translated during stress conditions (i.e. in the bloodstream or the midgut). On the other hand they may correspond to ncRNAs and for that reason they just might be categorized as ORFs and not CDSs.

In addition, features transcribed abundantly or moderately and that cannot be categorized as ORFs may correspond as well to ncRNAs. Therefore, another advantage that ribosome profiling offers is the use of the transcriptomic library to estimate the relative abundance of a ncRNAs. This can be done by comparing the number of reads obtained for a given putative ncRNA with the number of reads from other well-described genes. In the Figure 28 you can see an example after comparing the number of reads between the CTR and the *VSG*.

Given the importance of ncRNAs regulating mechanisms in different prokaryotic and eukaryotic organisms, I was interested in investigating the role of these transcripts in the physiology of *T. brucei*. Specifically, in 1997 Davies *et al.* observed that the deletion of the CTR, the regions just upstream of a *VSG*, led to a 100-fold increase in the *in situ* switching rate. Therefore, the authors suggest the presence of a stabilizing element in the CTR of the active BES. Discussing their results, the authors hypothesize the presence of a putative transcript or transcripts in this region that can play the role as stabilizing elements (Davies *et al.*, 1997). Transcripts from the regions upstream of *VSGs* were identified around

30 years ago (Scholler et al., 1988), and their transcriptional and translational potential was evaluated in a couple of studies in the late 1980s (Aline et al., 1989; Scholler et al., 1988). With the availability of modern methodologies such as ribosome profiling, I was able to describe the transcriptional and translational potential of the CTR.

Taking into account the potential role of a putative CTR transcript in *in situ* switching, it is relevant to re-examine the importance of this mechanism for the parasite survival. *T. brucei* avoids elimination by the host immune response by using a mechanism of antigenic variation, changing the composition of its surface coat that is composed of around 10 million copies of VSGs (Cross, 1975). In Lister 427, the expression of a VSG gene occurs from only one of 15 BESs that is actively transcribed at a time while the others are transcriptionally repressed (Hertz-Fowler et al., 2008). In addition, the *T. brucei* genome contains ~2,500 VSG genes and pseudogenes located at chromosome-internal sites and minichromosomes (Cross et al., 2014). In the parasite, antigenic variation is generated by genetic and epigenetic mechanisms (Horn, 2014). Among the genetic mechanisms we can find recombination events like duplicative gene conversion, segmental gene conversion and telomere exchange (Rudenko, 2010). The epigenetic mechanisms include the activation of a previously inactivated BESs and the corresponding silencing of the active one (*in situ* switching). Other peculiarities of the active BES are the decreased nucleosome content (Figueiredo and Cross, 2010; Stanne and Rudenko, 2010) and the fact that depletion of the chromatin-associated proteins or chaperons ASF, CAF, Dot1B, FACT and ISWI induce defects in the mutually exclusive expression of BESs (Alsford and Horn, 2012; Denninger and Rudenko, 2014; Figueiredo et al., 2008; Hughes et al., 2007;

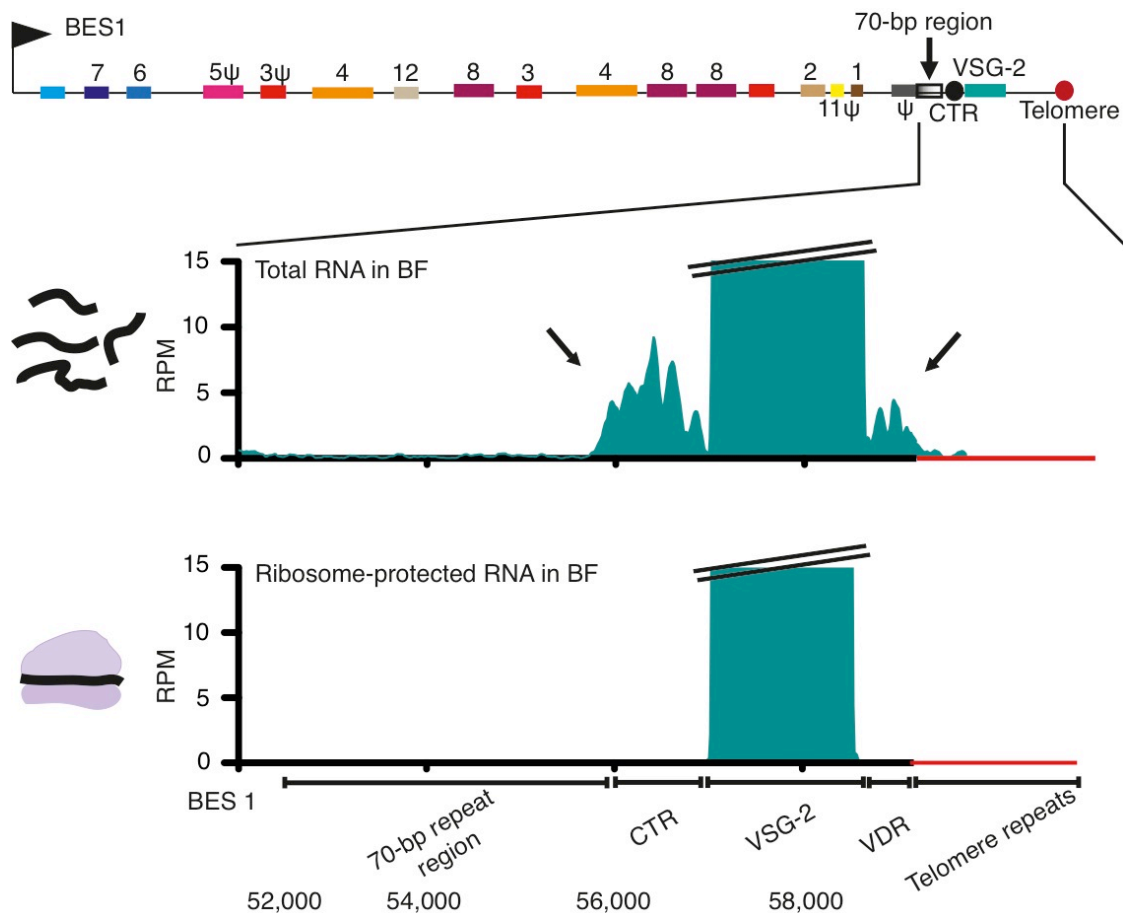
Janzen et al., 2006a; Stanne et al., 2011). The deletion of the CTR reported by Davis *et al.* produced an increase in the *in situ* switching rate and not in the mutually exclusive expression of BESs, the factor or factors controlling the *in situ* switching are not currently known and the role of the CTR in *VSG* switching has not been further investigated.

The main goal of this project is the identification of the stabilizing element present in the CTR of the active BES in Lister 427. As part of my thesis, I was interested in determine the number of transcripts transcribed from the CTR, the length of each of them and to identify the sequence in both the 5'- and the 3'-end of the transcripts. The information generated in this characterization will be used in the experimental design for the deletion of the CTR and in the resulting measurement of the *in situ* switching rate.

## **Results**

### **CTRs are transcribed into long non-coding RNAs (lncRNAs)**

After mapping the RNA-seq and ribosome-profiling data to the Lister 427 genome, I found that the CTR of the active BES in BF was moderately transcribed but not translated, which suggest the presence of ncRNAs (Figure 28). In addition, this analysis showed transcription in the region downstream of the *VSG* (region that I have called *VSG* downstream region [VDR]). Nevertheless, I did not perform the primary characterization of this region, which waits for the elucidation of whether is transcribed into one long or multiple short ncRNA transcripts.

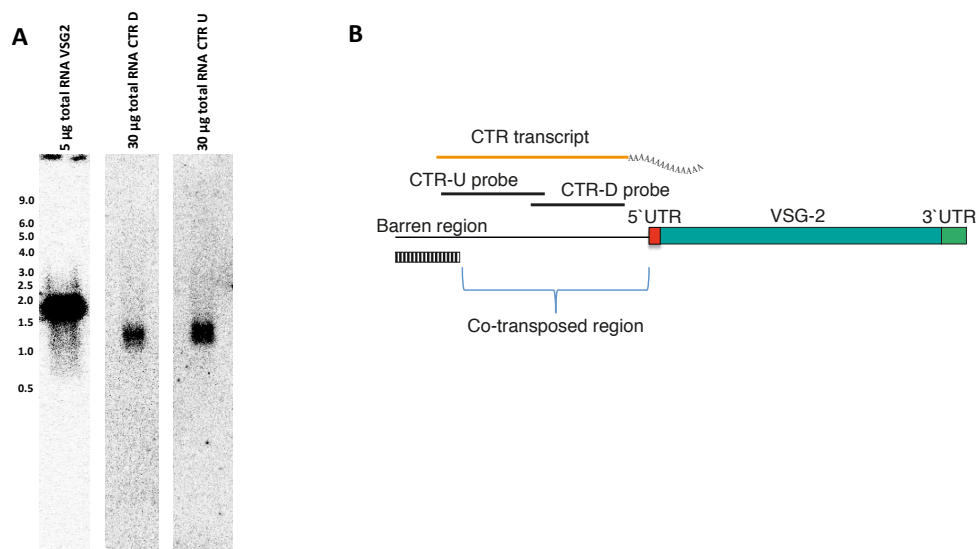


**Figure 28.** Representation of the RNA-seq and ribosome profiling data for the region encompassing the 70-bp repeat region and the telomere repeats in the BES1 of Lister 427. The top panel represents the entire BES1 (adapted from Hertz-Fowler *et al.* 2008) and the bottom ones are a zoom up of the region of interest in this study.  $\psi$  corresponds to pseudogenes. Black arrows in the RNA-seq panel indicate the moderate transcription in the CTR and the region downstream of the 3'-UTR of the *VSG-2* (VDR). The parallel diagonal bars indicate that the totality of reads for *VSG-2* is not shown.

The boundaries of the DNA sequence for the CTR comprise the region between the 70-bp repeat region and the SAS of *VSG-221* (also known as *VSG-2*), which corresponds to 1,072 nucleotides. The next step after analyzing the ribo-seq data was to perform a northern blot to determine the size and the number of transcripts produced from the CTR. Using two probes that spanned the entire



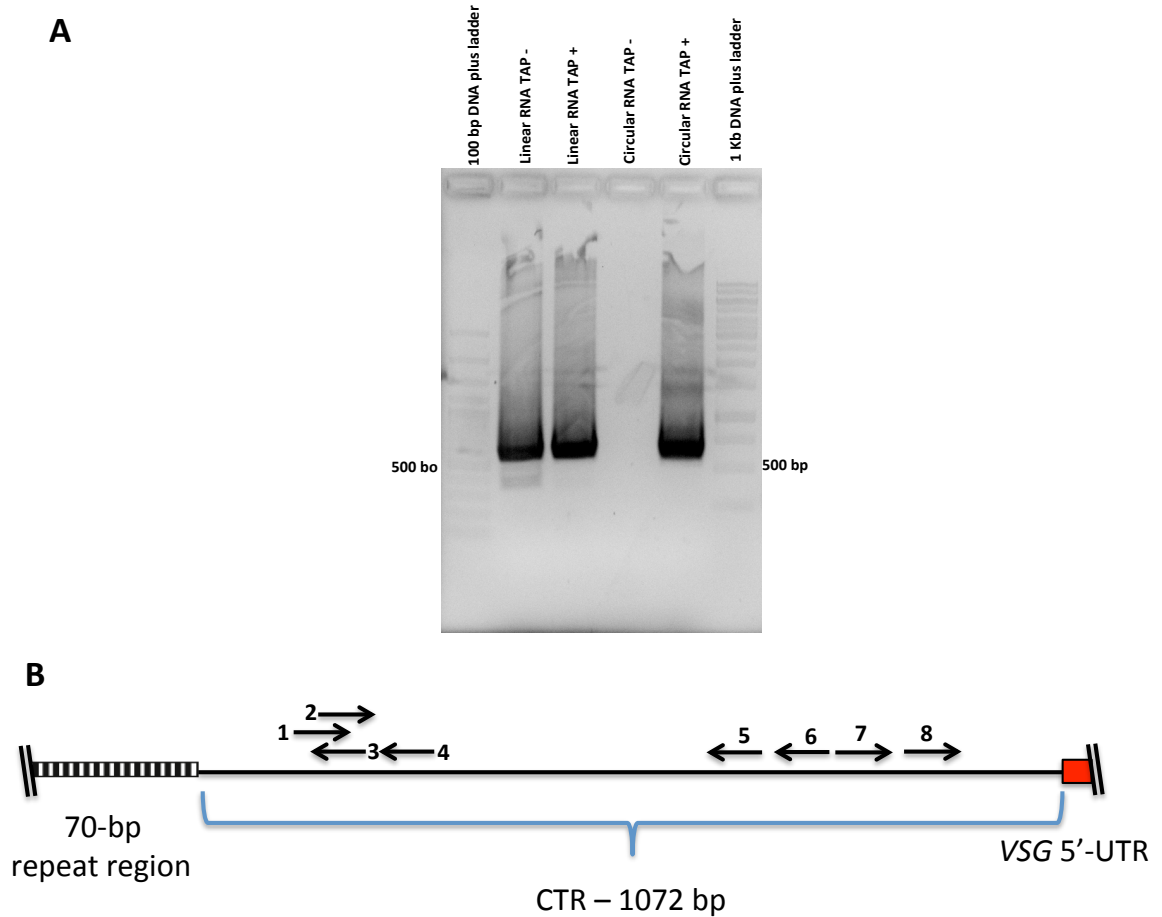
region of interest, I found only one transcript species of ~ 1,200-nt in length (Figure 29A).



**Figure 29.** Characterization of the ncRNA transcribed from the CTR in the active BES of *Listeria* 427. **(A)** Northern blot using total RNA from BF and one probe for *VSG-2* and two probes for CTR (CTR-U and CTR-D). The length of the mature *VSG-2* transcript is ~ 1,540-nt. Both probes for the CTR hybridize to a transcript of the same size ~ 1,200-nt. **(B)** Schematic representation of the CTR transcript and its genomic context. Circular RACE PCR suggested extensive polyadenylation at the 3'-end and the absence of the SL at the 5'-end.

The relative abundance of the CTR transcript was lower compared to the highly abundant *VSG* mRNA, which is in agreement with what was observed in the RNA-seq data (Figure 28). Furthermore, the signal obtained with the CTR-D probe was lower compared to the signal from CTR-U, which can be explained given the differences in length for the two probes (805-nt for CTR-U and 514-nt for CTR-D). The length of the detected transcript, ~ 1,200-nt, enables its classification as a lncRNA. The presence of a lncRNA transcript was confirmed by circular RACE experiments that allowed me to map its 5'- and 3'- boundaries. I analyzed samples treated or not with TAP (TAP cleaves the pyrophosphate bond

of the 5'-terminal methylated guanine nucleotide "cap" of eukaryotic messenger RNAs) or RNA 5'-Polyphosphatase (which sequentially removes the  $\gamma$  and  $\beta$  phosphates from 5'-triphosphorylated RNA [primary RNA transcripts] and 5'-diphosphorylated RNA). After running the nested PCR products on an agarose gel, a band was observed only in the circularized sample treated with TAP, which implies the presence of a cap at the 5'-end of the transcript (Figure 30). No bands were observed for the samples treated with RNA 5'-Polyphosphatase (instead of TAP) or without any of the two enzymes, which indicates the capped nature of the CTR transcript.



**Figure 30.** The 5'-end of the CTR transcript is capped. A. cDNA obtained from linear or circular RNA was used as a template for two rounds of PCR using nested primers to increase the specificity. A 1% agarose gel using 1x TAE was run at 90 V for 1 hour. The gel was stained with ethidium bromide for 10 minutes. The samples in which the cDNA was obtained from linear RNA served as a control for the presence of the CTR transcript. B. Schematic representation of the CTR showing the region of annealing for the primers used in the first PCR and the nested PCR. 1: 617\_3race\_F2, 2: 618\_3race\_F2\_n, 3: 1012\_CTRRDn, 4: 1011\_CTRRD, 5: 718\_CTR\_R3\_n, 6: 420\_CTR\_R3, 7: 1010\_CTRFDn and 8: 1009\_CTRFD.

Sanger sequencing of the RACE PCR product from the TAP+ treated RNA confirmed the presence of a unique transcript of 1,135-nt. The advantage of a circular RACE strategy resides in the identification of the sequences present at the ends of a transcript. By means of this strategy, I was able to identify a long poly(A) tail at the 3'- end but not the typical SL present at the 5'-end of the

mRNAs in *T. brucei*. These results indicated that the CTR transcript spans 130-nt in the 70-bp region and 1,005-nt in the CTR sequence. The 3'-end of the transcript is located 67-nt upstream of the annotated VSG-221 SAS (Figure 29B).

## Discussion

As part of the CTR characterization and its role in stabilizing the active BES, I was interested in answer three questions: How many transcripts were transcribed from the CTR? How long was each of them? Was the SL present at the 5'-end and the poly(A) tail at the 3'-end of them? I found only one transcript of ~ 1200 nt transcribed from the CTR and I was able to determine the exact start and end of the sequence. The transcript included a long poly(A) tail at the 3'-end but the absence of the SL at the 5'-end. This finding is surprising given the coupled nature of the trans-splicing and polyadenylation mechanisms in PTUs.

*In T. brucei*, RNA polymerase II transcribed genes (mRNA) and some RNA polymerase I transcribed genes (procyclins and VSGs) are trans-spliced and polyadenylated (Imboden et al., 1987; Johnson, 1987). Basically, in order to be efficiently processed, an immature transcript has to be trans-spliced near its 5'-UTR and at the same time a polyadenylation event has to occur on the transcript immediately upstream of the gene in an event that is temporally and spatially coupled (Matthews et al., 1994; Ullu et al., 1993). During trans-splicing, a 39-nt miniexon or spliced leader (SL) is trans spliced to the primary transcript from a capped precursor of ~140-nt. This process is necessary for the generation of the mature mRNA (together with polydenylation at the 3'-end) and the SL provides the 7-methyl guanosine (m<sup>7</sup>G) cap structure at the 5'-end of the transcript (see the reviews Clayton, 2002; Liang et al., 2003). Trans-splicing requires a GT motif

at the 5'-SAS (splice acceptor site), an adenosine branch point, a poly(Y) tract and AG at the 3'-SAS following two catalytic *trans* esterification reactions mediated by the spliceosome (reviewed in Liang et al., 2003). Polyadenylation of the 3'-end of the transcripts has been shown to occur at one of many spaced polyadenylation sites (PAS) and their structure is even more heterogeneous than trans-splicing sites (Kolev et al., 2010).

This implies that immature transcripts transcribed from the BES are processed by the coupled action of the trans-splicing and polyadenylation mechanisms described before. Therefore, immediately upstream of the SAS site on the immature *VSG* transcript has to be found the PAS of the CTR for the correct processing of the *VSG* transcript. On the other hand, the 5'-end of the transcript was identified in the 70-bp repeat region, which means that the SAS for the CTR transcript has to be located there. In addition, a PAS should be located in the same 70-bp repeat region or at the 3'-end of the pseudo*VSG* gene located upstream of the aforementioned repeated region. My results pointed that the immature CTR transcript is somehow processed at the 5'-end but not by the widespread mechanism of trans-splicing given the absence of the SL on it. To me this result was not expected, as all the genes transcribed from PTUs are trans-spliced in *T. brucei*. Nevertheless, in this parasite other ncRNAs are not trans-spliced as is the case of rRNAs (Bruderer et al., 2003) and tRNAs (Tan et al., 2002). Furthermore, snRNAs do not contain the SL but contain a trimethylated cap structure (Mottram et al., 1989; Tschudi et al., 1986). Therefore, the possibility that a not yet described mechanism is able to cap the CTR ncRNA in the absence of the SL cannot be completely ruled out.

As I mentioned in the introduction of this chapter, Davis *et al.* proposed that a transcript produced from the CTR might play a role in the increase of the switching rate. To my knowledge, the only paper that mentions the presence of transcripts from CTRs was published in 1988 (Scholler *et al.*, 1988). In that paper, the authors examined steady-state transcripts from the CTR sequences immediately upstream of the *VSG-1* and 11 genes present in the IsTaR 1 serodeme.

The CTR DNA sequence upstream of the *VSG-1* gene was of approximately 1.6 kb and at least 10.2 kb for the one upstream of the *VSG-11* gene. In both cases, they found several CTR transcripts of different length, many of them overlapping the sequence of the *VSG* gene. All the transcripts described were polyadenylated as determined by binding to oligo-d(T) cellulose. This finding agrees to what was published previously as the processing of the *VSG* precursor transcript produced polyadenylated transcripts from the CTRs (Kooter *et al.*, 1987; Pays *et al.*, 1982; Scholler *et al.*, 1988). They found the SL sequence in some of the transcripts upon performing sandwich dot blot hybridization with DNA from an M13 clone containing sequences complementary to the RNA of the CTRs. The transcripts were present at about one-tenth the abundance of the *VSG* mRNA and the 5'-end of some of them was located beyond the 70-bp repeat region.

There are two main differences between their results and mine: the number of CTR transcripts and the presence of the SL. The fact that the CTR transcripts were from different lengths and in some cases overlapped part of the *VSG* sequence could be explained because they corresponded to unprocessed mRNA before the coupled process of trans-splicing and polyadenylation

occurred. As mentioned before, Scholler *et al.* found that some but not all of the CTR transcript species presented the SL cap (Scholler *et al.*, 1988). Their experimental procedure was based in sandwich dot blot hybridization. Basically the hybrid selections were performed using DNA from an M13 clone containing sequences complementary to the CTR RNA. The probes were applied to a square nitrocellulose filter and poly(A) + RNA was hybridized. Then, a radioactive-labeled SL probe was hybridized and the selected RNA was eluted and ethanol precipitated. The results were confirmed by northern blot using the eluted RNA as a target and hybridizing again with the SL probe. It is possible that during the hybridization experiments there was some kind of cross-reactivity with trans-spliced transcripts and for this reason the SL was detected in some of the CTR transcripts in contrast to the results shown in my work. The success of this approach will depend whether the excised band only contains the CTR transcript and not other mRNAs that might hybridize with the SL probe.

Another difference consisted in that the strain used by them was different to the Lister 427 and therefore I cannot rule out that the presence of the SL is strain specific. Additionally, in that study the BF were isolated directly from irradiated rats and not from *in vitro* cultured parasites as was done in my study. In the end, it is supposed that the CTR confers a stabilizing effect to the active BES, a role that is directly related with the process of antigenic variation and the survival of the parasite in the mammalian host. Are there differences in the characteristics of the CTR transcripts between different strains or according to the infective status of the parasite? Determining the number of CTR transcripts and whether they include the SL and a poly-A tail in different strains cultivated *in*

*vitro* or obtained from infected mice may help to explain the differences observed between my study and the previous ones.

In a second paper published by the same group in 1989, the authors mentioned that the transcripts were associated with polyribosomes (Aline et al., 1989) but with limited coding potential due to their small ORFs (no longer than 49 amino acids). They performed sucrose gradients to collect the polysome fraction and northern blots to confirm the presence of the CTR transcripts. In addition, to proof that these transcripts are present in the polysomes rather than other complexes that co-sediment with polysomes, the polysome fraction from VAT 11 was treated with puromycin in order to release mRNAs from the ribosomes. Upon puromycin treatment, the majority of the CTR transcripts were released into the supernatant fraction and for this reason the authors concluded that they were present in polysomes and may be involved in translation. The results published implied that the transcripts of two *IsTat VSG* CTRs are exported from the nucleus and then recruited by the ribosomal machinery. Therefore, it is possible that some specific strains of *T. brucei* may produce CTRs transcripts capped with the SL and presenting coding potential that can have some modulatory role in the physiology of the parasite. Nowadays it is accepted the role of small proteins in the biology of several organisms including *T. brucei* (Ericson et al., 2014). Nevertheless, it is important to mention as well that the binding of ribosomes to ncRNAs has been reported in other organisms (Guttman et al., 2013; Ingolia et al., 2009; Pircher et al., 2014) but such interaction does not lead to translation of the ncRNA.

In an attempt to find a common feature with the CTRs present in the BES of other strains, they compared the coding potential of the CTR present in the



IsTat 1 *VSG* to putative proteins encoded by ORFs in the CTRs of the AnTat 1.1, IaTat 1.2, I1Tat 1.1, MITat 117, MITat 118, and MITat 221 *VSG* genes without finding any apparent conserved motif. One of the ORFs present in the CTR of the IsTat 1 *VSG* is rich in arginine and lysine residues; feature that is present in the ORFs of the CTRs in the MITat 117 and 118 *VSG* genes. But apart from this similarity, they did not find any amino acid sequence homologies, similarities in hydrophobicity, common codon biases or amino acid compositions.

Several hypotheses can be put forward to explain the nature and function of the stabilizing element contained in the CTR. One of them is that the stabilizing element resides in the DNA structure of the region. DNA coiling or a chromatin structure might interact with specific proteins that anchor the active BES to the ESB allowing a high level of gene transcription. Nevertheless, it is possible as well that the transcript characterized in this thesis confers the stabilizing effect exhibited by the CTR. In both scenarios, the deletion of this element present in the CTR should destabilize directly or indirectly the transcription of the active BES leading to *in situ* switching events.

The next step in characterizing the CTR would be to repeat the deletion performed 20 years ago by Davis *et al.* The difference is that modern technologies allow us to make marker-free genome editing even at nucleotide resolution, as is the case of the CRISPR-Cas9 system. In the next chapter of my thesis I will show the implementation of CRISPR-Cas9 in *T. brucei* PF cells.

In order to complete the characterization and establish the CTR function, the following aspects will have to be determined in the future: measuring of the *VSG* switching frequency, mapping of the precise location of the CTR element

that confers the stability and determining the interacting partners of the CTR transcript.

Laboratory-adapted strains like Lister 427 switch at a rate of  $1 \times 10^{-6}$  per generation and for this reason cultures present homogeneous VSG expression. Therefore, the classic way to measure VSG switching frequency employs the depletion of non-switched parasites followed by the counting of the remaining switched cells. Davies *et al.* for example employed immunolysis, by making use of a VSG-221-specific antibody and complement-containing serum (Davies *et al.*, 1997). Another strategy is based in the use of VSG-specific antibodies coupled to magnetic beads (Boothroyd *et al.*, 2009). In other instances a VSG switching reporter strain was generated to score different switching mechanisms (Kim and Cross, 2010). In the latter strategy, the parental strain is doubly marked with a blasticidin-resistance gene (BSD) downstream of the promoter and PUR-TK (puromycin resistant gene-thymidine kinase) at the 3'-end of the 70-bp repeat region. In the case of an *in situ* switching event, the cells become sensitive to BSD and resistant to ganciclovir (Kim and Cross, 2010). The advantage of the strategies based in VSG-specific antibodies consists in that they do not need to modify the active BES and the concomitant drug selection process. The variability in the quality of the complement-containing serum due to batch-to-batch differences or suppliers is the main disadvantage of the immunolysis-based method. Probably the best method to deplete the non-switched parasites is the one based in VSG-specific antibodies coupled to magnetic beads: it is not dependent of the complement-containing serum and it is not necessary to generate additional cell lines and the drug selection process as well.

Once the non-switchers are eliminated, the next step is the quantification of *VSG* transcripts. This can be performed by northern blot or qPCR but with these approaches only a small fraction of the *VSG* repertoire can be analyzed. In addition, the detection limit of these methodologies does not allow the detection of transcripts after a small increase in switching frequency. To improve the detection limit, the *VSG* switching frequency can be measured by combining immunoselection and *VSG*-specific RNA-seq.

Looking at perspective, it is important to determine the specific regions in the CTR DNA sequence or in the CTR transcript that interact with other genomic regions, transcripts or proteins and as a result generate the stabilizing phenotype. To precisely map the localization of the minimal element required for BES stability, the implementation of the CRISPR-Cas9 technology in BF will allow the systematic deletion of specific CTR sub-regions.

In case the CTR transcript confers the stabilizing function, the next step in characterizing function will be determining its interaction partners. The transcript may either interact in cis by recruiting specific transcription activators to the active BES or in trans by recruiting transcription repressors to the silenced BESs. In any case, once the CTR sub-regions are identified, an *in vitro* synthesized S1-tagged CTR transcript can be exposed to a BF cellular extract and then the RNA-protein complexes would be pulled-down. In this way, the interacting proteins may be determined by quantitative mass spectrometry (Scheibe et al., 2012). Furthermore, it would be possible to localize the CTR transcript by RNA-FISH. The expected result would be a co-localization with the active BES (cis acting) or with all the repressed BESs (trans acting).

During my thesis I performed some bioinformatics analyses searching for motifs involved in the formation of secondary structures in the CTRs. Therefore, I chose the putative sequences for all the CTRS located in the 15 BESs of Lister 427. I used the open access software Cofold, LocARNA and RNAstructure without any success. Nevertheless, I cannot rule out the presence of these structures even in the absence of sequence conservation. If there is any structural conservation, a mutational analysis of these motifs may provide valuable information for their deletion using the CRISPR-Cas9 approach and for the design of the RNA-FISH probes necessary for a more complete characterization of the stabilizing element present in the CTR. Furthermore, it would allow the synthesis of a shorter CTR synthetic transcript for the identification of the respective interacting partners. The use of a synthetic transcript including only the essential motifs conferring the stabilizing function would facilitate the screening by minimizing the score of false positives after pull-down the complexes.

Finally, in case that the RNA-based experiments do not produce any evidence about the role of the CTR transcript in the stabilizing function, the experiments should be focused on the DNA sequence. The first step would be to determine the nucleosomal density in the CTR making use of micrococcal nuclease to generate mononucleosomes and Southern blotting to determine the nucleosomal spacing. As it was mentioned in the introduction, mung-bean nuclease preferentially cleaves the DNA sequence between 300–800 bp upstream of the *VSG* CDS in the IaTat 1.2 strain. Therefore, It would be interesting to perform a similar experiment in Lister 427 and see whether there is preferential cleavage at the boundaries of gene conversion transposition. It would be exciting as well to observe the localization of the CTR sequence present

in the active BES. In order to achieve that, DNA FISH experiments would give some understanding about the position of the CTR regarding the active promoter or other regions of the BES.

As I mentioned at the beginning of this chapter, the ribosome profiling analysis showed moderate transcription in the region downstream of the *VSG* 3'-UTR (VDR). It would be interesting to characterize this region as well and to test whether is involved in the *in situ* switching process.

## **Chapter 5. Implementation of the CRISPR-Cas9 system for genome editing in *T. brucei***

### **Introduction**

Genome editing has become an essential tool used in multiple experimental situations (Gilles and Averof, 2014). The classical example is the deletion of genes and the replacement of the targeted region by a selective marker. After random double strand breaks (DSBs), double crossover events allow the introduction of linear DNA molecules bearing homology regions at the ends and then triggering recombination events. Cells have developed different DNA repair pathways upon DSBs such as homologous recombination (HR), non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ). In Trypanosomatids, the NHEJ pathway is very inefficient and for this reason HR is the prevailing pathway for DNA repair (Burton et al., 2007; Passos-Silva et al., 2010). Recently, it has been discovered that *T. brucei* and *T. cruzi* can use the MMEJ pathway in the absence of templates to repair DSBs (Burton et al., 2007; Peng et al., 2015). MMEJ generally uses 5- to 20-bp regions of microhomology to repair DSBs, generating deletions between the microhomology regions (McVey and Lee, 2008).

In *T. brucei*, around 60-70% of the genes are of unknown function, lacking homologues in other eukaryotes (Barry et al., 2007) and the presence of gene families with several members (i.e. *VSGs*, histones and tubulin arrays among others) is common. The *T. brucei* toolkit for genome engineering has been

gradually improved in the last years but is still limited and is necessary to complement it to have a better understanding of this pathogen.

*T. brucei* has the advantage, compared to *T. cruzi*, *L. major* and *L. donovani*, of having a functional RNA interference machinery (Ngô et al., 1998), which has helped in the elucidation of gene function (Kolev et al., 2011). Nevertheless, the RNAi system presents some pitfalls such as when the target mRNA is not sufficiently knocked down or when the transfected cells become resistant to the RNAi effect (reviewed in Boutros and Ahringer, 2008; Sledz and Williams, 2005). It is also possible that the observed phenotype cannot always be directly attributed to the targeted gene and instead to off-target effects (Jackson and Linsley, 2010)

Furthermore, *T. brucei* is amenable to the expression of exogenous genes or overexpression of endogenous genes and their deletion, but the methods are laborious and time-consuming (Barrett et al., 2004; Kelly et al., 2007; Oberholzer et al., 2006; Scahill et al., 2008). As mentioned above, the manipulation of the *T. brucei* genome depends on HR, which has low efficiency (Barnes and McCulloch, 2007; Wickstead et al., 2003), is limited to a single-allele knockout (KO) per drug selection marker (Kim et al., 2013) and the selection process per allele can take between 1 week in BF to 3 weeks in PF. Besides, the available number of drug-selectable markers is limited (G418, puromycin, hygromycin, phelomycin and blasticidin), which in turns restricts the number of manipulations that can be performed.

With the development of different genome editing systems based on engineered versions of nucleases such as zinc-finger nucleases, TALEN and

CRISPR-Cas9, a new spectrum of possibilities has been opened to manipulate genomes of different organisms, even those that were previously intractable for editing purposes (Gaj et al., 2013). In the specific case of the CRISPR-Cas9 system, the manipulation of different genomes is now more accurate, highly efficient, relatively easy and allows multiple modifications in parallel. At the moment of writing this thesis, none of these three systems had been implemented in *T. brucei*, but the CRISPR-Cas9 system has been successfully implemented in the protozoan parasites *P. falciparum*, *T. gondii*, *T. cruzi* and *L. major* and *L. donovani* (Ghorbal et al., 2014; Peng et al., 2015; Shen et al., 2014; Sollelis et al., 2015; Zhang and Matlashewski, 2015).

The CRISPR-Cas9 editing system consists of a chimeric single guide RNA (sgRNA), which guides the Cas9 endonuclease by RNA-DNA hybridization to induce a DSB in a specific region of the genome. In addition to the Cas9 binding domain, the sgRNA contains a customizable 20-nt protospacer sequence that confers the specificity for the target DNA.

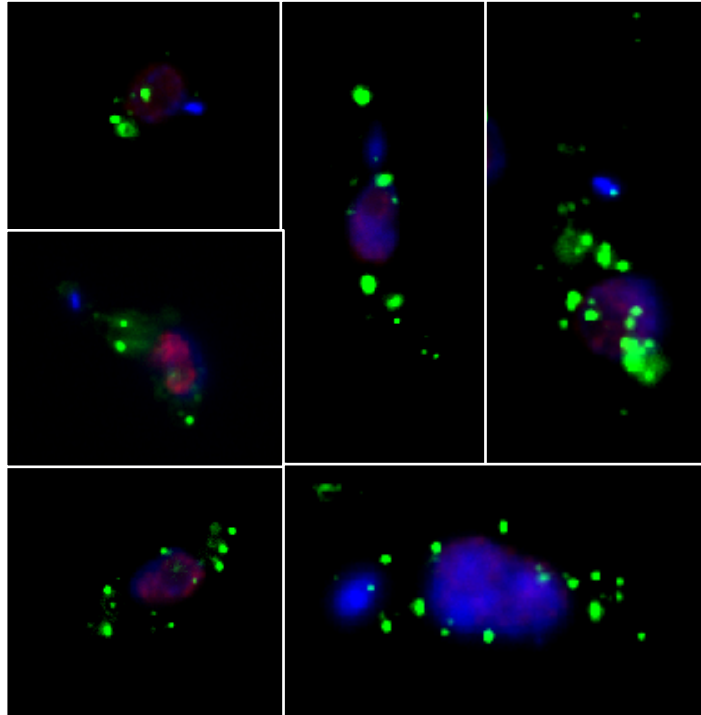
In this chapter of my thesis, I describe the establishment of the CRISPR-Cas9 system for the genome editing of *T. brucei* and demonstrate its efficient use in PF. In the next sections, I will show the rapid, efficient and marker-free insertion by homologous recombination (HR) of a tag containing the eGFP CDS at the 3'-end of the SCD6 CDS in PF. In addition, I confirmed the absence of off-target edition by whole genome sequencing, which is the main concern regarding genome-editing systems based in engineered nucleases.



## Results

### **The Cas9, sgRNA and repair sequence of the CRISPR-Cas9 system were cloned in two episomes for their implementation in PF of *T. brucei***

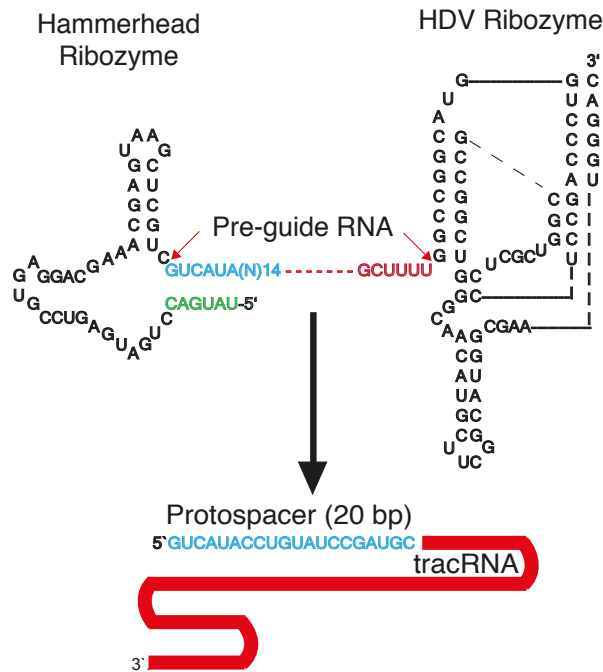
I adapted the CRISPR-Cas9 system in *T. brucei* by generating two episomal vectors: one containing the Cas9 endonuclease gene and the other the sgRNA and the repair DNA template. Human codon adapted SpCas9, bearing a SV40 derived nuclear localization signal, was expressed under the control of the PARP promoter in the episome pEV\_Cas9v2\_puro (conferring resistance to puromycin) (Figure 12, left panel). The expression and nuclear localization of SpCas9 was determined by IF using an antibody against the Ty1 tag (Figure 31). SpCas9 localized mainly in the cytoplasm but the fraction imported to the nucleus was enough to induce DSBs in the *T. brucei* genome. Previous efforts performed by MSc. Carolin Wedel in the Siegel laboratory to improve the SpCas9 nuclear localization by using constructs with different NLSs, or moving the NLSs from the 5'-end to the 3'-end of the SpCas9 were unsuccessful. Therefore, the construct used in this work provided the best SpCas9 localization among the several tested (data not shown).



**Figure 31.** Cellular localization of a human codon adapted SpCas9 expressed in PF. A specific Cas9 antibody was used to detect the localization of the endonuclease (green signal). An anti-H2A.Z antibody was used to stain the nucleus (red signal). DAPI stains the nucleus and the kinetoplast (blue signal). Representative images are shown. The Cas9 expression was not homogeneous in the cells evaluated under the microscope.

The first 20 nucleotides of the guide RNA transcript define the CRISPR target and thus precision at the 5'-end is critical to work properly. It is believed that U6 promoter transcription starts at the +1 position (23-nt after the TATA box), with G as the preferred initiation nucleotide (reviewed in Ma et al., 2014). Therefore, RNA polymerase III promoters can process sgRNAs more accurately than others such as RNA polymerase II promoters. In addition, the transcripts generated from the latter promoters usually enter the mRNA modification pathway (capping and poly-A tailing), which is undesirable in the sgRNAs. RNA polymerase III promoters, such as the U6 promoter, have been successfully used to implement the CRISPR-Cas9 system in other organisms, including *P.*

*falciparum* (Ghorbal et al., 2014) and *L. major* (Sollelis et al., 2015). For example, in *L. major* the promoter sequence used harbors a part of the sequence coding for the tRNAGln gene including the canonical boxes A and B flanked by the 5'-UTR of U6snRNA and the terminator comprising 120-bp of the U6snRNA 3'-UTR. In *T. brucei* a similar U6 RNA polymerase III promoter structure has been reported (Nakaar et al., 1997). However, taking into account that the exact sites for processing the U6snRNA transcript are not well defined and that this information is fundamental for the correct processing of the sgRNA, I decided to follow a different strategy based on ribozymes, which has been successfully used in other organisms (Gao and Zhao, 2014). In this strategy, the sgRNA is transcribed together with ribozyme sequences at both ends: hammer head (HH) at the 5'-end and hepatitis delta virus at the 3'-end (HDV). Then the primary transcript undergo self-catalyzed cleavage to generate the desired sgRNA, which in turn can efficiently guide the Cas9 endonuclease to induce sequence-specific cleavage of DNA targets. The primary transcript can be transcribed from any promoters. I used the PARP promoter for the transcription of the sgRNA transcript flanked by the ribozyme sequences in order to be precisely processed (Figure 32).



**Figure 32.** Processing of the sgRNA by the HH and HDV ribozymes. The blue sequence represents the protospacer and the red one the tracrRNA that together form the sgRNA. The green sequence corresponds to the 5'-end of the HH ribozyme, which anneals complementary to the first 6-nt of the protospacer. The HH ribozyme processes the 5'-end of the sgRNA and the HDV ribozyme the 3'-end. The pre-guide RNA is accurately processed to produce the mature form of the sgRNA (Figure adapted from Gao and Zhao, 2014).

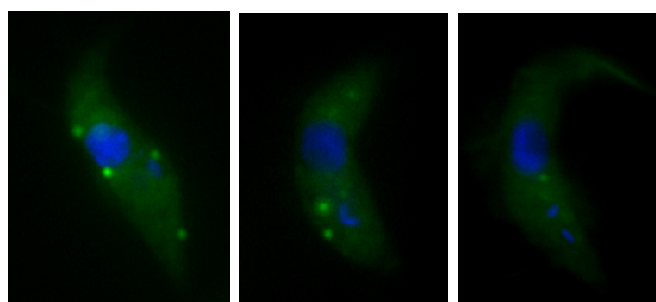
Each sgRNA and the repair DNA template for HR were placed in the same episome, pEV\_RS\_sgRNA\_G418 (Figure 12, right panel). To ensure maximal specificity and efficiency of the CRISPR-Cas9 editing system, I included two large homologous regions (~ 250-bp) to promote efficient HR. The sequence of each protospacer was included in the homologous regions preceding the NGG motif present in the targeted SCD6 sequence. The Zhang lab recommends using homology regions that are less than 15-100 bp away from the DSB site to increase the efficiency of recombination (<https://www.addgene.org/crispr/zhang/FAQ/>).

## **Efficient CRISPR-Cas9 mediated gene editing in *T. brucei***

To determine the ability of the CRISPR-Cas9 system for editing genes in *T. brucei*, I generated a transgenic cell line by tagging the 3'-end of the SCD6 CDS with an eGFP CDS. SCD6 is a protein located in P-bodies and responsible for granule assembly in trypanosomes (Kruger et al., 2013). I chose this protein because it localizes in cytoplasmic foci, it is clearly visible under the microscope and because my ribosome profiling results indicate that this protein is abundant in both PF and BF. The episome containing the sgRNA and the repair sequence SCD6\_eGFP was named pEV\_sgRNA\_SCD6\_GFP. This repair sequence consisted of two homology regions including the 3'-end of the SCD6 CDS (excluding the stop codon), the entire eGFP CDS with the exception of the start codon and part of the 3'-UTR of the SCD6 gene (Figure 12, right panel). Three sgRNAs were chosen because variability in editing efficiency has been reported according to the genomic context (Ran et al., 2013a) (Figure 13). In the approach described here, the SpCas9 protein was constitutively expressed from the pEV\_Cas9v2\_puro episome.

I tried co-transfection and sequential transfection approaches with the two episomes. In the co-transfection approach, the permanent selection with two antibiotics (puromycin and G418) caused a decrease in the growth rate to the point that only after one month I was able to have enough cells to evaluate them by microscopy. Additionally, the cells showed a clear growth defect, in special a rounded morphology, which can be explained by the simultaneous selection with puromycin and G418. In contrast, the sequential transfection of PF resulted in highly efficient SCD6-GFP expression and the green fluorescence was visible

even as early as day 2 after transfection (Figure 33). The cells were first transfected with pEV\_Cas9v2\_puro, which under continuous drug pressure (puromycin) allowed the cells to keep the episome for at least five months. It has been reported that cell lines expressing some exogenous proteins may occasionally have altered growth kinetics (Peng et al., 2015). Cas9 expression in *T. cruzi* and *S. cerevisiae* (Peng et al., 2015; Ryan et al., 2014) decreased the growth potential presumably due to the Cas9 ability to bind to unspecific target regions upstream of canonical PAMs (off-target effects). In my experiments, the doubling time of PF constitutively expressing SpCas9 was higher compared to wild type cells but the cells did not show any appreciable morphological defect. Therefore, the increase in the doubling time may be attributed just to the puromycin selection.



**Figure 33.** Localization of SCD6-eGFP expressed in cells sequentially transfected with episomes carrying the components of the CRISPR-Cas9 system. Upon Cas9 induced DSBs, one of the episomes provided a repair sequence for the insertion of eGFP between the last coding codon of the SCD6 CDS and the first codon of the SCD6 3'-UTR. Representative images are shown. Nucleus and kinetoplast stained with DAPI (blue). (Left and central panel) Images of PF successfully expressing the fused protein upon CRISPR-Cas9 genome editing. (Right panel) Image of PF cell transfected with pMOTag3G\_SCD6 expressing eGFP.

Upon the first round of transfection, the cells were transfected with the episome containing the repair sequence and the sgRNA (pEV\_sgRNA\_SCD6\_GFP).

After analyzing the cells by fluorescent microscopy, each of the three sgRNAs successfully guided the Cas9 to the insertion of the eGFP CDS between the SCD6 CDS and the SCD6 3'-UTR. The observation of fluorescent cells suggested the efficient cleavage and repair of the target sequence during the first few growth cycles after transfection. The sgRNA1 was the best ranked according to the <http://spot.colorado.edu/~slin/cas9.html> resource and no difference in the number of fluorescent cells were observed after comparing cells transfected with the three different sgRNAs. For this reason the cell line transfected with the sgRNA1 was chosen for further experiments.

The time window for plating out the cells after the second transfection with pEV\_sgRNA1\_SCD6\_GFP was standardized in order to obtain clonal populations and to estimate the editing efficiency and the allelic frequency of editing. Therefore, immediately after transfection the cells were transferred to a T25 flask with 5 ml of SDM79 (20% FSC) and the next day puromycin and G418 were added to the cultures. Then the cells were either sequentially diluted (1:1 x 10<sup>1</sup>, 1:1 x 10<sup>2</sup>, 1:1 x 10<sup>3</sup>, 1:1 x 10<sup>4</sup>, 1:1 x 10<sup>5</sup> and 1:1 x 10<sup>6</sup>) in 96-well plates or kept in T25 flasks between 5 and 17 days and then sequentially diluted as well. In all cases the cells were kept under drug pressure.

When the cultures were immediately plated out after drug addition, wells containing dense cell cultures were observed after ~ one month, especially for the more diluted plates where it is expected to find clonal populations. In contrast, after keeping the parasites in T25 flasks for 10 days and then plate them out, the cells grew faster. Around three weeks after transfection, clones in the 1:1 x 10<sup>6</sup> dilution were dense enough for being transferred to larger well

plates and eventually to flasks for next procedures. It is well known that for PF the cell density is crucial to keep the cultures healthy and that the release of some molecule(s) by the same cells is important for this process (Brun and Schöenberger, 1979). Burkhard *et al.* (Burkard et al., 2007) have reported that the survival rates for PF after transient transfection is ~59% and the percentage of fluorescent cells is 20%. It is possible that the number of parasites even in the 1:10 dilution is already low in the case of the cells plated out 24 hours after transfection. Furthermore, taking into account that only around 20% of the parasites may contain the second episome, the drug selection process can decrease the number of viable cells even more.

In parallel, each well was evaluated for the presence or absence of fluorescence using an EVOS® FL Imaging System for the cells kept initially in the flask and plated out after 10 days. These results already suggest that keeping the cells during continuous drug pressure for three weeks increase the editing frequency to ~ 89% (Table 5).

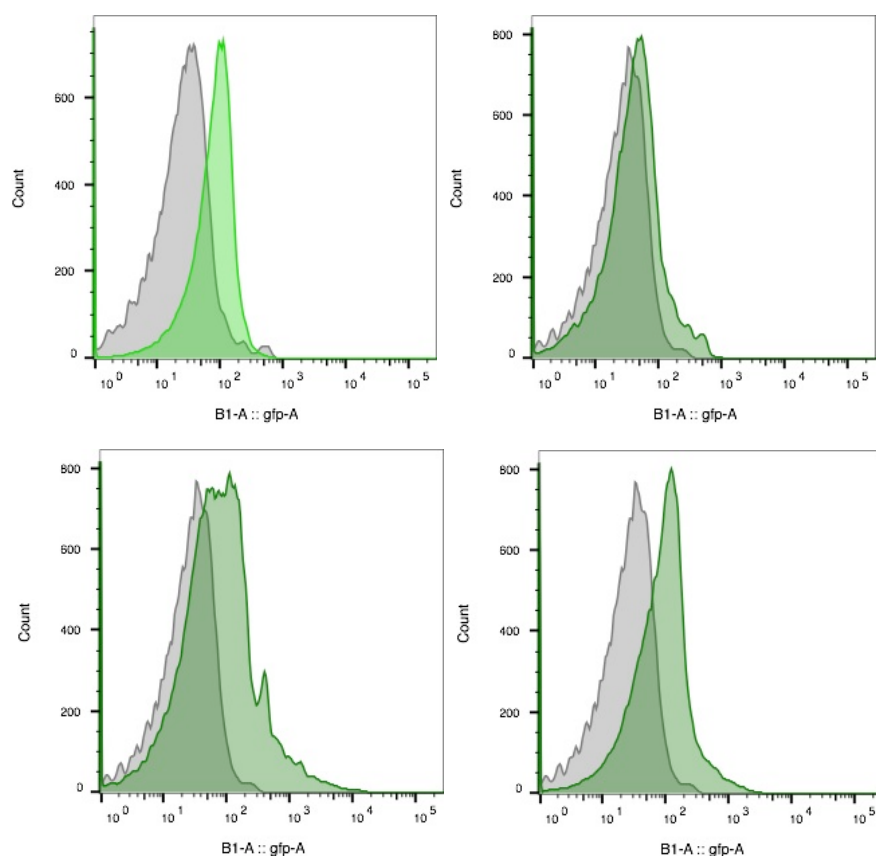
**Table 5.** Percentage of fluorescent cells three weeks after the second round of transfection.

Dilution	Absence of cells	Presence of cells		% Fluorescent cells
		Non fluorescent	Fluorescent	
1:1x10 <sup>1</sup>	0	ND	ND	ND
1:1x10 <sup>2</sup>	0	ND	ND	ND
1:1x10 <sup>3</sup>	12	9	75	89
1:1x10 <sup>4</sup>	45	6	45	88
1:1x10 <sup>5</sup>	78	1	17	94
1:1x10 <sup>6</sup>	77	2	17	89

Number of wells in which cells were present or absent and where fluorescent cells were detected or not using the EVOS® FL Imaging System. ND: not determined.



In order to have more confidence about the clonality of the populations, I chose 10 clones in the  $1:1 \times 10^6$  dilution and analyzed them in the MACS Quant (Figure 34). This approach allowed me to distinguish the two peaks that corresponded to fluorescent and non-fluorescent cells even though the two peaks overlap at some degree.

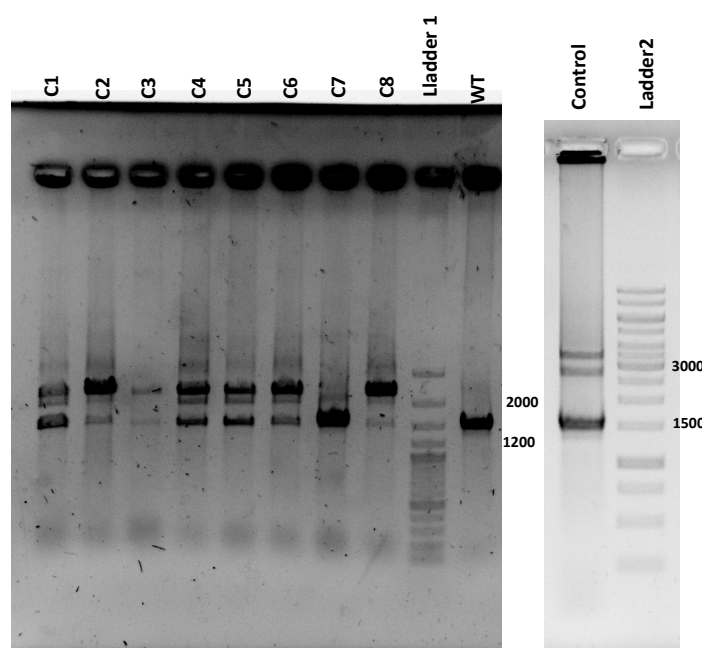


**Figure 34.** Flow cytometry analysis of PF transfected with pEV\_Cas9v2\_puro and pEV\_sgRNA1\_SCD6\_GFP. Shown are representative examples of 3 clones. Grey color represent wild type PF, bright green PF transfected with pMOTag3 and pale green PF sequentially transfected with the episomes containing the components of the CRISPR-Cas9 system. (Upper left panel) PF transfected with pMOTag3 and expressing SCD6-GFP represent a positive control for fluorescence. (Upper right panel) Sequentially transfected PF in which no fluorescence was detected. (Bottom left panel) Sequentially transfected PF probably representing a non-clonal population. (Bottom right panel) Sequentially transfected PF presenting a similar pattern compared to the positive control.

Nevertheless, to have a more accurate picture of the clonality of the selected populations, I took images of the cells present in 5 different fields using a Leica DMI6000B microscope with a HCX PL FLUOTAR dry objective with 40x0.6 magnification. Then, I counted the number of cells (DAPI staining) and the number of fluorescent cells (green fluorescence from SCD6-GFP). Around 86% of the cells were clearly fluorescent. Nevertheless, the quality of the images obtained was not the best given that not all the cells were in the same focal plane. Therefore, it is possible that the fluorescence of some cells was not detected, which can underrepresent the number of positive cells. This result indicates that the majority of the cells were modified by the CRISPR-Cas9 system but I cannot conclude with total certainty that ~ 100% of them are fluorescent. Further experiments improving the preparation of the samples are necessary to be more confident about the efficiency of editing on the selected populations.

The next step in the implementation of the CRISPR-Cas9 system was to analyze the allelic editing frequency in the selected samples. Therefore, the gDNA was extracted and PCR amplification performed using primers that annealed in the genomic DNA outside the repair sequence present in the episome pEV\_sgRNA\_SCD6\_GFP (see material and methods). PCR analysis showed a larger fragment (2186-bp compared to 1466-bp for the endogenous gene) corresponding to the insertion of the eGFP CDS through HR (Figure 35). Both bands were observed in six clones (1-6). The clone number 8 presented an intense band corresponding to the SCD6-GFP allele but a faint band was still observed for the endogenous allele (Figure 35). The band corresponding to the endogenous allele was sent for Sanger sequencing and a substitution was

observed in the third position of the PAM sequence (CGC instead of CCG). This result suggests that apparently the selection process selected for clones expressing one of the non-tagged alleles and that it was not possible to modify both SCD6 alleles simultaneously.



**Figure 35.** Evaluation of the allelic editing frequency by PCR. gDNA from 8 sequentially transfected clones were amplified using primers annealing in the *T. brucei* genome outside of the repair sequence present in the episome pEV\_sgRNA1\_SCD6\_GFP. WT: wild type PF. Control: gDNA extracted from PF transfected with pMOTag3\_SCD6\_GFP. Ladder 1: 100 bp plus DNA ladder (Thermo Scientific). Ladder 2 :1 Kb DNA ladder (Thermo Scientific). Expected size of the bands: The band for the endogenous allele corresponds to 1466-bp and for SCD6\_GFP to 2186-bp. In the pMOTag3 cell line the band corresponding to SCD6-GFP is 3599-bp. Clone number 7 correspond to transfected cells in which almost no fluorescent cells were observed.

In summary, the CRISPR-Cas9 strategy described here allowed the successful insertion of an eGFP tag to the 3'-end of the SCD6 CDS in PF cells, obtaining ~ 90% of editing efficiency. Nevertheless, it was not possible to obtain homozygous clones for the two SCD6 alleles tagged with the eGFP sequence. The

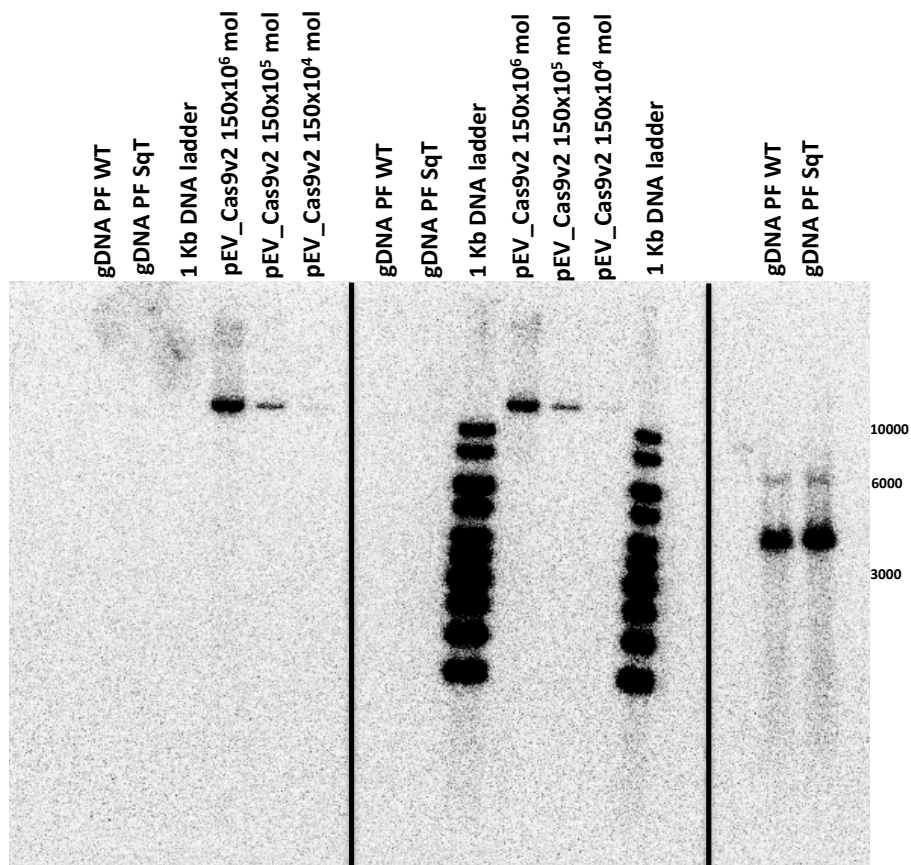
complete process of transfection, selection and screening can be performed in around one month.

### **Episome stability in *T. brucei***

The episomes generated in this work were modified from the pT13-11 episome constructed by Patnaik *et al.* (Patnaik et al., 1993). In that paper, it was mentioned that even without drug selection the pT13-11 episome was quite stable for several generations. In order to address the stability of my constructs, I removed the drug pressure from parasites sequentially transfected with both episomes. The cells were under no-drug pressure for two months and then they were plated out in serial dilutions with or without drug pressure. Living cells were observed in all the dilution plates where drug pressure was included again. It means that even after two months with no-drug selection, the cellular machinery replicated and segregated both episomes at least for a fraction of the population.

The plasmid maintenance sequence (PMS) present in pT13-11 and pEV includes sequences corresponding to calpain-like cysteine peptidase (chromosome XI),  $\beta$ -tubulin chaperone, ubiquitin-like domain containing protein, TaD related DNase (chromosome III) and a pseudo *VSG*. Therefore, after a random DSB in any of these regions in the genomic DNA of the parasites, the episome may integrate into the genome. To evaluate the integration of the episomes in the genomic DNA, I performed Southern blotting using probes against a region in the SpCas9 and the lactamase sequence. Genomic DNA from 150 million cells was extracted from wild type cells and from cells sequentially

transfected with the episomes and digested with the six-base cutter restriction enzyme KpnI. pEV\_Cas9v2\_puro has digested as well with KpnI and amounts of DNA equivalent to  $150 \times 10^4$ ,  $150 \times 10^5$  and  $150 \times 10^6$  molecules were loaded into the gel. According to Patnaik *et al.* each transfected cell contains on average one copy of the episome (Patnaik *et al.*, 1993). The results showed that no episomal integration was detected after loading the gDNA equivalent to 150 million of cells (Figure 36). For the Cas9 probe, I can conclude that less than  $150 \times 10^4$  cells (10%) contain the pEV\_Cas9v2 episome. For the lactamase probe, less than  $150 \times 10^4$  cells contains at least one of the episomes.



**Figure 36.** Testing the episome stability in PF by Southern blot. (Left panel) Samples hybridized with the Cas9 probe. (Central panel) Samples hybridized with the lactamase probe. (Right panel) Samples hybridized with the  $\beta$ -tubulin probe. WT: wild type PF. SqT: PF sequentially transfected with the two episomes. The screen was exposed for 1 hour.

The two previous results suggested that the episomes were stable maintained even without drug selection but at the same time they were not frequently integrated into the genomic DNA.

To expand the applicability of the CRISPR-Cas9 toolkit for editing the *T. brucei* genome, it would be ideal to be able to remove the episomes in order to avoid additional alterations in the genome, off-target effects or toxicity caused by the Cas9 endonuclease. In addition, it would be advantageous to be able to make new rounds of editing with other episomes. Therefore, I generated episomes with the puromycin/HSVTK negative selection system to evaluate the possibility of killing cells containing these episomes upon addition of ganciclovir. However, even with ganciclovir doses higher than 100 µg/µl the parasites still survive the drug selection. In PF, Valdes *et al.* (Valdés et al., 1996) reported a reversion rate of 10<sup>-6</sup> per cell-generation to a ganciclovir resistant phenotype. In general, PF cells need higher concentration of drugs than BF and ganciclovir is not the exception. In addition, other reports mention that large amounts of ganciclovir (several hundred µg/ml) are not enough to kill PF effectively (Hee-Sook Kim, personal communication). An alternative for the elimination of the episomes is mentioned in the discussion section.

### **Off-target mutation analysis**

Potential off-target regions were identified in the Lister 427 genome using the Protospacer Workbench and the third party tool Cas-OFFinder. I searched for sequences that matched the sgRNA1 targeting sequence for SCD6 with less than 10 mismatches and excluding any mismatches in the PAMs since

these motifs are required for Cas9 DSBs (Cho et al., 2014). The overall quality of the sgRNA1 design was determined by ranking the score of each off-target site. Using these parameters, the sgRNA1 had a total of 69,269 potential off-target sites in the *T. brucei* genome. This analysis showed decreasing scores as the rank increases and differences between the intended target and the next best target. The top 10 potential off-target sites are showed in Table 6.

**Table 6.** List of putative off-target sites and their location regarding indels for the sgRNA1 (SCD6-GFP)

Chromosome	Position	Strand	Sequence	PAM	Mismatches	Score	<20-bp indel
Tb427_11_01_v4	140551	+	AGCGGCCGTTGTAACGTCCC	CgG	0	1	
Tb427_07_v4	1348504	+	ttCtGcTtTcGTAACGTCCC	AaG	6	0.003781263	NO
Tb427_11_01_v4	4490022	+	AaaGGtCGTgGTAACGTCCt	TgG	5	0.002261262	NO
Tb427_06_v4	478380	+	AGCGGcTtTTGccACGTCCC	CgG	4	0.001775052	NO
Tb427_05_v4	1346351	+	AtgtGCCtccaTAACGTCCC	TaG	7	0.001341647	NO
Tb427_05_v4	1352886	+	AtgtGCCtccaTAACGTCCC	TaG	7	0.001341647	NO
Tb427_04_v4	940606	-	gGtaaCgcTTGTcACGTCCC	TgG	7	0.001161574	NO
Tb427_03_v4	210929	+	AaataCCGTTGacACGTCCC	AaG	6	0.001149468	NO
Tb427_10_v5	3023133	-	AGaGGCaacgGaAACGTCCC	CgG	6	0.001120779	NO
Tb427_11_01_v4	2614623	-	cGgcaCCtccaTAACGTCCC	AgG	8	0.001023911	NO
Tb427_11_01_v4	3675931	+	AttGtttGTTGTtACGTCCC	GaG	6	0.000965401	NO

<20-bp indel: putative off-target sites less than 20-bp from a specific indel in the sequentially transfected cell line

Genomic DNA from wild type and sequentially transfected parasites was extracted to prepare libraries for whole genome sequencing and for searching PAM-associated indels uniquely present in the sequentially transfected (mutant) cell line. The software VarScan2 (Koboldt et al., 2012) detected 4,205 and 1,751 unique indels in the wild type and the mutant cell line respectively and 2,436 in common. This analysis left 4,205 indels potentially caused by the CRISPR-Cas9 editing. However, none of these indels fall within 20-nt of any putative off-target site (up to 10 mismatches) precluding the possibility that they were caused by Cas9 off-target activity. Therefore, the results indicated the accuracy and

reliability of the CRISPR-Cas9 system to modify the target without the inclusion of off-target effects, at least not within the limits evaluated here.

## **Discussion**

To my knowledge, this is the first study of CRISPR-Cas9 genome editing in *T. brucei*. My work demonstrates that the CRISPR-Cas9 system mediated the *in situ* tagging of the SCD6 gene with an eGFP tag. Although it was not possible to obtain clonal populations containing both alleles tagged with eGFP, even two days post transfection the fluorescence of the cells was evident. The episomes used as carriers of the CRISPR-Cas9 elements were kept in the cells for several months even with no drug selection but they were not integrated into the genome of the PF. In this part of my thesis, I demonstrate that the CRISPR-Cas9 system is a versatile and powerful tool for genome engineering in *T. brucei*, broadening the possibilities of discerning the functions of previously uncharacterized genes and manipulating genes families containing several copies.

Understanding the complexities of a parasite like *T. brucei* is greatly facilitated by the ability of manipulating its genome by introducing deletions, insertions or substitutions of new or enhanced functions. The production of null mutants for single-copy genes requires the deletion of each allele in two rounds of transfection, which takes between 1-2 months depending of the life cycle stage under study. In addition, the deletion of several genes in the same cell line is complicated given the limited number of available resistance genes. Removal of one of the previous resistant markers used can be a solution to the problem but



implies more rounds of transfection. Other limitation is the editing of gene families comprising several members such as *VSGs* or histones.

The CRISPR-Cas9 system provides several advantages over the available methods for the editing of the *T. brucei* genome. One of these advantages is the site-specific introduction of DSBs by SpCas9 that greatly increase HR between a supplied template and the specific target in the genome. By introducing a DSB via the RNA-guided Cas9, I was able to introduce the eGFP CDS at the 3'-end of the SCD6 CDS at efficiencies higher than the obtained by spontaneous DSBs.

Gene disruption was highly efficient in PF, with up to 90% of the population exhibiting a mutant phenotype. In addition, once the cells contained the pEV\_Cas9v2\_puro episome, the selection process until obtaining clonal populations took on average one month. According to my results, keeping the cells in a flask after the transfection promotes a faster growth of the cells before diluting and plating them out in 96-well plates. The high rate of gene editing provided by the system allowed to test the fluorescence of the cells as early as 2 days after transfection. Additionally, cell sorting can be performed for rapid isolation of knock-in fluorescent protein expressing cells. Nevertheless, the observation by microscopy of a minority of non-fluorescent cells may imply that the non-fluorescent cells survived the antibiotic selection because they still contained the two episomes. In *L. major*, Sollelis *et al.* used FACS to analyze GFP-positive and GFP-negative parasites following transfection with a sgRNA for eGFP disruption and measured Cas9 levels by ELISA (Sollelis *et al.*, 2015). The non-modified parasites had significantly lower levels of Cas9 protein than parasites in which GFP was disrupted by sgRNA transfection. These results

suggest that the observed efficiency of gene disruption of less than 100% may be due to variable levels of Cas9 expression. Taking into account that the localization of Cas9 in my system is mainly cytoplasmic, it is possible that differences in the level of the endonuclease in the nucleus explain the observation of non-fluorescent cells. Sollelis *et al.* suggest that the sgRNA transcription driven by the U6snRNA promoter may be weak as well (Sollelis *et al.*, 2015). In my system, the PARP promoter drives the sgRNA transcription, which is highly active in PF for the production of the surface-binding procyclins (Rudenko *et al.*, 1990).

Another advantage of the CRISPR-Cas9 system is the possibility of editing both alleles of a gene after only one round of transfection. It was not possible to obtain cells containing both alleles tagged with the eGFP CDS because the second allele presented a mutation in the PAM sequence. It has been shown that upon RNAi induction targeting the SCD6 transcript, the cumulative cell number of PF decreases compared to the non-induced treatment (Kramer *et al.*, 2012). In addition, SCD6-eYFP/- PF (one SCD6 allele deleted and the other tagged with eYFP) grew slower compared to SCD6 + / + PF (containing both endogenous alleles) (Kruger *et al.*, 2013). Therefore, it is possible that PF with both alleles tagged and therefore expressing only the fused protein may have a growth disadvantage compared to the cells expressing the two types of proteins (tagged and non-tagged). For this reason cells containing one tagged allele and a mutation in the PAM sequence on the other allele could grow faster and be the predominant type in the populations selected.

In my work I also analyzed the potential for off-target sequence disruption. When using genome editing tools such as zinc-finger nucleases, TALEN and CRISPR-Cas9, off-target events are always a concern and it is recommended to evaluate their impact in the respective genome (Fu et al., 2013). Using the <http://spot.colorado.edu/~slin/cas9.html> resource I was able to select gene specific sgRNAs to target the *T. brucei* genome. At present, the Protospacer Workbench (MacPherson and Scherf, 2015) is able to analyze the Lister 427 genome for the selection of highly gene specific sgRNAs with low sequence identity to other sites in the *T. brucei* genome. This tool also offers the possibility to analyze the presence of NGG motifs close to the off-target sites in order to minimize their effects. In my study, I confirmed the lack of off-target effects analyzing deep sequencing data from mutant parasites with the software VarScan2 (Koboldt et al., 2012). None of the indels detected matched the predicted off-target sites by Cas-OFFinder (Bae et al., 2014). Additionally, Southern blot experiments also indicated the lack of integration of the two episomes into the parasite genome. The lack of evidence for off-targeting activity is not surprising because the *T. brucei* NHEJ pathway has low efficiency, which may cause the cell death upon off-target DNA cleavage. In conclusion, gene editing can be performed in *T. brucei* without production of detectable off-target mutations.

The stability of the episomes might preclude the possibility of performing further editing of the parasites with the CRISPR-Cas9 system. The probability of cells keeping more than two different episomes and resist the selection of several drugs is low. Using a negative selection system (HSVTK) for killing cells

carrying the episomes was not possible even at the high doses of ganciclovir added to the cultures. Nevertheless, the same CRISPR-Cas9 system can be used to inactivate the resistant marker present in the episome containing the repair sequence and the sgRNA (G418). Once the first editing event has been performed and screened, the cells can be transfected with a new episome that confers resistance to another antibiotic and includes a repair sequence for a new target. In addition, this episome should contain two sgRNAs: one for the modification of the new target and another for the deletion of G418. In this way, the parasites will contain the Cas9 endonuclease, the new episome and will be resistant to only two drugs (puromycin and another drug different to G418). This strategy would open the possibility of performing further manipulation of the genome in additional rounds of transfection.

The tagging of SCD6 with eGFP offered an example for the introduction of a gene in the genome of *T. brucei* but the CRISPR-Cas9 system can be used for the deletion of genes as well. For example it is possible to target H3.V and H4.V for their deletion in PF. In addition to H2A.Z and H2B.V, *T. brucei* contains variant forms of H3 and H4, whose functions are just beginning to be elucidated. IF microscopy and standard ChIP followed by dot-blot readout has revealed that H3.V is enriched at telomeric repeats (Lowell and Cross, 2004) and transcription termination sites (Siegel et al., 2009). No lethal phenotype was detected for a cell line lacking both alleles of H3.V and H4.V (Lowell, 2005; Lowell and Cross, 2004; Siegel et al., 2009). Taking into account that H3.V and H4.V may play a role in transcription termination and that they apparently are not essential for the viability of the parasite, they can be good models for evaluating the deletion of

genes using the CRISPR-Cas9 system. In case these experiments prove to be successful, it would be possible to generate KO cell lines in a shorter period of time than with the available tools requiring two sequential rounds of transfections.

Another interesting application of the CRISPR-Cas9 system in *T. brucei* is the editing of multigene families such as *VSG*,  $\alpha$ - and  $\beta$ -tubulin, actin and rDNA arrays. Using traditional approaches, the editing of these arrays may be time consuming and sometimes not feasible depending on the genomic context and differences in sequence of each copy. Additionally, there are available only five selectable markers that work in *T. brucei* and therefore after introducing all of them in the genome of the parasite, it would be necessary to remove some of them in order to continue the editing of the other copies in the multigene array. One example, which is relevant for the epigenetic control of gene expression in the parasite, is the distribution of H4 in a tandem array of at least 17 haploid copies in chromosome V as suggested by the most recent analysis of this multigene family in Lister 427 (Dr. Raul Cosentino, personal communication). The CDSs of the different members in this gene family are almost identical and only minor differences are present in their UTRs making it possible to target the whole gene family. In addition, it would be interesting to introduce a non-synonymous conservative substitution in the H4 array. Janzen *et al.* (Janzen et al., 2006b) have found that ~ 80% of H4 is acetylated at K4 and ~ 10% is acetylated at K10. In a subsequent study, Siegel *et al.* (Siegel et al., 2007) characterized the acetylation of lysine 4 of histone H4 (H4K4) by generating highly specific antibodies to acetylated and unmodified H4K4. They showed that the histone

acetyltransferase HAT3 is responsible for H4K4 acetylation. Therefore, the idea is to substitute H4K4 for H4R4 by exchanging the codon AAG (Lys) for CGC (Arg). In this case, the repair sequence should contain the H4 CDS including the aforementioned substitution as well as the synonymous substitution of the PAM target sequence (also called shield mutation) to avoid the permanent cleavage of the Cas9 in the same allele. It is important to mention that the shield and the desired mutation should be close to each other to minimize the possibility that only the shield modification is incorporated. Given the high number of alleles to modify (34 copies), it is possible that the period of selection has to be increased until the majority of alleles include the substitution. To screen the number of copies presenting the substitution, the gDNA of several clones will be extracted, the region that includes the substitution will be amplified by PCR and the corresponding band excised from the gel, cleaned up and sent for Sanger sequencing. The rationale of this screening is that if several of the 34 alleles are modified, two peaks should be observed in the Sanger chromatogram for each of the three nucleotides present in the fourth codon: those for the original codon that codifies for lysine and those that codify for arginine. If the Sanger chromatogram evidence that the majority of peaks corresponds to CGC (Arg) and not to AAG (Lys), the strategy will be at least partially successful and the next step will be to generate a library to determine how many copies of the alleles were modified.

Another possibility is to evaluate if there is a phenotype after the substitution of H4K4 for H4R4 using an indirect enzymatic assay. As was already mentioned, H4 is acetylated by histone acetyltransferase 3 (HAT3) and in turn HAT3 and SIR2rp1 are required for efficient RAD51-dependent homologous recombination

and therefore for DNA repair (Glover and Horn, 2014). Basically, the idea is to induce DNA damage with methyl methanesulphonate (MMS) or phleomycin. If after the action of the CRISPR-Cas9 system there are few lysines at position 4 in H4 to be acetylated by HAT3; then the DNA repair efficiency would decrease leading to a growth defect.

The implementation of the CRISPR-Cas9 system in *T. brucei* opens opportunities for new applications like the use of catalytically inactive dCas9 fused to repressive or activating chromatin-modifying effectors (Gilbert et al., 2013). Another example was provided by the high multiplexing capacity of the CRISPR-Cas9 system in *T. cruzi*, where it was possible to knock down the expression of a gene family consisting of 65 members using three sgRNAs (Peng et al., 2015). Therefore, in the future it would be interesting to transfect *T. brucei* cells with an episome containing several sgRNA that in turn target Cas9 to different regions of a multicopy gene array and then evaluate the downregulation of the protein product. Another major improvement, especially when studying essential genes, would be the development of an inducible Cas9 gene system (CRISPRi) that would provide the opportunity to generate null mutants in essential genes and monitor them in specific time points as protein level decrease. The idea behind CRISPRi is to make the knockdown of genes more reliable given that RNAi can sometimes exhibit significant off-target effects and toxicity (Qi et al., 2016) and it is not existent in all organisms as is the case of *T. cruzi* (DaRocha et al., 2004), *L. major* and *L. donovani* (Robinson and Beverley, 2003) or divergent and difficult to implement as in *P. falciparum*, *T. gondii* and *Giardia intestinalis* (reviewed in Kolev et al., 2011).

CRISPR-Cas9 has revolutionized genome editing in multiple species, some of them previously intractable, and in this part of my thesis, I have shown the feasibility for use in the human pathogen *T. brucei*. I expect that the CRISPR-Cas9 system will rapidly become a routine laboratory-editing tool for the *T. brucei* research community. In the future this system will be a powerful tool for manipulation of essential genes, large gene families and transcription control for a greater understanding of the physiology of this parasite and its interactions with its animal hosts.



## Chapter 6. Conclusions and future directions

### **Ribosome profiling enables a genome-wide translation analysis in different life cycle stages of *T. brucei***

The paper titled “Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages” (Vasquez et al., 2014) was the first genome-wide analysis of protein synthesis for a eukaryotic pathogen and an organism without transcriptional control. It was also the first strand-specific analysis of RNA transcript levels in *T. brucei*. We found that translational efficiency varied between genes in each life cycle stage analyzed and around 100-fold between the PF and the BF. According to these results, translational regulation contributes to gene expression to a similar extent as RNA stability. The sub-codon resolution of ribosome profiling revealed extensive translation from uORF and led to the identification of around 2,000 un-annotated putative CDSs. Twenty of them were validated upon comparison with a previous proteomic dataset. Genome-wide RNAi data suggested an important role for >200 of those CDSs in parasite survival, especially in the form that is infective to mammals. Thus, ribosome profiling data demonstrate the importance of translational control in the two life cycle stages of *T. brucei* most commonly used for *in vitro* studies. Additionally, the dataset is now available at TriTryp database (<http://tritrypdb.org/tritrypdb/getDataset.do?display=detail>) for the scientific community interested in the study of the influence of translational regulation in this parasite.

Ribosome profiling has demonstrated to be a powerful application and therefore additional strategies can produce new information about translational control. It is possible to re-annotate sequences with the data we made publicly available. For example, the start codon can now be accurately localized for those CDSs that contain consecutive ATG codons at the 5'-end. In our work we also showed extensive upstream translation at non-AUG codons in PF. It would be interesting to see if this phenomenon is related to a more relaxed translational machinery given the temperature PF are exposed to (27°C) and if it has biological significance for the control of gene expression. The widespread presence of canonical uORF raises the question of whether these short sequences affect the initiation of translation of the downstream CDS and whether they achieve this putative regulation by coding small peptides as has been described in yeast, mammals and other organisms (Andrews and Rothnagel, 2014; Somers et al., 2013).

The absence of aligned footprint reads to several non-annotated features and annotated CDSs was observed in our work. The former may be attributed to the presence of ncRNA specially if there is evidence of at least moderate transcription and low coding potential in the sequence. For example, in this work I was able to characterize the ncRNA transcribed from the CTR but several other ncRNA may play important roles in the physiology of *T. brucei*. In the latter case, it is possible that these CDSs are translated in other life cycle stages like epimatigotes, metacyclics or stumpy BF. It would be worthy to perform ribosome profiling in all these stages as long as enough mRNA and footprint material could be obtained. This effort may produce a complete picture of the translational

landscape in *T. brucei*. Ideally, the material extracted should come from parasites present in mammalian hosts and tsetse vectors as was partially done by Jensen *et al.* in animal-derived slender BF (Jensen *et al.*, 2014). Finally, it is important to mention the possibility that many of the CDSs found to be not translated in my analysis, may be translated under different stress conditions. Again, ideally the ribosome profiling experiments have to be performed in an *in vivo* model to reproduce as accurately as possible the conditions that affect the parasite in the insect vector and the mammalian host. The different life cycle stages of *T. brucei* live under dissimilar conditions of temperature (mammalian host [37°C] vs tsetse fly [27°C]), availability of nutrients and environment (bloodstream vs digestive tract). For example, it is known that heat shock induces changes in translation, which is sometimes controlled by uORFs (Shalgi *et al.*, 2013).

Recently, it has become clear that small peptides play important roles in different cells and under different conditions (Andrews and Rothnagel, 2014). Ribosome profiling can provide valuable information about the small proteome potentially coded from the *T. brucei* genome. This methodology offers as well the possibility to study the translational elongation speed that, at least in mice, has been shown is affected by determinants of the coding sequence such as codon adaptation, local mRNA folding and the local charge of amino acids (Dana and Tuller, 2012). Finally, other levels of gene control that can be studied with ribosome profiling include out-of-frame internal initiation for induction of nonsense-mediated mRNA decay (Arribere and Gilbert, 2013) and the sequential and coordinated co-translational folding of proteins (Han *et al.*, 2012).

## **A polyadenylated, non-spliced ncRNA is transcribed from the CTR in the Lister 427 strain of *T. brucei***

RNA-seq datasets enable the detection of changes in basal transcription levels and ribosome profiling enables to determine the coding potential for those regions. Therefore, both datasets can be combined to predict putative ncRNAs. Given that transcription in PTUs is constitutive (including the active BES), it is possible that some intergenic regions may correspond to ncRNAs. Our analysis showed clearly the lack of translation in the CTR but at the same time a moderate transcription in this region compared to the basal transcription levels observed in non-coding regions of the active BES. The characterization of the transcript potential of the CTR allowed me to identify a single transcript of ~1200 nt in length. This transcript contains a poly(A) tail at the 3'-end but not the SL, which is present at the 5'-end of all mRNAs. The coding potential of the region is low and, as mentioned before, there is no evidence of translation by the ribosomal machinery. However, this characterization does not solve the question whether the CTR is involved in the regulation of *in situ* switching and if the proposed stabilizing element is of RNA or DNA nature.

The first step to elucidate its role would be to repeat the experiment performed by Davies *et al.* in 1997 by deleting part of the CTR. This can be performed as they did, that is by introducing a reporter gene (*VSG-117* or luciferase) and a resistant marker. However, I have been able to implement the CRISPR-Cas9 system in PF and at the moment of writing this thesis I am trying to adapt the system to BF. The CRISPR-Cas9 system allows the editing of genomic regions without the need of introducing reporters and selection markers. This

advantage cannot be underestimated given the importance of the active BES for the survival of the parasite and the introduction of exogenous sequences in the active BES might involve the death of the cells or the activation of a new BES.

The episomes derived from the pEV series developed by Patnaik *et al.* (Patnaik et al., 1993, 1994), work well in PF but not in BF. They developed as well the episome pT11bs that is almost identical to pT13-11 and pEV but the PAS, SAS and UTRs flanking the G418 resistant cassette are the ones from the fructose-bisphosphate aldolase gene. This small change surprisingly enabled the pT11bs to behave like an episome in BF and to be kept in this life cycle stage for several weeks (Patnaik et al., 1993), (Dr. Christian Janzen, personal communication). At the time of writing this thesis, I obtained the pT11bs episomes containing the resistant markers G418 or Phleomycin (kind gift from Dr. Christian Janzen). In order to test the system as fast as possible, the idea is to introduce the SCD6-GFP cassette in one of the episomes. The Cas9v2 might be expressed from a second episome or from a copy stably integrated into the 2T1 cell line. The first possibility is going to be challenging because the transfection efficiency in BF is lower than in PF (Burkard et al., 2007) and I don't know whether the two episomes can be maintained simultaneously in BF cells. According to Patnaik's reports (Patnaik et al., 1993), the pT11bs episome is less stable than the pT13-11 that worked so well in PF. For this reason, the second approach would be more feasible. In this approach, pRPa plasmids developed by Alsford and Horn (Alsford and Horn, 2008) may be used to transfect the 2T1 cell line. pRPa plasmids allow the integration at a landing pad locus in a ribosomal spacer on chromosome IIa. In this way, the Cas9v2 can be stably integrated into

the genome of BF cells and its expression induced with doxycycline. In case that I can reproduce the results obtained with the SCD6-GFP system in PF, the next step will be the integration in pT11bs of a repair template containing a truncated version of the CTR sequence (pT11bs\_CTR\_sgRNA). The 2T1 cells expressing the Cas9v2 will be transfected with pT11bs\_CTR\_sgRNA and exposed to drug pressure according to the time window the cells be resistant to the antibiotics. Then, the cells are going to be plated out in limited dilutions in order to obtain clonal populations. Some of these clones are going to be tested by PCR for the deletion of the CTR.

In case this strategy proves to be successful, it will be necessary to measure the effect of the deletion on the *in situ* switching rate. According to Davies *et al.*, upon deletion of the CTR, there was a 100-fold increase in the rate of switching. If the reported rate of switching in laboratory-adapted BF is  $\sim 1 \times 10^{-7}$  per generation, the CTR deletion may increase the number to  $\sim 1 \times 10^{-5}$ . Therefore, it will be necessary to deplete the culture of the cells that still express the VSG-221. The depletion has been done adding guinea-pig serum containing antibodies anti-VSG-221 to cells plated in 96-well plates (Davies *et al.*, 1997; Engstler *et al.*, 2007). Nevertheless, depending on the quality and to batch-to-batch differences of the serum, the depletion of the VSG-221-expressing cells is not always complete, which in turn affects the reproducibility of this approach. Another possibility is based in the addition of rabbit anti-VSG-221 serum without complement (Boothroyd *et al.*, 2009). After the addition of the serum, the cells are washed with HMI-11 medium and incubated with goat anti-rabbit microbeads, washed again with HMI-11 medium and applied to a MidiMACS

Separator Column (Miltenyi Biotech). In this way, the majority of VSG-221-expressing cells are retained in the column and the cells expressing other VSGs are recovered in the flow-through fraction (Boothroyd et al., 2009). An alternative approach is to lyse the cells and deplete the VSG-221 RNA instead of the whole parasites. The resulting lysate is therefore added to streptavidin-coupled Dynabeads and biotinylated oligos complementary to the VSG-221 mRNA. In this way, the mRNAs from other VSGs should be present in the fraction that did not bind to the complementary oligos.

The resulting mRNA will be extracted and used to prepare libraries suitable for NGS. The PCR step will be performed using primers designed in such a way that only VSG transcripts will be amplified. To obtain such enrichment of VSG transcripts, the forward primer will have to anneal in the SL and the reverse primer to a region present close to the 3'-end of the CDS that is conserved in most of the VSGs. NGS can provide quantitative information as long as spiked-in control transcripts are included in the preparation of the library. Comparison of the spiked-in data to known spike-in concentrations and ratios would allow to reliably quantify the abundance of each VSG species.

The purpose of the experiments mentioned above is to answer the question: Is there an increase in the transcription levels of the VSGs located in other BES after the deletion of the CTR present in the active BES? In case the answer is negative, the specific strategy used by Davis *et al.* would have to be repeated in order to see if the introduction of exogenous elements instead of the CTR deletion generated the increase in the *in situ* switching. This means to delete exactly the same region in the CTR and to include the same construct containing

VSG-117 or luciferase and the resistant marker. Then, the depletion assay would have to be repeated as similar as possible, which could be difficult given the differences in the quality of the serum used almost 20 years ago.

In case the answer is affirmative, it would be necessary to perform a variety of assays in order to identify the role played by the CTR. Once the CRISPR-Cas9 system is implemented in BF, it will be possible to systematically delete regions in the CTR to map the localization of the minimal element required for BES stability. More rigorous bioinformatics analysis should be performed to complement the CTR mapping and to find common motifs in the different CTR sequences present in the other BES and chromosome internal regions. In order to identify the putative target(s) of the CTR transcript(s), an RNA-pulldown assay might be performed for RNA-protein interactions or ChIRP (Chromatin isolation by RNA purification). The latter would allow the identification of chromatin factors associated to the CTR ncRNAs (Chu et al., 2012). In addition, the location of the ncRNA should be determined by RNA-FISH; and the location of the CTR transcript and its putative partner(s) by CHART (Capture Hybridization Analysis of RNA Targets) whereby genomic binding sites of a specific ncRNA can be detected (Simon, 2013). These experiments may be complemented by determining the nucleosomal density in the CTR and by DNA FISH analysis to know the relative position of the CTR regarding the active promoter and other regions of the BES. Finally, the role and structure of other CTRs should not be forgotten. Therefore, besides the deletion of the CTR present in the active BES1, it would be interesting to delete CTRs present in cell lines that actively transcribe other BESs. Furthermore, in case the



CTR transcript plays an active role in *in situ* switching, it would be necessary to address what happens if these ncRNAs are overexpressed or if two different CTRs are transcribed simultaneously. For example, in my thesis work I cloned the CTR sequences present upstream of the *VSG-3*, *VSG-6* and *VSG-13* genes in the pRPa vector for their overexpression in 2T1 cells. Then I extracted the total RNA and performed qPCRs using primers specific to these sequences. Nevertheless, after analyzing the data I did not find strong evidence of an increase in transcription for *VSG-3*, *VSG-6* and *VSG-13*. I performed the experiments only twice so I cannot totally rule out the possibility of an increase in transcription for those *VSGs*. In addition, it is possible that qPCR is not sensitive enough for the detection of a moderate increase in the transcription levels for the aforementioned *VSGs*. Therefore, it would be necessary to analyze more replicates and to evaluate the transcription levels by NGS.

It would be interesting as well to observe the location of the CTR transcript compared to the location of the active BES. One way to perform that would be the insertion of a reporter system upstream of the BES promoter. In order to magnify the signal to be visible by microscopy, it is possible to modify non-transcribed regions upstream the active BES promoter with arrays of lac operators and in turn express a lacR-GFP fusion protein. If feasible, it would be possible to observe the active BES in the ESB and together with a RNA-FISH approach, to determine the localization of the CTR transcript.

## **The CRISPR-Cas9 system allows the efficient genome editing in procyclic forms**

As I mentioned in the previous section, the CRISPR-Cas9 system might be used in BF to delete the CTR and specific parts of its sequence in order to determine the regions that confer the stabilizing function. The implementation of the CRISPR-Cas9 technology in BF would open exciting possibilities for genome editing given the biological importance of the BF and the fact that many researchers have focused their work in this life cycle stage.

In the chapter concerning the CRISPR-Cas9 system, I mentioned the possibility of substituting the K4 codon in the histone H4 gene with an R4. Given the high number of H4 gene copies (at least 34), it is possible that extending the drug selection process might lead to the increase in the number of modified copies. In *T. cruzi*, the  $\beta$ -galactofuranosyl glycosyltransferase ( $\beta$ -GalGT) family of 65 annotated genes was targeted with three sgRNAs and flow cytometric analysis showed the reduction in surface  $\beta$ -galactosyl residues (Peng et al., 2015). The authors introduced the sgRNAs in three sequential transfections and observed a progressive reduction in the residues after each round. This result prompted me to follow a similar approach for H4 in case the proposed strategy failed to target a significant number of copies of the tandem array. Histone acetylation and possibly of other PTMs plays an important role for the epigenetic control of gene expression in *T. brucei* and therefore the successful editing of these marks or “histone codes” might allow the identification of the factors involved in the deposition of histones and the recruitment of chromatin remodelers and other associated factors.

The CRISPR-Cas9 system has shown to be a rapid and versatile approach for editing CDSs, ncRNAs, UTRs, promoters, enhancers and other genomic features without the unnecessary addition of resistance markers in the target region. With this approach it is now possible to perform *in situ* tagging of genes in *T. brucei*, keeping the endogenous UTRs intact, as was shown when the SCD6 CDS was tagged at the 3'-end with the eGFP sequence. Indeed, it is now possible to evaluate the impact of UTRs in the regulation of RNA stability and translational efficiency *in situ*, which was virtually impossible with the methods previously available. Furthermore, the endogenous UTRs of a specific gene may now be replaced for those UTRs comprehensively studied such as fructose-bisphosphate aldolase or PARP in a marker-free way. Another possibility would be to delete part of the UTR for the identification of motifs that can play regulatory roles or introduce active motifs in other genes downregulated in a specific life cycle stage. Besides, in the ribosome profiling study we showed the presence of several canonical and non-canonical uORF and the CRISPR-Cas9 platform can be used to precisely delete or modify them in order to obtain insight about the putative role of these short sequences in the control of translation initiation of the main CDS.

Other applications of the CRISPR-Cas9 system to be developed in the future include the multiplexing capacity for genome-wide functional screens and the use of dCas9 (Cas9 with no endonuclease activity) as a transcriptional activator or repressor. Traditional techniques for labeling DNA, such as FISH are unable to capture live processes. The use of fluorescently tagged Cas9 to target

specific DNA loci is a technique already developed for live-cell-imaging (Chen et al., 2013).

## **List of collaborators**

### **Ribosome profiling**

Dr. Tim Nicolai Siegel conceived and supervised the project. He also contributed to the data analyses and to write the paper.

Dr. Chung-Chau Hon contributed to the data analyses.

Dr. Jens T. Vanselow and Dr. Andreas Schlosser contributed to the proteomic analyses.

### **CRISPR-Cas9 system**

MSc. Carolin Wedel generated the Cas9 constructs and performed the IFs to evaluate the cellular localization of the endonuclease.

Dr. Raul Cosentino contributed to the off-target analysis.

## References (Literaturverzeichnis)

Aline, R., MacDonald, G., Brown, E., Allison, J., Myler, P., Rothwell, V., and Stuart, K. (1985). (TAA)<sub>n</sub> within sequences flanking several intrachromosomal variant surface glycoprotein genes in *Trypanosoma brucei*. *Nucleic Acids Res.* *13*, 3161–3177.

Aline, R.F., Scholler, J.K., and Stuart, K. (1989). Transcripts from the co-transposed segment of variant surface glycoprotein genes are in *Trypanosoma brucei* polyribosomes. *Mol. Biochem. Parasitol.* *32*, 169–178.

Alsford, S., and Horn, D. (2008). Single-locus targeting constructs for reliable regulated RNAi and transgene expression in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *161*, 76–79.

Alsford, S., and Horn, D. (2012). Cell-cycle-regulated control of VSG expression site silencing by histones and histone chaperones ASF1A and CAF-1b in *Trypanosoma brucei*. *Nucleic Acids Res.* *40*, 10150–10160.

Alsford, S., Turner, D.J., Obado, S.O., Sanchez-Flores, A., Glover, L., Berriman, M., Hertz-Fowler, C., and Horn, D. (2011). High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Res.* *21*, 915–924.

Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* *14*, 35–48.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* *11*, R106.

- Andreev, D.E., O'Connor, P.B.F., Fahey, C., Kenny, E.M., Terenin, I.M., Dmitriev, S.E., Cormican, P., Morris, D.W., Shatsky, I.N., and Baranov, P. V (2015). Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* 4, e03971.
- Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* 15, 193–204.
- Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3889–3894.
- Archer, S.K., Luu, V.-D., de Queiroz, R.A., Brems, S., and Clayton, C. (2009). *Trypanosoma brucei* PUF9 Regulates mRNAs for Proteins Involved in Replicative Processes over the Cell Cycle. *PLoS Pathog.* 5, e1000565.
- Arribere, J.A., and Gilbert, W. V (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.* 23, 977–987.
- Aurrecochea, C., Barreto, A., Brestelli, J., Brunk, B.P., Cade, S., Doherty, R., Fischer, S., Gajria, B., Gao, X., Gingle, A., et al. (2013). EuPathDB: The Eukaryotic Pathogen database. *Nucleic Acids Res.* 41, D684–D691.
- Bae, S., Park, J., and Kim, J.-S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* 30, 1473–1475.
- Bakheet, T., Williams, B.R.G., and Khabar, K.S.A. (2003). ARED 2.0: an update of

AU-rich element mRNA database. *Nucleic Acids Res.* *31*, 421–423.

Barnes, R.L., and McCulloch, R. (2007). Trypanosoma brucei homologous recombination is dependent on substrate length and homology, though displays a differential dependence on mismatch repair as substrate length decreases. *Nucleic Acids Res.* *35*, 3478–3493.

Barrett, B., LaCount, D.J., and Donelson, J.E. (2004). Trypanosoma brucei: a first-generation CRE-loxP site-specific recombination system. *Exp. Parasitol.* *106*, 37–44.

Barry, D., McCulloch, R., Mottram, J., and Acosta-Serrano, A. (2007). *Trypanosomes After the Genome* (Horizon Bioscience, Wymondham, Norfolk, UK). p. 423.

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science* *336*, 233–237.

Benne, R., Van Den Burg, J., Brakenhoff, J.P.J.J., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* *46*, 819–826.

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* *10*, 973–982.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renault, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., et al. (2005). The genome of



the African trypanosome *Trypanosoma brucei*. *Science* 309, 416–422.

Bhaya, D., Davison, M., and Barrangou, R. (2011). CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. *Annu. Rev. Genet.* 45, 273–297.

Boothroyd, C.E., Dreesen, O., Leonova, T., Ly, K.I., Figueiredo, L.M., Cross, G.A.M., and Papavasiliou, F.N. (2009). A yeast-endonuclease-generated DNA break induces antigenic switching in *Trypanosoma brucei*. *Nature* 459, 278–281.

Boutros, M., and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. *Nat Rev Genet* 9, 554–566.

Brar, G.A., and Weissman, J.S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 16, 651–664.

Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T., and Weissman, J.S. (2012). High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science*. 335, 552–557.

Brecht, M., and Parsons, M. (1998). Changes in polysome profiles accompany trypanosome development. *Mol. Biochem. Parasitol.* 97, 189–198.

Brown, K.H., Brentano, S.T., and Donelson, J.E. (1986). Mung bean nuclease cleaves preferentially at the boundaries of variant surface glycoprotein gene transpositions in trypanosome DNA. *J. Biol. Chem.* 261, 10352–10358.

Bruce, D. (1895). Preliminary report on the tsetse fly disease or nagana in Zululand. (Bennett & Davis), p. 36.

Bruderer, T., Tu, L.-C., and Lee, M.G.-S. (2003). The 5' end structure of transcripts

derived from the rRNA gene and the RNA polymerase I transcribed protein coding genes in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *129*, 69–77.

Brun, R., and Schönenberger, C. (1979). Cultivation and in vitro cloning or procyclic culture forms of *Trypanosoma brucei* in a semi-defined medium. Short communication. *Acta Trop* *36*, 289–292.

Burkard, G., Fragoso, C.M., and Roditi, I. (2007). Highly efficient stable transformation of bloodstream forms of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *153*, 220–223.

Burton, P., McBride, D.J., Wilkes, J.M., Barry, J.D., and McCulloch, R. (2007). Ku Heterodimer-Independent End Joining in *Trypanosoma brucei* Cell Extracts Relies upon Sequence Microhomology. *Eukaryot. Cell* *6*, 1773–1781.

Butter, F., Bucerius, F., Michel, M., Cicova, Z., Mann, M., and Janzen, C.J. (2013). Comparative Proteomics of Two Life Cycle Stages of Stable Isotope-labeled *Trypanosoma brucei* Reveals Novel Components of the Parasite's Host Adaptation Machinery. *Mol. Cell. Proteomics* *12*, 172–179.

Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 7507–7512.

Camby, I. (2006). Galectin-1: a small protein with major functions. *Glycobiology* *16*, 137R – 157R.

Campbell, D.A., van Bree, M.P., and Boothroyd, J.C. (1984). The 5'-limit of transposition and upstream barren region of a trypanosome VSG gene: tandem

76 base-pair repeats flanking (TAA)<sup>90</sup>. *Nucleic Acids Res.* *12*, 2759–2774.

Capewell, P., Monk, S., Ivens, A., MacGregor, P., Fenn, K., Walrad, P., Bringaud, F., Smith, T.K., and Matthews, K.R. (2013). Regulation of *Trypanosoma brucei* Total and Polysomal mRNA during Development within Its Mammalian Host. *PLoS One* *8*, e67069.

Carlson, D.F., Tan, W., Lillico, S.G., Stverakova, D., Proudfoot, C., Christian, M., Voytas, D.F., Long, C.R., Whitelaw, C.B.A., and Fahrenkrug, S.C. (2012). Efficient TALEN-mediated gene knockout in livestock. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 17382–17387.

Caro, F., Ahyong, V., Betegon, M., and DeRisi, J.L. (2014). Genome-wide regulatory dynamics of translation in the *Plasmodium falciparum* asexual blood stages. *Elife* *3*, e04106.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* *155*, 1479–1491.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. (2005). Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science*. *308*, 1149–1154.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., and Kim, J.-S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* *24*, 132–141.

- Chu, C., Quinn, J., and Chang, H.Y. (2012). Chromatin Isolation by RNA Purification (ChIRP). *J. Vis. Exp.* 3912.
- Clancy, S. (2008). RNA splicing: introns, exons and spliceosome. *Nat. Educ.* 1, 31.
- Clayton, C.E. (2002). Life without transcriptional control? From fly to man and back again. *EMBO J.* 21, 1881–1888.
- Clayton, C.E. (2014). Networks of gene expression regulation in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 195, 96–106.
- Clayton, C., and Estevez, A. (2010). The Exosomes of Trypanosomes and Other Protists. In *RNA Exosome SE - 4*, T. Jensen, ed. (Springer US), pp. 39–49.
- Clayton, C.E., and Michels, P. (1996). Metabolic compartmentation in African trypanosomes. *Parasitol. Today* 12, 465–471.
- Cliffe, L.J., Siegel, T.N.N., Marshall, M., Cross, G. A. M., and Sabatini, R. (2010). Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of *Trypanosoma brucei*. *Nucleic Acids Res.* 38, 3923–3935.
- Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 339, 819–823.
- Consortium, T.E.P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.
- Consortium, R.G.E.R.G. and G.S.G. (Genome N.P.C.G. and the F., Katayama, S.,

Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., et al. (2005a). Antisense Transcription in the Mammalian Transcriptome. *Science*. *309*, 1564–1566.

Consortium, T.F., Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., et al. (2005b). The Transcriptional Landscape of the Mammalian Genome. *Science*. *309*, 1559–1563.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* *26*, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A, Olsen, J. V, and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* *10*, 1794–1805.

Cross, G. (1975). Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of *Trypanosoma brucei*. *Parasitology Dec*; *71*, 393–417.

Cross, G. A. M., Kim, H.-S., and Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol. Biochem. Parasitol.* *195*, 59–73.

Dana, A., and Tuller, T. (2012). Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. *PLoS Comput. Biol.* *8*, e1002755.

Daniels, J.-P., Gull, K., and Wickstead, B. (2010). Cell biology of the trypanosome

genome. *Microbiol. Mol. Biol. Rev.* *74*, 552–569.

DaRocha, W.D., Otsu, K., Teixeira, S.M.R., and Donelson, J.E. (2004). Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. *Mol. Biochem. Parasitol.* *133*, 175–186.

Davies, K.P., Carruthers, V.B., and Cross, G. A (1997). Manipulation of the vsg co-transposed region increases expression-site switching in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *86*, 163–177.

de Klerk, E., Fokkema, I.F.A.C., Thiadens, K.A.M.H., Goeman, J.J., Palmblad, M., den Dunnen, J.T., von Lindern, M., and 't Hoen, P.A.C. (2015). Assessing the translational landscape of myogenic differentiation by ribosome profiling. *Nucleic Acids Res.* *43*, 4408–4428.

Deitsch, K.W., Lukehart, S. A., and Stringer, J.R. (2009). Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat. Rev. Microbiol.* *7*, 493–503.

Denninger, V., and Rudenko, G. (2014). FACT plays a major role in histone dynamics affecting *VSG* expression site control in *Trypanosoma brucei*. *Mol. Microbiol.* *94*, 945–962.

van Deursen, F.J., Shahi, S.K., Turner, C.M., Hartmann, C., Guerra-Giraldez, C., Matthews, K.R., and Clayton, C.E. (2001). Characterisation of the growth and differentiation in vivo and in vitro-of bloodstream-form *Trypanosoma brucei* strain TREU 927. *Mol. Biochem. Parasitol.* *112*, 163–171.

Dever, T.E., Feng, L., Wek, R.C., Cigan, A.M., Donahue, T.F., and Hinnebusch, A.G. (1992). Phosphorylation of initiation factor 2 $\alpha$  by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* 68, 585–596.

Dhalia, R., Reis, C.R.S., Freire, E.R., Rocha, P.O., Katz, R., Muniz, J.R.C., Standart, N., and de Melo Neto, O.P. (2005). Translation initiation in *Leishmania major*: characterisation of multiple eIF4F subunit homologues. *Mol. Biochem. Parasitol.* 140, 23–41.

Dhalia, R., Marinsek, N., Reis, C.R.S., Katz, R., Muniz, J.R.C., Standart, N., Carrington, M., and de Melo Neto, O.P. (2006). The two eIF4A helicases in *Trypanosoma brucei* are functionally distinct. *Nucleic Acids Res.* 34, 2495–2507.

Droll, D., Minia, I., Fadda, A., Singh, A., Stewart, M., Queiroz, R., and Clayton, C. (2013). Post-Transcriptional Regulation of the Trypanosome Heat Shock Response by a Zinc Finger Protein. *PLoS Pathog.* 9, e1003286.

Engstler, M., Thilo, L., Weise, F., Grünfelder, C.G., Schwarz, H., Boshart, M., and Overath, P. (2004). Kinetics of endocytosis and recycling of the GPI-anchored variant surface glycoprotein in *Trypanosoma brucei*. *J. Cell Sci.* 117, 1105–1115.

Engstler, M., Pfohl, T., Herminghaus, S., Boshart, M., Wiegertjes, G., Heddergott, N., and Overath, P. (2007). Hydrodynamic Flow-Mediated Protein Sorting on the Cell Surface of Trypanosomes. *Cell* 131, 505–515.

Erben, E.D., Fadda, A., Lueong, S., Hoheisel, J.D., and Clayton, C. (2014). A genome-wide tethering screen reveals novel potential post-transcriptional regulators in *Trypanosoma brucei*. *PLoS Pathog.* 10, e1004178.

- Ericson, M., Janes, M. A., Butter, F., Mann, M., Ullu, E., and Tschudi, C. (2014). On the extent and role of the small proteome in the parasitic eukaryote *Trypanosoma brucei*. *BMC Biol.* *12*, 14.
- Evans, D.A., and Brown, R.C. (1972). The Utilization of Glucose and Proline by Culture Forms of *Trypanosoma brucei*. *J. Protozool.* *19*, 686–690.
- Fadda, A., Färber, V., Droll, D., and Clayton, C. (2013). The roles of 3'-exoribonucleases and the exosome in trypanosome mRNA degradation. *RNA* *19*, 937–947.
- Fadda, A., Ryten, M., Droll, D., Rojas, F., Färber, V., Haanstra, J.R., Merce, C., Bakker, B.M., Matthews, K., and Clayton, C. (2014). Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. *Mol. Microbiol.* *94*, 307–326.
- Fairlamb, A.H., and Opperdoes, F.R. (1986). Carbohydrate Metabolism in Cultured Cells. M.J. Morgan, ed. (Boston, MA: Springer US), pp. 183–224.
- Fenn, K., and Matthews, K.R. (2007). The cell biology of *Trypanosoma brucei* differentiation. *Curr. Opin. Microbiol.* *10*, 539–546.
- Ferguson, M. A., Homans, S.W., Dwek, R. A., and Rademacher, T.W. (1988). Glycosyl-phosphatidylinositol moiety that anchors *Trypanosoma brucei* variant surface glycoprotein to the membrane. *Science.* *239*, 753–759.
- Fernandes, A P., Nelson, K., and Beverley, S.M. (1993). Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of



parasitism. *Proc. Natl. Acad. Sci. U. S. A.* *90*, 11608–11612.

Fernández-Moya, S.M., Carrington, M., and Estévez, A.M. (2014). A short RNA stem-loop is necessary and sufficient for repression of gene expression during early logarithmic phase in trypanosomes. *Nucleic Acids Res.* *42*, 7201–7209.

Fèvre, E.M., Wissmann, B. V., Welburn, S.C., and Lutumba, P. (2008). The Burden of Human African Trypanosomiasis. *PLoS Negl. Trop. Dis.* *2*, e333.

Figueiredo, L.M., and Cross, G.A.M. (2010). Nucleosomes Are Depleted at the VSG Expression Site Transcribed by RNA Polymerase I in African Trypanosomes. *Eukaryot. Cell* *9*, 148–154.

Figueiredo, L.M., Janzen, C.J., and Cross, G. A. M. (2008). A Histone Methyltransferase Modulates Antigenic Variation in African Trypanosomes. *PLoS Biol.* *6*, e161.

Fletcher, J.C. (1999). Signaling of Cell Fate Decisions by CLAVATA3 in Arabidopsis Shoot Meristems. *Science.* *283*, 1911–1914.

Florent, I., Baltz, T., Raibaud, A., and Eisen, H. (1987). On the role of repeated sequences 5' to variant surface glycoprotein genes in African trypanosomes. *Gene* *53*, 55–62.

Freire, E.R., Dhalia, R., Moura, D.M.N., da Costa Lima, T.D., Lima, R.P., Reis, C.R.S., Hughes, K., Figueiredo, R.C.B.Q., Standart, N., Carrington, M., et al. (2011). The four trypanosomatid eIF4E homologues fall into two separate groups, with distinct features in primary sequence and biological properties. *Mol. Biochem. Parasitol.* *176*, 25–36.

Freire, E.R., Malvezzi, A.M., Vashisht, A.A., Zuberek, J., Saada, E.A., Langousis, G., Nascimento, J.D.F., Moura, D., Darzynkiewicz, E., Hill, K., et al. (2014).

Trypanosoma brucei Translation Initiation Factor Homolog EIF4E6 Forms a Tripartite Cytosolic Complex with EIF4G5 and a Capping Enzyme Homolog. *Eukaryot. Cell* 13, 896–908.

Freyman, D., Down, J., Carrington, M., Roditi, I., Turner, M., and Wiley, D. (1990).

2.9 A resolution structure of the N-terminal domain of a variant surface glycoprotein from Trypanosoma brucei. *J. Mol. Biol.* 216, 141–160.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* 31, 822–826.

Furger, A, Schürch, N., Kurath, U., and Roditi, I. (1997). Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of Trypanosoma brucei by modulating RNA stability and translation. *Mol. Cell. Biol.* 17, 4372–4380.

Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31, 397–405.

Gao, Y., and Zhao, Y. (2014). Self-processing of ribozyme-flanked RNAs into guide RNAs *in vitro* and *in vivo* for CRISPR-mediated genome editing. *J. Integr. Plant Biol.* 56, 343–349.

Garneau, J.E., Dupuis, M.-E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468,

67–71.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* *109*, E2579–E2586.

Gaudenzi, J. De, Frasch, A.C., and Clayton, C. (2005). RNA-Binding Domain Proteins in Kinetoplastids : a Comparative Analysis. *Eukaryot. Cell* *4*, 2106–2114.

Gerashchenko, M. V, Lobanov, A. V, and Gladyshev, V.N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 17394–17399.

Geurts, A.M., Cost, G.J., Freyvert, Y., Zeitler, B., Miller, J.C., Choi, V.M., Jenkins, S.S., Wood, A., Cui, X., Meng, X., et al. (2009). Knockout Rats Produced Using Designed Zinc Finger Nucleases. *Science* *325*, 433.

Ghorbal, M., Gorman, M., Macpherson, C.R., Martins, R.M., Scherf, A., and Lopez-Rubio, J.-J. (2014). Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. *Nat. Biotechnol.* *32*, 819–821.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* *154*, 442–451.

Gilles, A.F., and Averof, M. (2014). Functional genetics for all: engineered nucleases, CRISPR and the gene editing revolution. *Evodevo* *5*, 43.

Glover, L., and Horn, D. (2014). Locus-specific control of DNA resection and

suppression of subtelomeric VSG recombination by HAT3 in the African trypanosome. *Nucleic Acids Res.* *42*, 12600–12613.

de Godoy, L.M.F., Olsen, J. V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* *455*, 1251–1254.

Gunasekera, K., Wüthrich, D., Braga-Lagache, S., Heller, M., and Ochsenreiter, T. (2012). Proteome remodelling during development from blood to insect-form *Trypanosoma brucei* quantified by SILAC and mass spectrometry. *BMC Genomics* *13*, 556.

Günzl, A., Bruderer, T., Laufer, G., Schimanski, B., Tu, L.-C., Chung, H.-M., Lee, P.-T., and Lee, M.G.-S. (2003). RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot. Cell* *2*, 542–551.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* *154*, 240–251.

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V.C., Graveley, B.R., Terns, R.M., et al. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* *45*, 292–302.

Hall, J.P.J., Wang, H., and Barry, J.D. (2013). Mosaic VSGs and the Scale of *Trypanosoma brucei* Antigenic Variation. *PLoS Pathog.* *9*, e1003502.

Han, Y., David, A., Liu, B., Magadán, J.G., Bennink, J.R., Yewdell, J.W., and Qian, S.-B. (2012). Monitoring cotranslational protein folding in mammalian cells at codon resolution. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 12467–12472.

Hastings, M.L., and Krainer, A.R. (2001). Pre-mRNA splicing in the new millennium. *Curr. Opin. Cell Biol.* *13*, 302–309.

Hehl, A, Vassella, E., Braun, R., and Roditi, I. (1994). A conserved stem-loop structure in the 3' untranslated region of procyclin mRNAs regulates expression in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* *91*, 370–374.

Helm, J., Wilson, M.E., and Donelson, J.E. (2008). Different trans RNA splicing events in bloodstream and procyclic *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *159*, 134–137.

Helm, J.R., Wilson, M.E., and Donelson, J.E. (2009). Differential expression of a protease gene family in African trypanosomes. *Mol. Biochem. Parasitol.* *163*, 8–18.

Hertz-Fowler, C., Figueiredo, L.M., Quail, M. A., Becker, M., Jackson, A., Bason, N., Brooks, K., Churcher, C., Fahkro, S., Goodhead, I., et al. (2008). Telomeric Expression Sites Are Highly Conserved in *Trypanosoma brucei*. *PLoS One* *3*, e3527.

Hinnebusch, A.G. (1993). Gene-specific translational control of the yeast GCN4 gene by phosphorylation of eukaryotic initiation factor 2. *Mol. Microbiol.* *10*, 215–223.

Hoek, M., Engstler, M., and Cross, G.A. (2000). Expression-site-associated gene 8

(ESAG8) of *Trypanosoma brucei* is apparently essential and accumulates in the nucleolus. *J. Cell Sci.* *113*, 3959–3968.

Hood, H.M., Neafsey, D.E., Galagan, J., and Sachs, M.S. (2009). Evolutionary Roles of Upstream Open Reading Frames in Mediating Gene Regulation in Fungi. *Annu. Rev. Microbiol.* *63*, 385–409.

Horn, D. (2008). Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics* *9*, 2.

Horn, D. (2014). Antigenic variation in African trypanosomes. *Mol. Biochem. Parasitol.* *195*, 123–129.

Hotz, H.R., Hartmann, C., Huober, K., Hug, M., and Clayton, C. (1997). Mechanisms of developmental regulation in *Trypanosoma brucei*: a polypyrimidine tract in the 3'-untranslated region of a surface protein mRNA affects RNA abundance and translation. *Nucleic Acids Res.* *25*, 3017–3026.

Hovel-Miner, G. A., Boothroyd, C.E., Mugnier, M., Dreesen, O., Cross, G. A. M., and Papavasiliou, F.N. (2012). Telomere Length Affects the Frequency and Mechanism of Antigenic Variation in *Trypanosoma brucei*. *PLoS Pathog.* *8*, e1002900.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotech* *31*, 827–832.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* *157*, 1262–1278.

<http://code.google.com/p/cutadapt/> Retrieved in November 2013. Cutadapt.

<http://spot.colorado.edu/~slin/cas9.html> Retrieved in December 2015. Jack Lin's CRISPR/Cas9 gRNA finder.

<http://tritrypdb.org> Retrieved in December 2015. TryTripDB Kinetoplastid Genomic resource.

<http://tritrypdb.org/tritrypdb/getDataset.do?display=detail> Retrieved in December 2015. TryTripDB Kinetoplastid Genomic resource.

<http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html> Retrieved in November 2013. Bioconductor. Open source software for bioinformatics.

<http://www.cdc.gov/parasites/sleepingsickness/biology.html> Retrieved in December 2015. CDC Centers for disease control and prevention.

<https://www.addgene.org/crispr/zhang/FAQ/> Retrieved in February 2016. Zhang Lab's CRISPR Frequently Asked Questions.

Huang, H., Yoon, H., Hannig, E.M., and Donahue, T.F. (1997). GTP hydrolysis controls stringent selection of the AUG start codon during translation initiation in *Saccharomyces cerevisiae*. *Genes Dev.* *11*, 2396–2413.

Hughes, K., Wand, M., Foulston, L., Young, R., Harley, K., Terry, S., Ersfeld, K., and Rudenko, G. (2007). A novel ISWI is involved in VSG expression site downregulation in African trypanosomes. *EMBO J.* *26*, 2400–2410.

Iacono, M., Mignone, F., and Pesole, G. (2005). uAUG and uORFs in human and rodent 5' untranslated mRNAs. *Gene* *349*, 97–105.

Ibrahim, B.S., Kanneganti, N., Rieckhof, G.E., Das, A., Laurents, D. V, Palenchar, J.B., Bellofatto, V., and Wah, D.A. (2009). Structure of the C-terminal domain of transcription factor IIB from *Trypanosoma brucei*. *Proc. Natl. Acad. Sci.* *106*, 13242–13247.

Imboden, M.A., Laird, P.W., Affolter, M., and Seebeck, T. (1987). Transcription of the intergenic regions of the tubulin gene cluster of *Trypanosoma brucei*: evidence for a polycistronic transcription unit in a eukaryote. *Nucleic Acids Res.* *15*, 7357–7368.

Ingolia, N.T. (2010). Genome-Wide Translational Profiling by Ribosome Footprinting. *Methods Enzymol.* (Elsevier Inc.). 470:119-42.

Ingolia, N.T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* *15*, 205–213.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science.* *324*, 218–223.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* *147*, 789–802.

Ingolia, N.T., Brar, G. A, Rouskin, S., McGeachy, A.M., and Weissman, J.S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* *7*, 1534–1550.

International Human Genome Sequencing Consortium (2004). Finishing the



euchromatic sequence of the human genome. *Nature* 431, 931–945.

Jackson, A.L., and Linsley, P.S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov* 9, 57–67.

Jackson, R.J., Hellen, C.U.T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* 11, 113–127.

Jang, C., Lahens, N.F., Hogenesch, J.B., and Sehgal, A. (2015). Ribosome profiling reveals an important role for translational control in circadian gene expression. *Genome Res.* 25, 1836-47.

Janzen, C.J., Hake, S.B., Lowell, J.E., and Cross, G.A.M. (2006a). Selective Di- or Trimethylation of Histone H3 Lysine 76 by Two DOT1 Homologs Is Important for Cell Cycle Regulation in *Trypanosoma brucei*. *Mol. Cell* 23, 497–507.

Janzen, C.J., Fernandez, J.P., Deng, H., Diaz, R., Hake, S.B., and Cross, G. A. M. (2006b). Unusual histone modifications in *Trypanosoma brucei*. *FEBS Lett.* 580, 2306–2310.

Jensen, B.C., Brekken, D.L., Randall, A.C., Kifer, C.T., and Parsons, M. (2005). Species Specificity in Ribosome Biogenesis : a Nonconserved Phosphoprotein Is Required for Formation of the Large Ribosomal Subunit in *Trypanosoma brucei*. *Eukariot. Cell* 4, 30–35.

Jensen, B.C., Sivam, D., Kifer, C.T., Myler, P.J., and Parsons, M. (2009). Widespread variation in transcript abundance within and across developmental stages of

*Trypanosoma brucei*. BMC Genomics 10, 482.

Jensen, B.C., Ramasamy, G., Vasconcelos, E.J.R., Ingolia, N.T., Myler, P.J., and Parsons, M. (2014). Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*. BMC Genomics 15, 1–21.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 337, 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife 2013, 1–9.

Johnson, P. (1987). Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. Cell 51, 273–281.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V (2014). Evolutionary conservation of long noncoding RNAs; sequence, structure, function. Biochim. Biophys. Acta 1840, 1063–1071.

Juntawong, P., Girke, T., Bazin, J., and Bailey-Serres, J. (2014). Translational dynamics revealed by genome-wide profiling of ribosome footprints in *Arabidopsis*. Proc. Natl. Acad. Sci. U. S. A. 111, E203–E212.

Kelly, S., Reed, J., Kramer, S., Ellis, L., Webb, H., Sunter, J., Salje, J., Marinsek, N.,

Gull, K., Wickstead, B., et al. (2007). Functional genomics in *Trypanosoma brucei*: a collection of vectors for the expression of tagged proteins from endogenous and ectopic gene loci. *Mol. Biochem. Parasitol.* *154*, 103–109.

Kim, H.-S., and Cross, G. A. M. (2010). TOPO3 $\alpha$  Influences Antigenic Variation by Monitoring Expression-Site-Associated VSG Switching in *Trypanosoma brucei*. *PLoS Pathog.* *6*, e1000992.

Kim, H.-S., Li, Z., Boothroyd, C., and Cross, G. A. M. (2013). Strategies to construct null and conditional null *Trypanosoma brucei* mutants using Cre-recombinase and loxP. *Mol. Biochem. Parasitol.* *191*, 16–19.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* *22*, 568–576.

Kolev, N.G., Franklin, J.B., Carmi, S., Shi, H., Michaeli, S., and Tschudi, C. (2010). The Transcriptome of the Human Pathogen *Trypanosoma brucei* at Single-Nucleotide Resolution. *PLoS Pathog.* *6*, e1001090.

Kolev, N.G., Tschudi, C., and Ullu, E. (2011). RNA Interference in Protozoan Parasites: Achievements and Challenges. *Eukaryot. Cell* *10*, 1156–1163.

Kolev, N.G., Ullu, E., and Tschudi, C. (2014). The emerging role of RNA-binding proteins in the life cycle of *Trypanosoma brucei*. *Cell. Microbiol.* *16*, 482–489.

Kong, J., and Lasko, P. (2012). Translational control in cellular and developmental processes. *Nat Rev Genet* *13*, 383–394.

Kooter, J.M., van der Spek, H.J., Wagter, R., d'Oliveira, C.E., van der Hoeven, F., Johnson, P.J., and Borst, P. (1987). The anatomy and transcription of a telomeric expression site for variant-specific surface antigens in *T. brucei*. *Cell* 51, 261–272.

Koumandou, V.L., Natesan, S., Sergeenko, T., and Field, M.C. (2008). The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. *BMC Genomics* 9, 1–28.

Kramer, S., Marnef, A., Standart, N., and Carrington, M. (2012). Inhibition of mRNA maturation in trypanosomes causes the formation of novel foci at the nuclear periphery containing cytoplasmic regulators of mRNA fate. *J. Cell Sci.* 125, 2896–2909.

Kruger, T., Hofweber, M., and Kramer, S. (2013). SCD6 induces ribonucleoprotein granule formation in trypanosomes in a translation-independent manner, regulated by its LSM and RGG domains. *Mol. Biol. Cell* 24, 2098–2111.

Lahav, T., Sivam, D., Volpin, H., Ronen, M., Tsigankov, P., Green, a., Holland, N., Kuzyk, M., Borchers, C., Zilberstein, D., et al. (2011). Multiple levels of gene regulation mediate differentiation of the intracellular pathogen *Leishmania*. *FASEB J.* 25, 515–525.

Lamont, G.S., Tucker, R.S., and Cross, G.A.M. (1986). Analysis of antigen switching rates in *Trypanosoma brucei*. *Parasitology* 92, 355–367.

Lander, N., Li, Z.-H., Niyogi, S., and Docampo, R. (2015). CRISPR/Cas9-Induced Disruption of Paraflagellar Rod Protein 1 and 2 Genes in *Trypanosoma cruzi* Reveals Their Role in Flagellar Attachment. *mBio* 6, e01012.

De Lange, T., Michels, P.A.M., Veerman, H.J.G., Cornelissen, A.W.C.A., and Borst, P. (1984). Many trypanosome messenger RNAs share a common 5' terminal sequence. *Nucleic Acids Res.* *12*, 3777–3790.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* *9*, 357–359.

van Leeuwen, F., Kieft, R., Cross, M., and Borst, P. (2000). Tandemly repeated DNA is a target for the partial replacement of thymine by  $\beta$ -d-glucosyl-hydroxymethyluracil in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *109*, 133–145.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annu. Rev. Genet.* *46*, 43–68.

Leung, K.F., Riley, F.S., Carrington, M., and Field, M.C. (2011). Ubiquitylation and Developmental Regulation of Invariant Surface Protein Expression in Trypanosomes. *Eukaryot. Cell* *10*, 916–931.

Li, G.-W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* *484*, 538–541.

Li, Z., Zou, C.-B., Yao, Y., Hoyt, M.A., McDonough, S., Mackey, Z.B., Coffino, P., and Wang, C.C. (2002). An Easily Dissociated 26 S Proteasome Catalyzes an Essential Ubiquitin-mediated Protein Degradation Pathway in *Trypanosoma brucei*. *J. Biol. Chem.* *277*, 15486–15498.

Liang, X.H., Haritan, A., Uliel, S., and Michaeli, S. (2003). trans and cis splicing in trypanosomatids: Mechanism, factors, and regulation. *Eukaryot. Cell* *2*, 830–840.

Ligtenberg, M.J., Bitter, W., Kieft, R., Steverding, D., Janssen, H., Calafat, J., and Borst, P. (1994). Reconstitution of a surface transferrin binding complex in insect form *Trypanosoma brucei*. *EMBO J.* *13*, 2565–2573.

Liu, a Y., Van der Ploeg, L.H., Rijsewijk, F. a, and Borst, P. (1983). The transposition unit of variant surface glycoprotein gene 118 of *Trypanosoma brucei*. Presence of repeated elements at its border and absence of promoter-associated sequences. *J. Mol. Biol.* *167*, 57–75.

Liu, X., Jiang, H., Gu, Z., and Roberts, J.W. (2013). High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc. Natl. Acad. Sci. U. S. A.* *110*, 11928–11933.

Lowell, J.E. (2005). Histone H2AZ dimerizes with a novel variant H2B and is enriched at repetitive DNA in *Trypanosoma brucei*. *J. Cell Sci.* *118*, 5721–5730.

Lowell, J.E., and Cross, G.A.M. (2004). A variant histone H3 is enriched at telomeres in *Trypanosoma brucei*. *J. Cell Sci.* *117*, 5937–5947.

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* *380*, 2095–2128.

Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E.M. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* *25*, 117–124.

Ma, H., Wu, Y., Dang, Y., Choi, J.-G., Zhang, J., and Wu, H. (2014). Pol III Promoters

to Express Small RNAs: Delineation of Transcription Initiation. *Mol Ther Nucleic Acids* 3, e161.

MacPherson, C.R., and Scherf, A. (2015). Flexible guide-RNA design for CRISPR applications using Protospacer Workbench. *Nat Biotech* 33, 805–806.

Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H.O., Bishop, J.R., Falcone, F.H., Gavrilescu, C., Montgomery, J.L., Santori, M.I., et al. (2000). A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA* 6, 163–169.

Mair, G.R., Braks, J.A.M., Garver, L.S., Wiegant, J.C.A.G., Hall, N., Dirks, R.W., Khan, S.M., Dimopoulos, G., Janse, C.J., and Waters, A.P. (2006). Regulation of Sexual Development of Plasmodium by Translational Repression. *Science*. 313, 667–669.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. *Nat Rev Micro* 9, 467–477.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science*. 339, 823–826.

Manful, T., Cristodero, M., and Clayton, C. (2009). DRBD1 is the Trypanosoma brucei homologue of the Spliceosome-Associated Protein 49. *Mol. Biochem. Parasitol.* 166, 186–189.

Manful, T., Fadda, a., and Clayton, C. (2011). The role of the 5'-3' exoribonuclease XRNA in transcriptome-wide mRNA degradation. *RNA* 17, 2039–2047.

Mani, J., Güttinger, A., Schimanski, B., Heller, M., Acosta-Serrano, A., Pescher, P., Späth, G., and Roditi, I. (2011). Alba-domain proteins of *Trypanosoma brucei* are cytoplasmic RNA-binding proteins that interact with the translation machinery. *PLoS One* 6, e22463.

Marcello, L., and Barry, J.D. (2007). Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.* 17, 1344–1352.

Matsui, M., Yachie, N., Okada, Y., Saito, R., and Tomita, M. (2007). Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett.* 581, 4184–4188.

Matthews, K.R., and Gull, K. (1994). Evidence for an interplay between cell cycle progression and the initiation of differentiation between life cycle forms of African trypanosomes. *J. Cell Biol.* 125, 1147–1156.

Matthews, K.R., Tschadi, C., Ullu, E., Tschudi, C., and Ullu, E. (1994). A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes Dev.* 8, 491–501.

Matthews, K.R., Ellis, J.R., and Paterou, A. (2004). Molecular regulation of the life cycle of African trypanosomes. *Trends Parasitol.* 20, 40–47.

Mayho, M., Fenn, K., Craddy, P., Crosthwaite, S., and Matthews, K. (2006). Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in *Trypanosoma brucei*: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements. *Nucleic Acids Res.* 34, 5312–5324.



McCulloch, R., and Barry, J.D. (1999). A role for RAD51 and homologous recombination in *Trypanosoma brucei* antigenic variation. *Genes Dev.* *13*, 2875–2888.

McGrath, P.T. (2011). Characterizing cDNA ends by circular RACE. *Methods Mol Biol.* *772*:257-65

McNicoll, F., Drummelsmith, J., Müller, M., Madore, É., Boilard, N., Ouellette, M., and Papadopoulou, B. (2006). A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics* *6*, 3567–3581.

McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* *24*, 529–538.

Melville, S.E., Leech, V., Navarro, M., and Cross, G.A. (2000). The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427. *Mol. Biochem. Parasitol.* *111*, 261–273.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. *Nat Rev Genet* *10*, 155–159.

Miao, J., Li, J., Fan, Q., Li, X., Li, X., and Cui, L. (2010). The Puf-family RNA-binding protein PfPuf2 regulates sexual development and sex differentiation in the malaria parasite *Plasmodium falciparum*. *J. Cell Sci.* *123*, 1039–1049.

Michel, A.M., and Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip. Rev. RNA* *4*, 473–490.

Mignone, F., Gissi, C., Liuni, S., and Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biol.* 3, REVIEWS0004.

Millerioux, Y., Ebikeme, C., Biran, M., Morand, P., Bouyssou, G., Vincent, I.M., Mazet, M., Riviere, L., Franconi, J.-M., Burchmore, R.J.S., et al. (2013). The threonine degradation pathway of the *Trypanosoma brucei* procyclic form: the main carbon source for lipid biosynthesis is under metabolic control. *Mol. Microbiol.* 90, 114-29.

Mocanu, M., Mocanu, V., Parker, C., Warren, M., and Dicheva, N. (2009). Mass Spectrometry for Post-Translational Modifications. In *Neuroproteomics*, (CRC Press), pp. 93–113.

Mojica, F.J.M., Diez-Villasenor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* 36, 244–246.

Mony, B.M., MacGregor, P., Ivens, A., Rojas, F., Cowton, A., Young, J., Horn, D., and Matthews, K. (2014). Genome-wide dissection of the quorum sensing signalling pathway in *Trypanosoma brucei*. *Nature* 505, 681–685.

Moore, M.J. (2005). From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science*. 309, 1514–1518.

Morris, K. V, and Mattick, J.S. (2014). The rise of regulatory RNA. *Nat Rev Genet* 15, 423–437.

Mottram, J., Perry, K.L., Lizardi, P.M., Lührmann, R., Agabian, N., and Nelson, R.G. (1989). Isolation and sequence of four small nuclear U RNA genes of

*Trypanosoma brucei* subsp. *brucei*: identification of the U2, U4, and U6 RNA analogs. *Mol. Cell. Biol.* 9, 1212–1223.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2014). Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* 7, 548–548.

Nakaar, V., Günzl, A., Ullu, E., and Tschudi, C. (1997). Structure of the *Trypanosoma brucei* U6 snRNA gene promoter. *Mol. Biochem. Parasitol.* 88, 13–23.

Narayanan, M.S., and Rudenko, G. (2013). TDP1 is an HMG chromatin protein facilitating RNA polymerase I transcription in African trypanosomes. *Nucleic Acids Res.* 41, 2981–2992.

Navarro, M., and Gull, K. (2001). A pol I transcriptional body associated with VSG monoallelic expression in *Trypanosoma brucei*. *Nature* 414, 759–763.

Ngô, H., Tschudi, C., Gull, K., and Ullu, E. (1998). Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14687–14692.

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., and Ochsenreiter, T. (2010). Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of *Trypanosoma brucei*. *PLoS Pathog.* 6, e1001037.

Oberholzer, M., Morand, S., Kunz, S., and Seebeck, T. (2006). A vector series for rapid PCR-mediated C-terminal in situ tagging of *Trypanosoma brucei* genes.

Mol. Biochem. Parasitol. *145*, 117–120.

Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G., et al. (2011). Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo. *Cell* *147*, 1295–1308.

Opperdoes, F.R. (1987). Compartmentation of Carbohydrate Metabolism in Trypanosomes. *Annu. Rev. Microbiol.* *41*, 127–151.

Paindavoine, P., Rolin, S., van Assel, S., Geuskens, M., Jauniaux, J.-C., Dinsart, C., Huet, G., and Pays, E. (1992). A Gene from the Variant Surface Glycoprotein Expression Site Encodes One of Several Transmembrane Adenylate Cyclases Located on the Flagellum of *Trypanosoma brucei*. *Mol. Cell. Biol.* *12*, 1218–1225.

Palenchar, J.B., and Bellofatto, V. (2006). Gene transcription in trypanosomes. *Mol. Biochem. Parasitol.* *146*, 135–141.

Pang, K.C., Frith, M.C., and Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* *22*, 1–5.

Parsons, M., Nelson, R.G., Watkins, K.P., and Agabian, N. (1984). Trypanosome mRNAs share a common 5' spliced leader sequence. *Cell* *38*, 309–316.

Parsons, M., Ramasamy, G., Vasconcelos, E.J.R., Jensen, B.C., and Myler, P.J. (2015). Advancing *Trypanosoma brucei* genome annotation through ribosome profiling and spliced leader mapping. *Mol. Biochem. Parasitol.* *202*, 1–10.

Passos-Silva, D.G., Rajão, M.A., Nascimento de Aguiar, P.H., Vieira-da-Rocha, J.P., Machado, C.R., and Furtado, C. (2010). Overview of DNA Repair in *Trypanosoma cruzi*, *Trypanosoma brucei*, and *Leishmania major*. *J. Nucleic Acids* *2010*, 840768.

Patnaik, P.K., Kulkarni, S.K., and Cross, G.A. (1993). Autonomously replicating single-copy episomes in *Trypanosoma brucei* show unusual stability. *EMBO J.* *12*, 2529–2538.

Patnaik, P.K., Fang, X., and Cross, G.A. (1994). The region encompassing the procyclic acidic repetitive protein (PARP) gene promoter plays a role in plasmid DNA replication in *Trypanosoma brucei*. *Nucleic Acids Res.* *22*, 4111–4118.

Pays, E., Lheureux, M., and Steinert, M. (1982). Structure and expression of a *Trypanosoma brucei* gambiense variant specific antigen gene. *Nucleic Acids Res.* *10*, 3149–3163.

Pays, E., Lips, S., Nolan, D., Vanhamme, L., and Pérez-Morga, D. (2001). The VSG expression sites of *Trypanosoma brucei*: Multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol. Biochem. Parasitol.* *114*, 1–16.

Peacock, L., Ferris, V., Sharma, R., Sunter, J., Bailey, M., Carrington, M., and Gibson, W. (2011). Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. *Proc. Natl. Acad. Sci.* *108*, 3671–3676.

Peng, D., Kurup, S.P., Yao, P.Y., Minning, T.A., and Tarleton, R.L. (2015). CRISPR-Cas9-Mediated Single-Gene and Gene Family Disruption in *Trypanosoma cruzi*. *mBio* *6*, e02097-14.

Perry, K.L., Watkins, K.P., and Agabian, N. (1987). Trypanosome mRNAs have unusual “cap 4” structures acquired by addition of a spliced leader. *Proc. Natl. Acad. Sci.* *84*, 8190–8194.

Pircher, A., Gebetsberger, J., and Polacek, N. (2014). Ribosome-associated

ncRNAs: An emerging class of translation regulators. *RNA Biol.* *11*, 1335–1339.

Povelones, M.L., Gluenz, E., Dembek, M., Gull, K., and Rudenko, G. (2012). Histone H1 Plays a Role in Heterochromatin Formation and VSG Expression Site Silencing in *Trypanosoma brucei*. *PLoS Pathog.* *8*, e1003010.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2016). Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* *152*, 1173–1183.

Ran, F.A., Hsu, P.P.D., Wright, J., Agarwala, V., Scott, D. A, and Zhang, F. (2013a). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* *8*, 2281–2308.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., Konermann, S., Trevino, A., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., et al. (2013b). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* *154*, 1380–1389.

Reed, R. (2000). Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* *12*, 340–345.

Reeves, R. (2010). Nuclear functions of the HMG proteins. *Biochim. Biophys. Acta - Gene Regul. Mech.* *1799*, 3–14.

Robinson, K.A., and Beverley, S.M. (2003). Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite *Leishmania*. *Mol. Biochem. Parasitol.* *128*, 217–228.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* *26*, 139–140.

Roditi, I., and Liniger, M. (2002). Dressed for success: the surface coats of insect-borne protozoan parasites. *Trends Microbiol.* *10*, 128–134.

Rooijers, K., Loayza-Puch, F., Nijtmans, L.G., and Agami, R. (2013). Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat. Commun.* *4*, 2886.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H., et al. (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* *40*, D912–D917.

Rudenko, G. (2010). Epigenetics and transcriptional control in African trypanosomes. *Essays Biochem.* *48*, 201–219.

Rudenko, G. (2011). African trypanosomes: the genome and adaptations for immune evasion. *Essays Biochem.* *51*, 47–62.

Rudenko, G., Le Blancq, S., Smith, J., Lee, M.G., Rattray, A, and Van der Ploeg, L.H. (1990). Procyclic acidic repetitive protein (PARP) genes located in an unusually small alpha-amanitin-resistant transcription unit: PARP promoter activity assayed by transient DNA transfection of *Trypanosoma brucei*. *Mol. Cell. Biol.* *10*, 3492–3504.

Ryan, O.W., Skerker, J.M., Maurer, M.J., Li, X., Tsai, J.C., Poddar, S., Lee, M.E., DeLoache, W., Dueber, J.E., Arkin, A.P., et al. (2014). Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife* *3*, e03703.

Saas, J., Ziegelbauer, K., von Haeseler, a., Fast, B., and Boshart, M. (2000). A

Developmentally Regulated Aconitase Related to Iron-regulatory Protein-1 Is Localized in the Cytoplasm and in the Mitochondrion of *Trypanosoma brucei*. *J. Biol. Chem.* *275*, 2745–2755.

Saleh-Gohari, N., and Helleday, T. (2004). Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. *Nucleic Acids Res.* *32*, 3683–3688.

Sambrook, J. and Russell, D. (2001). *Molecular Cloning a Laboratory Manual*. Cold Spring Harbor Laboratory Press. Third edition.

Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.-L., and Weiss, D.S. (2013). A CRISPR-CAS System Mediates Bacterial Innate Immune Evasion and Virulence. *Nature* *497*, 254–257.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for genome editing, regulation and targeting. *Nat. Biotechnol.* *32*, 347–355.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* *39*, 9275–9282.

Scahill, M.D., Pastar, I., and Cross, G. A. M. (2008). CRE recombinase-based positive–negative selection systems for genetic manipulation in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *157*, 73–82.

Scheibe, M., Butter, F., Hafner, M., Tuschl, T., and Mann, M. (2012). Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Res.* *40*, 9897–9902.



Scholler, J.K., Aline Jr., R.F., and Stuart, K.D. (1988). Variant specific transcripts from the co-transposed segments of variant surface glycoprotein genes in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 29, 89–103.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* 473, 337–342.

Schwede, A., Manful, T., Jha, B.A., Helbig, C., Bercovich, N., Stewart, M., and Clayton, C. (2009). The role of deadenylation in the degradation of unstable mRNAs in trypanosomes. *Nucleic Acids Res.* 37, 5511–5528.

Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C.B. (2013). Widespread regulation of translation by elongation pausing in heat shock. *Mol. Cell* 49, 439–452.

Shapiro, S.Z., Naessens, J., Liesegang, B., Mooloo, S.K., and Magundu, J. (1984). Analysis by flow cytometry of DNA synthesis during the life cycle of African trypanosomes. *Acta Trop* 41, 313–323.

Sharma, R., Peacock, L., Gluenz, E., Gull, K., Gibson, W., and Carrington, M. (2008). Asymmetric Cell Division as a Route to Reduction in Cell Length and Change in Cell Morphology in Trypanosomes. *Protist* 159, 137–151.

Shen, B., Brown, K.M., Lee, T.D., and Sibley, L.D. (2014). Efficient Gene Disruption in Diverse Strains of *Toxoplasma gondii* Using CRISPR/CAS9. *mBio* 5 e01114-14.

Sidik, S.M., Hackett, C.G., Tran, F., Westwood, N.J., and Lourido, S. (2014). Efficient Genome Engineering of *Toxoplasma gondii* Using CRISPR/Cas9. *PLoS One* 9,

e100450.

Siegel, T.N., Tan, K.S.W., and Cross, G. A. M. (2005). Systematic Study of Sequence Motifs for RNA trans Splicing in *Trypanosoma brucei* Systematic Study of Sequence Motifs for RNA trans Splicing in *Trypanosoma brucei*. *Mol. Cell. Biol.* 25, 9586–9594.

Siegel, T.N., Kawahara, T., DeGrasse, J. A., Janzen, C.J., Horn, D., and Cross, G. A. M. (2007). Acetylation of histone H4K4 is cell cycle regulated and mediated by HAT3 in *Trypanosoma brucei*. *Mol. Microbiol.* 67, 762–771.

Siegel, T.N., Hekstra, D.R., Kemp, L.E., Figueiredo, L.M., Lowell, J.E., Fenyó, D., Wang, X., Dewell, S., and Cross, G. A. M. (2009). Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.* 23, 1063–1076.

Siegel, T.N., Hekstra, D.R., Wang, X., Dewell, S., and Cross, G. A. M. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* 38, 4946–4957.

Siegel, T.N., Gunasekera, K., Cross, G. A. M., and Ochsenreiter, T. (2011). Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing. *Trends Parasitol.* 27, 434–441.

Simon, M.D. (2013). Capture Hybridization Analysis of RNA Targets (CHART). In *Current Protocols in Molecular Biology*, (John Wiley & Sons, Inc.). Chapter 21, Unit 21.25.

- Sledz, C.A., and Williams, B.R.G. (2005). RNA interference in biology and disease. *Blood* 106, 787–794.
- Smircich, P., Eastman, G., Bispo, S., Duhagon, M.A., Guerra-Slompo, E.P., Garat, B., Goldenberg, S., Munroe, D.J., Dallagiovanna, B., Holetz, F., et al. (2015). Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in *Trypanosoma cruzi*. *BMC Genomics* 16, 443.
- Smith, M.A., Gesell, T., Stadler, P.F., and Mattick, J.S. (2013). Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.* 4, 8220-36.
- Sollelis, L., Ghorbal, M., MacPherson, C.R., Martins, R.M., Kuk, N., Crobu, L., Bastien, P., Scherf, A., Lopez-Rubio, J.-J., and Sterkers, Y. (2015). First efficient CRISPR-Cas9-mediated genome editing in *Leishmania* parasites. *Cell. Microbiol.* 17, 1405–1412.
- Somers, J., Pöyry, T., and Willis, A.E. (2013). A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* 45, 1690–1700.
- Spicuglia, S., Maqbool, M.A., Puthier, D., and Andrau, J.-C. (2013). An update on recent methods applied for deciphering the diversity of the noncoding RNA genome structure and function. *Methods* 63, 3–17.
- Stadler, M., and Fire, A. (2011). Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* 17, 2063–2073.
- Stanne, T.M., and Rudenko, G. (2010). Active VSG Expression Sites in *Trypanosoma brucei* Are Depleted of Nucleosomes. *Eukaryot. Cell* 9, 136–147.
- Stanne, T.M., Kushwaha, M., Wand, M., Taylor, J.E., and Rudenko, G. (2011).

TbISWI Regulates Multiple Polymerase I (Pol I)-Transcribed Loci and Is Present at Pol II Transcription Boundaries in *Trypanosoma brucei*. *Eukaryot. Cell* *10*, 964–976.

Steitz, J.A. (1969). Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature* *224*, 957–964.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T.K., Hein, M.Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., Hengel, H., et al. (2012). Decoding Human Cytomegalovirus. *Science*. *338*, 1088–1093.

Steverding, D. (2003). The significance of transferrin receptor variation in *Trypanosoma brucei*. *Trends Parasitol.* *19*, 125–127.

Steverding, D. (2010). The development of drugs for treatment of sleeping sickness: a historical review. *Parasit. Vectors* *3*, 15.

Štros, M. (2010). HMGB proteins: Interactions with DNA and chromatin. *Biochim. Biophys. Acta - Gene Regul. Mech.* *1799*, 101–113.

Tan, T.H.P., Pach, R., Crausaz, A., Ivens, A., and Schneider, A. (2002). tRNAs in *Trypanosoma brucei*: Genomic Organization, Expression, and Mitochondrial Import. *Mol. Cell. Biol.* *22*, 3707–3717.

Tebas, P., Stein, D., Tang, W.W., Frank, I., Wang, S.Q., Lee, G., Spratt, S.K., Surosky, R.T., Giedlin, M.A., Nichol, G., et al. (2014). Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV. *N. Engl. J. Med.* *370*, 901–910.

Tiengwe, C., Marcello, L., Farr, H., Dickens, N., Kelly, S., Swiderski, M., Vaughan, D., Gull, K., Barry, J.D., Bell, S.D., et al. (2012). Genome-wide Analysis Reveals

Extensive Functional Interaction between DNA Replication Initiation and Transcription in the Genome of *Trypanosoma brucei*. *Cell Rep.* *2*, 185–197.

Tschudi, C., Richards, F.F., and Ullu, E. (1986). The U2 RNA analogue of *Trypanosoma brucei gambiense*: implications for a splicing mechanism in trypanosomes. *Nucleic Acids Res.* *14*, 8893–8903.

Ullu, E., Matthews, K.R., and Tschudi, C. (1993). Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts. *Mol. Cell. Biol.* *13*, 720–725.

Urbaniak, M.D., Guther, M.L.S., and Ferguson, M. A. J. (2012). Comparative SILAC Proteomic Analysis of *Trypanosoma brucei* Bloodstream and Procyclic Lifecycle Stages. *PLoS One* *7*, e36619.

Valdés, J., Taylor, M.C., Cross, M. A, Ligtenberg, M.J., Rudenko, G., and Borst, P. (1996). The viral thymidine kinase gene as a tool for the study of mutagenesis in *Trypanosoma brucei*. *Nucleic Acids Res.* *24*, 1809–1815.

VanDenAbbeele, J., Claes, Y., VanBockstaele, D., LeRay, D., and Coosemans, M. (1999). *Trypanosoma brucei* spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. *Parasitol. Today* *118*, 416–478.

Vanhamme, L., and Pays, E. (1995). Control of gene expression in trypanosomes. *Microbiol. Rev.* *59*, 223–240.

Vasquez, J.J., Hon, C.C., Vanselow, J.T., Schlosser, A., and Siegel, T.N. (2014).

Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res.* *42*, 3623–3637.

Vassella, E., Reuner, B., Yutzy, B., and Boshart, M. (1997). Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway. *J. Cell Sci.* *110*, 2661–2671.

Vassella, E., Acosta-Serrano, A., Studer, E., Lee, S.H., Englund, P.T., and Roditi, I. (2001). Multiple procyclin isoforms are expressed differentially during the development of insect forms of *Trypanosoma brucei*. *J. Mol. Biol.* *312*, 597–607.

Vattem, K.M., and Wek, R.C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* *101*, 11269–11274.

Vickerman, K. (1969). On the surface coat and flagellar adhesion in trypanosomes. *J. Cell Sci.* *5*, 163–193.

Wagner, J.C., Platt, R.J., Goldfless, S.J., Zhang, F., and Niles, J.C. (2014). Efficient CRISPR-Cas9-mediated genome editing in *Plasmodium falciparum*. *Nat. Methods* *11*, 915–918.

Walrad, P., Paterou, A., Acosta-Serrano, A., and Matthews, K.R. (2009). Differential Trypanosome Surface Coat Regulation by a CCCH Protein That Co-Associates with procyclin mRNA cis-Elements. *PLoS Pathog.* *5*, e1000317.

Walrad, P.B., Capewell, P., Fenn, K., and Matthews, K.R. (2012). The post-transcriptional trans-acting regulator, TbZFP3, co-ordinates transmission-stage enriched mRNAs in *Trypanosoma brucei*. *Nucleic Acids Res.* *40*, 2869–2883.

Wang, Q.-P., Kawahara, T., and Horn, D. (2010). Histone deacetylases play distinct roles in telomeric VSG expression site silencing in African trypanosomes. *Mol. Microbiol.* 77, 1237–1245.

Watanabe, T., Ochiai, H., Sakuma, T., Horch, H.W., Hamaguchi, N., Nakamura, T., Bando, T., Ohuchi, H., Yamamoto, T., Noji, S., et al. (2012). Non-transgenic genome modifications in a hemimetabolous insect using zinc-finger and TAL effector nucleases. *Nat. Commun.* 3, 1017.

Webb, H. (2005). Developmentally regulated instability of the GPI-PLC mRNA is dependent on a short-lived protein factor. *Nucleic Acids Res.* 33, 1503–1512.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A., and Leutz, A. (2014). uORFdb—a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res.* 42, D60–D67.

Wharton, R.P., and Aggarwal, A. K. (2006). mRNA Regulation by Puf Domain Proteins. *Sci. STKE* 2006, pe37–pe37.

WHO (2015). <http://www.who.int/mediacentre/factsheets/fs259/en/>  
Retrieved April 2015. Trypanosomiasis, Human African (sleeping sickness).

Wickstead, B., Ersfeld, K., and Gull, K. (2003). The frequency of gene targeting in *Trypanosoma brucei* is independent of target site copy number. *Nucleic Acids Res.* 31, 3993–4000.

Wilusz, C.J., and Wilusz, J. (2004). Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet.* 20, 491–497.

Wirtz, E., Leal, S., Ochatt, C., and Cross, G.M. (1999). A tightly regulated inducible

expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *99*, 89–101.

Wolin, S.L., and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.* *7*, 3559–3569.

Wright, J.R., Siegel, T.N., and Cross, G. A. M. (2010). Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* *172*, 141–144.

Yang, X., Figueiredo, L.M., Espinal, A., Okubo, E., and Li, B. (2009). RAP1 Is Essential for Silencing Telomeric Variant Surface Glycoprotein Genes in *Trypanosoma brucei*. *Cell* *137*, 99–109.

Yoffe, Y., Zuberek, J., Lewdorowicz, M., Zeira, Z., Keasar, C., Orr-Dahan, I., Jankowska-Anyszka, M., Stepinski, J., Darzynkiewicz, E., and Shapira, M. (2004). Cap-binding activity of an eIF4E homolog from *Leishmania*. *RNA* *10*, 1764–1775.

Yoffe, Y., Zuberek, J., Lerer, A., Lewdorowicz, M., Stepinski, J., Altmann, M., Darzynkiewicz, E., and Shapira, M. (2006). Binding Specificities and Potential Roles of Isoforms of Eukaryotic Initiation Factor 4E in *Leishmania*. *Eukaryot. Cell* *5*, 1969–1979.

Yun, O., Priotto, G., Tong, J., Flevaud, L., and Chappuis, F. (2010). NECT Is Next: Implementing the New Drug Combination Therapy for *Trypanosoma brucei* gambiense Sleeping Sickness. *PLoS Negl Trop Dis* *4*, e720.

Zhang, W., and Matlashewski, G. (2015). CRISPR-Cas9-Mediated Genome Editing in *Leishmania donovani*. *MBio* *6*, e00861.



Ziegelbauer, K., Quinten, M., Schwarz, H., Pearson, T.W., and Overath, P. (1990). Synchronous differentiation of *Trypanosoma brucei* from bloodstream to procyclic forms in vitro. *Eur. J. Biochem.* *192*, 373–378.

Zinoviev, A., and Shapira, M. (2012). Evolutionary Conservation and Diversification of the Translation Initiation Apparatus in Trypanosomatids. *Comp. Funct. Genomics* *2012*, 1–10.

Zomerdijk, J.C., Kieft, R., Shiels, P.G., and Borst, P. (1991). Alpha-amanitin-resistant transcription units in trypanosomes: a comparison of promoter sequences for a VSG gene expression site and for the ribosomal RNA genes. *Nucleic Acids Res.* *19*, 5153–5158.

Zoschke, R., Watkins, K.P., and Barkan, A. (2013). A Rapid Ribosome Profiling Method Elucidates Chloroplast Ribosome Behavior in Vivo. *Plant Cell* *25*, 2265–2275.

## **Curriculum vitae**

**Name:** Juan José Vásquez Ospina

**Place of birth:** Bogotá, Colombia

**Date of birth:** 05 April 1979

### **Personal information**

---

Work Address: University of Würzburg, Institute for Molecular Infection Biology (IMIB), Josef Schneider Str. 2, Bau D15 97080, Würzburg, Germany.

Telephone number: +49-931-31-81746

E-mail: [juan.vasquez\\_ospina@uni-wuerzburg.de](mailto:juan.vasquez_ospina@uni-wuerzburg.de)

### **Education**

---

Nov 2012 – April 2016: University of Würzburg, Germany, Institute for Molecular Infection Biology (IMIB). PhD. candidate (pending thesis defense by April 2016).

Aug 2009 – Dec 2011: Universidad de los Andes, Colombia. M.Sc. Biological Sciences, area Microbiology.

Jan 1996 – June 2002: Universidad de los Andes, Colombia. B.Sc. Biology

### **Experience**

---

Nov 2012 – present: University of Würzburg, Germany, Institute of Molecular Infection Biology (IMIB). Trypanosoma gene regulation group.

PhD. thesis: "Development of tools for the study of gene regulation in *Trypanosoma brucei*".

March 2012 – Sept 2012: Universidad de los Andes, Colombia. Laboratory of Mycology and Plant Pathology and Research Group in Biochemistry and Molecular Biology of Parasites.

Research assistant. Cloned, expressed and purified the *de novo* pyrimidine biosynthesis pathway enzymes in *Solanum tuberosum*.

Jan 2011 – Feb 2012: Universidad de los Andes, Colombia.

Graduate assistant. Coordinated the search of grants and funding opportunities for the Faculty of Science. Participated in the design of the Funding Support Office web page for the Faculty of Science.

Aug 2010 – Dec 2011: Universidad de los Andes, Colombia. Laboratory of Mycology and Plant Pathology and Research Group in Biochemistry and Molecular Biology of Parasites.

M.Sc. thesis: "Characterization of a pyrimidine salvage enzyme of the potato late blight *Phytophthora infestans* and its host *Solanum tuberosum*".

Jan 2010 – June 2010: Universidad de los Andes, Colombia.

Graduate teaching assistant. Taught complementary lectures, lead discussion groups, prepared and graded exams and class activities in the course Deciphering the Human Genome.

April 2005 – Jan 2009: National Center for Coffee Research (Cenicafé), Colombia. Plant Breeding and Biotechnology Unit.

Research Assistant. Characterized enzymes involved in the coffee bean digestion by the coffee berry borer *Hypothenemus hampei*.

April 2003 – June 2004: International Center for Medical Research and Training (CIDEIM), Colombia. Biochemistry and Molecular Biology Unit.

Young researcher. Identified molecular markers associated with resistance to antimonial drugs and differentially regulated proteins in two parasite stages of *Leishmania panamensis*

Aug 2001 – June 2002: International Center for Tropical Agriculture (CIAT), Colombia.

B.Sc. thesis: “Morphological, phenological and genetic characterization of red rice biotypes collected in the municipality of Saldaña (Tolima”).

## **Publications**

---

**Vasquez JJ**, Hon CC, Vanselow JT, Schlosser A, Siegel TN. Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. Nucleic Acids Res. 2014 Jan 17.

García-Bayona L, Garavito MF, Lozano GL, **Vasquez JJ**, Myers K, Fry WE, Bernal A, Zimmermann BH, Restrepo S. De novo pyrimidine biosynthesis in the oomycete plant pathogen *Phytophthora infestans*. Gene. 2014. Mar 10, 537(2):312-21.

Aguilera-Gálvez C, **Vásquez-Ospina JJ**, Gutiérrez-Sánchez P, Acuña-Zornosa R. Cloning and biochemical characterization of an endo-1,4- $\beta$ -mannanase from the coffee berry borer *Hypothenemus hampei*. BMC Res Notes. 2013 Aug 22;6:333.

Walker J, Gongora R, **Vasquez JJ**, Drummelsmith J, Burchmore R, Roy G, Ouellette

M, Gomez MA, Saravia NG. Discovery of factors linked to antimony resistance in *Leishmania panamensis* through differential proteome analysis. *Mol Biochem Parasitol.* 2012 Jun;183(2):166-76.

Walker, J., **Vásquez, J.J.**, Gómez, M.A., Drummelsmith, J., Burchmore, R., Girard, I. and Ouellette, M. Identification of developmentally-regulated proteins in *Leishmania panamensis* by proteome profiling of promastigotes and axenic amastigotes. *Mol. Biochem. Parasitol.* 2006. 147(1): 64-73.

### **Posters and presentations**

---

**Vásquez, J.J.**, Wedel C, Siegel T.N. Does ncRNAs modulate antigenic variation in *Trypanosoma brucei*? Kinetoplastid Molecular Cell Biology Meeting (KMCBM), Woods Hole, Massachusetts, USA, April 25-29, 2015. Poster.

BSP Spring Meeting, 7<sup>th</sup> April, 2014, Cambridge University, UK. Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. **Oral presentation.**

**Vásquez, J.J.**, Garavito, M.F., García Bayona, L., Tabima J.F, Bernal, A., Zimmermann, B.H., Restrepo, S. Uracil Phosphoribosyltransferase inhibition as a complementary target for *Phytophthora infestans* control. Oomycete Molecular Genetics Network Congress, Asilomar, California, USA, March 13-15, 2011. Poster.

**Vásquez, J.J.**, Acuña, R., Ossa, G., Rubio, D., Padilla, B. Differential Gel Electrophoresis (DIGE) analysis: a new tool applied to the functional genomic research on insects. XXXIII Entomology Colombian Society Meeting, SOCOLEN. July 26-28, 2006. Manizales, Colombia. Oral presentation.

## **Awards and honors**

---

Semestral Fellowship, Universidad de los Andes, 2009-2011. This fellowship is awarded by the Faculty of Science to students that show outstanding level during their Master thesis.

Hernando Alcaraz Viecco Award. Granted for the best work presented by professional researchers in the XXXII Entomology Colombian Society Meeting. Ibagué, July 27-29 2005. Work title: Proteomic analysis of different developmental stages of the coffee berry borer *Hypothenemus hampei*.

Young Researcher Fellowship. From April 2003 to June 2004. This fellowship is granted to promising young recent graduate researchers. Granted by the Administrative Department of Science, Technology and Innovation (Colciencias).

Semestral Excellence Fellowship, Universidad de los Andes, 2000. This fellowship is awarded to the higher GPA between students of the Faculty of Science.

## **Languages**

---

Native language: Spanish (fluent)

Other languages: English (fluent), German (beginner)

Signature

Place and date