Revealing the architecture of post transcriptional gene regulatory circuits in Trypanosomatids

Vahid H. Gazestani

Institute of Parasitology

McGill University, Montreal

August 2016

A thesis submitted to McGill University

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Table of Contents

Abstract

The highly diverged trypanosomatid parasites cause devastating diseases in humans and animals and continue to pose a major challenge in the drug development. During transmission between mammalian host and fly vector, these parasites face major environmental changes including available energy sources, immune system, temperature, and pH that require the fine regulation of gene expression patterns for adaptation and survival. Intriguingly, trypanosomatid gene regulation occurs almost entirely at the post transcriptional level. RNA-binding proteins play an important role in this process. Due to high evolutionary distance from other eukaryotes, supported by experimental data, the binding affinities of trypanosomatid RNA-binding proteins are diverged from other eukaryotes, including their human hosts, representing an ideal pathway for development of anti-parasitic drug leads. However, the current knowledge on the gene regulatory pathways of these parasites is only rudimentary. The focus of this thesis is on the inference of regulatory mechanisms that are essential for the differentiation and adaptation of the parasite in different life stages.

First, I review previously employed techniques for identification and characterization of RNA-binding proteins and their cognate mRNA binding sites (*cis*-regulatory elements). I argue that the limitations of small-scale studies on post transcriptional regulatory circuits have some fundamental limitations that hinder their use to chart the regulatory network of the parasite. Then, I propose several genome-wide techniques that can provide a global picture by circumventing the limitations of small-scale data. As the first step, by using a novel graph-based approach, I predicted 88 high confidence *cis*-regulatory elements (estimated accuracy of more than 60% based on benchmarking on human data) that are potentially involved in the transmission of the *Trypanosoma brucei* from the vector to the host. Importantly, 64% (56 out of 88) of the predicted RNA motifs have more than 50% A/U nucleotides in their sequence composition. However, these motifs mostly target different transcripts (as judged by motif co-occurrence profiles) and show different responses during the life cycle of *T. brucei*, suggesting a potentially distinct and diverse role for A-/U-rich motifs in the gene regulatory network of the parasite. I also developed a novel next generation sequencing based approach to investigate the effect of the 3′-tails variation on the mitochondrial transcripts. To the best of our knowledge, this approach provides highest depth of tail analysis compared to other techniques developed thus far. Application of this approach on *T. brucei* mitochondrial mRNAs led to new insights on RNA

polycistron processing and identification of transcript-specific tail variations. Next, benefiting from in-depth fractionation techniques coupled with mass spectrometry, I introduce the first experimentally derived protein co-complex map of *T. brucei*. Obtained results led to the assignment of 716 proteins including 635 protein groups that lacked experimental annotation to protein complexes. The constructed map was highly informative on the composition of protein complexes involved in RNA transport, translation, and editing. As an illustration, we were able to show that *T. brucei* RNA editing machinery is composed of multiple protein complexes that are loosely connected with each other and also identified new subunits of the machinery. The quality of predictions was verified by independent follow up experiments on newly characterized proteins associated with RNA editing machinery. Lastly, I developed a protein interaction-centric database to automatically integrate the obtained knowledge from the previous aims with other available resources accompanied with related statistical analysis. Benefiting from powerful asp.Net framework, the developed database can reliably represent the results in a user friendly and intuitive web-interface. The database automatically performs inter-species mapping of available data and information among 16 trypanosomatid parasites to help better characterize the queried proteins in the species of interest. Based on the built in features, the database is able to help researchers on their interactome related experiments to distinguish between the likely binding partners of a protein from confounding elements identified in their experiments and also suggest other potentially interacting proteins that are missing from the list of queried proteins. Collectively, the obtained results not only provide a multifaceted picture of trypanosomatid gene regulatory circuits, but also can be used to identify new leads for their in-depth characterization. Moreover, the developed tools and strategies can be used for dissection of gene regulatory networks in most non-model organism, for many of which no other alternatives are available.

Résumé

Les parasites trypanosomatides hautement déviés provoquent des maladies dévastatrices chez les êtres humains et les animaux et continuent de poser un défi majeur dans le développement de médicaments. Pendant la transmission entre l'hôte mammifère et la mouche vecteur, ces parasites sont confrontés à des changements environnementaux majeurs, tels que les sources d'énergie disponibles, le système immunitaire, la température et le pH qui nécessite la régulation très affinée des modèles d'expression génétique pour l'adaptation et la survie. Curieusement, la régulation des gènes trypanosomatide se produit presque entièrement au niveau post transcriptionnel avec un accent important sur les protéines de liaison d'ARN. En raison de grandes distances évolutives d'autres eucaryotes, appuyées par des données expérimentales, les affinités de liaison des protéines de liaison d'ARN trypanosomatides sont différentes de celles d'autres eucaryotes, y compris leurs hôtes humains, ce qui représente une voie idéale pour le développement de pistes anti-trypanosomiase. Cependant, les connaissances actuelles sur les voies de régulation de gènes de ces parasites ne sont que rudimentaires. Nous développons ici de nouveaux outils pour la modélisation systématique de tels circuits réglementaires.
Nous examinons d'abord les techniques précédemment employées pour identifier et caractériser les éléments et leurs RBP parentes cis-régulatrices. Nous soutenons que les limites des études ciblées et à petite échelle sur les circuits réglementaires suite à leur transcription ont des limitations fondamentales qui entravent leur utilisation pour tracer le réseau réglementaire du parasite. Par conséquent, nous proposons plusieurs techniques de l'ensemble du génome qui peuvent fournir une image globale en contournant les limites de données à petite échelle. Comme première étape, en utilisant une nouvelle approche basée sur des graphique, nous avons prédit 88 éléments cis-régulateurs à haute confiance (précision estimée de plus de 60 % sur la base de l'analyse comparative sur les humains) qui sont potentiellement impliqués dans la transmission de T. brucei à partir du vecteur à l'hôte. Fait plus important, nous avons constaté que 64% (56 sur 88) des motifs prédits par notre approche ont plus de 50% de A et / ou U dans leur composition. Toutefois, ces motifs ciblent pour la plupart des transcriptions différentes (telles que jugées par des profils à motif de co-occurrence) et montrent des réponses différentes au cours du cycle de vie de T. brucei, suggérant un rôle potentiellement divers et distinct pour A et / ou de motifs riches en U dans le réseau de gène de régulation du parasite. Dans une autre direction, nous avons mis au point une nouvelle approche à base de séquençage de nouvelle génération afin

d'étudier l'effet des RBP sur les queues des transcriptions mitochondriales. L'approche élaborée, au mieux de notre connaissance, fournit une analyse de la plus haute profondeur de la queue par rapport à d'autres techniques développées jusqu'ici. Les applications de cette approche sur les transcriptions mitochondriaux T. brucei ont conduit à de nouvelles perspectives sur le traitement d'ARN de polycistron et l'identification des étapes de vie, ainsi qu'à des variations de la queue spécifiques à la transcription. Puis, profitant de techniques de fractionnement en profondeur associées à la spectrométrie de masse, nous avons tracé la carte complexe de T. Brucei de la première protéine dérivée expérimentalement. Nos résultats ont conduit à l'attribution de nombreuses protéines non définies auparavant à des complexes. La carte construite était très instructive sur la composition des complexes de protéines impliquées dans le transport de l'ARN, ainsi que sur la traduction et l'édition. À titre d'illustration, nous avons pu montrer que les machines d'édition T. brucei de l'ARN sont composées de complexes protéiques multiples qui sont faiblement connectés les uns avec les autres et également identifié de nouvelles sous-unités de la machine. La qualité des prédictions ont été vérifiées par des expériences de suivi indépendantes sur les protéines nouvellement caractérisés associés aux machines l'édition de l'ARN. Comme dernier objectif, nous avons développé une base de données d'interaction centrée sur la protéine afin d'intégrer automatiquement les connaissances obtenues à partir des objectifs précédents avec d'autres ressources disponibles accompagnés d'analyses statistiques connexes. Bénéficiant d'un cadre de asp.Net puissant, la base de données développée effectue rapidement et de manière fiable et représente les résultats dans une interface web intuitive et facile à utiliser. La base de données effectue automatiquement la cartographie inter-espèces des données et informations disponibles parmi 16 parasites trypanosomatide pour mieux aider à la caractérisation des protéines cherchées dans les espèces d'intérêt. Enfin, sur la base des fonctions intégrées, la base de données est en mesure d'aider les chercheurs dans leurs expériences liées aux interactome à distinguer entre les partenaires susceptibles de liaison d'une protéine à partir d'éléments confondants identifiés dans leurs expériences et proposer d'autres protéines d'interaction potentielles qui sont absents de la liste des protéines cherchées. Collectivement, les outils et les stratégies développées peuvent être utilisés pour la dissection des réseaux de régulation des gènes dans un organisme non-modèle et les résultats obtenus fournissent non seulement une image à multiples facettes des circuits de régulation des gènes de

trypanosomatide, mais peut aussi être utilisé pour identifier de nouvelles pistes pour leur caractérisation.

Contributions of Authors

Chapters 2, 3, 4, and 5 of this thesis contain materials from previously published manuscripts (see references (Gazestani et al. 2014; Gazestani and Salavati 2015; Gazestani et al. 2016a; Gazestani et al. 2016b)). Reference (Gazestani et al. 2014) was co-authored by me, Lucy Lu, and my supervisor, Reza Salavati. As the primary author, I contributed to the data analysis and manuscript preparation. Lucy Lu contributed to the manuscript preparation. Reference (Gazestani and Salavati 2015) was co-authored by me and Reza Salavati. As the primary author, I developed the method, analyzed the data and wrote the manuscript. Reference (Gazestani et al. 2016a) was coauthored by me, Marshall Hampton, Juan E Abrahante, Reza Salavati, and Sara L Zimmer. I contributed to method development, data analysis, and manuscript preparation. Marshall Hampton contributed to the data analysis. Juan E Abrahante contributed to the experimental part. Sara L Zimmer conceived the study and contributed to the data analysis and manuscript preparation. Reference (Gazestani et al. 2016b) was co-authored by me, Najmeh Nikpour, Vaibhav Mehta, Hamed S Najafabadi, Houtan Moshiri, Armando Jardim, and Reza Salavati. I contributed to experimental design, data analysis, and drafted the manuscript. Najmeh Nikpour, Vaibhav Mehta, Houtan Moshiri, and Armando Jardim contributed to the experimental design and experiments. Hamed S Najafabadi contributed to the experimental design. Reza Salavati provided intellectual input in all these studies, and contributed to experimental design and manuscript preparation.

Acknowledgements

This dissertation would not be possible without the help of so many people who challenged, supported, and stuck with me along this journey. First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Reza Salavati. I was incredibly fortunate to have an advisor that put his student's satisfaction, professionally and personally, at the first priority.

I would also thank our collaborator, Dr. Sara Zimmer that gave me the opportunity to add to my skill sets. Each one of my advisory committee member has helped me immensely in my projects. Dr. Jeff Xia, Dr. Robin Beech, and Dr. Jacek Majewski have guided me all throughout my PhD project. I thank them all for providing me with many useful inputs on many aspects of my research.

I would also like to thank current and former lab members of Salavati's lab especially Chun Wai Yip, Vaibhav Mehta, Lucy Lu, Najmeh Nikpour, Bhaskar Anand Jha, George Wu, and Keshika Prematilake for their contributions in all aspects of the presented chapters. I also thank all the faculty members, staff, and students at the Institute of Parasitology.

I also thank two of my closest friends at Montreal who their unwavering friendship helped me go through problems personally and professionally, Mehran Karimzadeh and Nafiseh Yavari. I would also like to thank Leila Pirhaji, my old friend who introduced me to the field of bioinformatics and have spiritually supported me in all matters.

Last but not the least, I would like to thank my family for giving me the confidence that I can accomplish anything and supported me even across the ocean. I thank my dad for teaching me to think critically and the right way to approach the problems that life throws to us. I also thank my mom who formed my mind in childhood with familiarizing me with philosophy and taught me how to be calm and collected during hard times.

To my parents Mohammad H Gazestani and Narges M Ebrahim,

for their immeasurable support, encouragement, and love.

Introduction

The highly diverged trypanosomatid parasites are responsible for various life-threatening diseases in humans and major production losses (e.g., meat, milk, and fertility) in animals (Brun et al. 2010; Rassi et al. 2010; Alvar et al. 2012). Various closely related species of trypanosomatid parasites have been identified throughout the world, most notably in Latin America (*T. cruzi* and *T. vivax*), sub-saharan Africa (*T. brucei*, *T. vivax*, and *T. congolense*), and parts of Asia (*T. evansi*) (World Health 2012). The high adaptation capabilities of these parasites have allowed them to continuously invade new hosts and new geographical regions. For example, *T. cruzi* that could only infect wild animals, has gained ability to infect humans and domestic animals during evolution (Hamilton et al. 2012). Consistently, *T. cruzi* infections are increasingly reported in North America, most of Europe, and parts of western pacific countries (World Health 2012). As another example, *T. evansi* which originally diverged from *T. brucei* and infect animals like horses and camels, no longer requires a fly vector for completion of its life cycle (Lai et al. 2008). More importantly, a recent medical report demonstrates that *T. evansi* may have gained ability to infect humans (Joshi et al. 2005). Although posing global threats, various issues are associated with available drugs (including tolerability, cost, and resistance), necessitating the identification of novel essential parasite-specific pathways/genes as potential drug targets (World Health 2012).

Availability of genome sequences for *T. brucei*, *T. cruzi*, and *Leishmania major* since 2005 has revolutionized the depth of every research in these parasites by not only accelerating the pace of focused studies, but also providing opportunities for large-scale, systematic analyses such as gene discovery and various comparisons including with their hosts and other model organisms (Berriman et al. 2005; El-Sayed et al. 2005a; El-Sayed et al. 2005b; Ivens et al. 2005). Information on genomic sequence also helped researchers to effectively interrogate molecular pathways by allowing the efficient use of high-throughput techniques such as microarray gene expression and protein mass spectrometry (Jensen et al. 2009; Kabani et al. 2009; Panigrahi et al. 2009; Queiroz et al. 2009; Gunasekera et al. 2012; Urbaniak et al. 2012; Gazestani et al. 2016b). The combination of these studies has led to detailed knowledge on the RNA and protein composition, active molecular pathways, protein translational rates and post translational modifications, and remodeling of surface proteins in different life stages and in response to varying environmental conditions (Kramer et al. 2008; Jensen et al. 2009; Kabani et al. 2009;

Panigrahi et al. 2009; Queiroz et al. 2009; Kramer et al. 2010b; Alsford et al. 2011; Gunasekera et al. 2012; Urbaniak et al. 2012; Urbaniak et al. 2013). However, systems level insights on the underlying mechanisms of differentiation and adaptation during the life cycle of trypanosomatid parasites are still missing, hindering the efforts for finding potential drug targets against trypanosomiasis.

In the absence of transcriptional regulation, trypanosomatids rely extensively on post transcriptional regulatory mechanisms, particularly RNA-binding proteins (RBPs), to control their gene expression patterns (Kramer and Carrington 2011; Kramer 2012; Schwede et al. 2012). This major distinction between trypanosomatids and other eukaryotes including their hosts, renders these pathways as treasure trove for drug development. However, as mentioned above, the current knowledge on this scope is poorly understood, and much more needs to be discovered before being able to effectively target the regulatory pathways. The main goal of this work is to use integrated computational and experimental strategies to chart the RBP-mediated gene regulatory and protein interaction networks which impact the trypanosomatid RNA processing and abundance. We sought to fill this gap by both RBP-centric as well as RNA *cis*-regulatory element (RRE)-centric strategies. Namely, after critically reviewing the current knowledge on the trypanosomatid RREs in the second chapter, we predicted high confidence *cis*-regulatory elements that are potentially involved in the transmission process of the *T. brucei*. In addition, a novel deep sequencing based approach was developed to investigate the effect of RBPs on the tails of mitochondrial transcripts, by which we could uncover new aspects of trypanosomatid mitochondrial transcript processing as well as translational regulation. Moreover, we investigated the network context of trypanosomatid RBPs by charting the first experimentally derived protein co-complex map of *T. brucei* using a combined biochemical-mass spectrometry strategy. Finally, as the last aim, to make the obtained results in this study widely accessible and useful to the researchers in the field, we developed a web-based database that not only can dynamically represent the results in a query-based manner, but also integrate them with other available resources on trypanosomatids.

Due to ease of genetic manipulation as well as high genome similarity with other members of the family, *T. brucei* is the main model organism to study the molecular biology of parasites from Trypanosomatidae family. In this work, we focus on *T. brucei* to investigate the gene regulatory pathways of trypanosomatids. Nevertheless, the high degree of similarity with other members of

the family makes it possible to transfer the developed techniques and the obtained knowledge to the related parasites.

This work is innovative as it proposes novel and multifaceted solutions for the challenging task of trypanosomatid gene regulatory inference that are tailored to the current molecular knowledge on trypanosomatid parasites. Importantly, the developed strategies and methods are applicable to most, if not all, non-model organisms that lack systems-level insights on the underlying gene regulatory and protein interaction networks.

Chapter 1

Literature Review

Trypanosoma brucei

T. brucei is a unicellular parasite of that causes African trypanosomiasis commonly referred to as sleeping sickness in humans and nagana in livestock (Brun et al. 2010). Sleeping sickness occurs in 36 sub-Saharan Africa countries with poor socio-economic status. The fact that *T. brucei* can infect both humans and livestock impedes development of the affected regions and echoes the poverty. The vector-borne parasite *T. brucei* is transmitted by the tsetse fly (Glossina genus). So the parasite has two distinct stages during its life cycle: the procyclic stage (PS) in which the parasite reproduces within the tsetse fly, and the bloodstream stage (BS) which is responsible for disease in the mammalian host (Matthews 2005; Fenn and Matthews 2007). The BS life cycle starts with a specific developmental form of the parasite, known as long slender, that establishes the disease in the mammal host (Rico et al. 2013). Long slender parasites are extracellular, highly proliferative, and evade the host immune system by regular antigenic variation that leads to cyclic symptoms in the host such as fever, headaches, and joint pains (Stich et al. 2002). In the extreme cases of infection, long slender cells can penetrate extravascular tissues, including the central nervous system (Stich et al. 2002). By sensing the high density of parasite cells in the bloodstream of the host, some of long slender cells differentiate into unproliferative, short stumpy form cells that are ready to be taken up by the tsetse fly vector (Proto et al. 2013). *T. brucei* parasites undergo different differentiation processes in the vector as well that not only maximizes their chance of survival in the new environment, but also prepare them to invade a new mammal host. *T. brucei* parasites can be transmitted by 31 species and subspecies of Glossina genus of tsetse. The fly vectors can be classified based on the habitat to the three groups of: fusca group (forest), morsitans group (savannah) and palpalis group (riverine and forest) (Nash et al. 1972; Jordan 1973; Jordan 1978; Dame and Jordan 1981). Once infected by a parasite, the fly remains so for the rest of its life (Welburn and Maudlin 1992).
T. brucei gambiense and *T. brucei rhodesiense* are responsible for human African trypanosomasis that can be fatal if untreated, while *T. brucei brucei* is sensitive to human serum (Simarro et al. 2008; Steinmann et al. 2015). In Africa, approximately 70 million people distributed over an area of 1.55 million km^2 are at risk of infection. More than 97% of African trypanosomasis incidences and chronic infections are attributed to *T. b. gambiense* parasites that can be found in 24 countries in west and central Africa. A person can be infected with this parasite for several years without showing major symptoms. Therefore, it is usually diagnosed in

late stages where the central nervous system infection has occurred. Although *T. b. gambiense* can live in many Glossina species under the lab conditions, the parasite is present almost exclusively in both male and female flies of the palpalis group in nature (World Health 2012). In addition to the transmission by the fly vector, reports indicate the possibility of mechanical transmission of infection such as Congenital (Lindner and Priotto 2010). The main reservoir of *T. b. gambianse* is humans. The average 3 years presence of the parasite in the bloodstream of patients provide more than enough time to complete the human-fly-human life cycle (Checchi et al. 2008). *T. b. rhodesiense* develops rapidly in the body and causes acute disease in the patients with possible death within six months. It is present in 13 countries of eastern and southern Africa and relies exclusively on the fly vector for the transmission (Traub et al. 1978). A single bite by an infected vector is sufficient to infect in the mammal (Thuita et al. 2008). Likewise, taking up only one parasite in a blood meal is sufficient for the infection of the fly (Maudlin and Welburn 1989). *T. b. rhodesiense* is zoonotic and therefore, has non-human reservoirs as well (Wardrop et al. 2010; Clerinx et al. 2012; Wardrop et al. 2012). In advanced cases, *T. b. rhodesiense* parasites can be found in the central nervous system as well (Ueno and Lodoen 2015).

Three epidemics of African trypanosomasis have been reported since late 19th century (One between 1896 and 1906, mostly in Uganda and the Congo Basin; One in 1920 in African countries; The other one started in 1970 and lasted until the late 1990s). However, thanks to the trypanocidal drugs, cases are now decreasing. But, resistance to the currently administered drugs is emerging, indicating the high risk for the resurgence of the corresponding parasitemia (Barrett 2001; Gehrig and Efferth 2008; Barrett et al. 2011; Stich et al. 2013). Although *T. b. gambiense* causes most of recorded infection cases, *T. b. rhodesiense* is the one with highest probability of epidemic due to the lack of effective therapeutics (World Health 2013).

Drugs used for the treatment of human African trypanosomiasis

Melarsoprol: This drug is from the arsenicals family. Melarsoprol is a prodrug (is metabolized within the body into a pharmacologically active drug) and is usually administered for late stage treatments (Kuepfer et al. 2012). Melarsoprol has adverse side effects similar to arsenic poisoning (Williamson 1962). Melarsoprol removes trypanosomes from blood and lymph within one day in most cases. HPLC analysis has led to the discovery of melarsen oxide as the main

melarsoprol metabolite (100). Melarsen oxide has strong protein binding activity and the total serum protein binding of 79% has been measured by ultrafiltration (Keiser et al. 2000). TbAT1 transporter in the parasite is responsible for its uptake, and resistance is emerging by mutations in this promoter (Stewart et al. 2010).

Pentamidine: This drug is from diamidine family and has chemical similarity with phenformin antibiotic. Initially diamidines were used as trypanocidal due to their hypoglycaemic effect on the parasite, i.e., they kill the parasite by its starving from glucose. But later it was discovered that they also have direct trypanocidal activity (King et al. 1937). Pentamidine is usually used for the early stages of the infection. It is soluble but cannot be absorbed orally. The mode of action is still unknown (Wang 1995). It is suggested that medicine might kill the parasite by DNA binding and kinetoplastid disruption (Shapiro and Englund 1990). The main reason for its effectiveness in the parasite is that it can accumulate in *T. brucei* in millimolar concentrations (Damper and Patton 1976). Although resistance to diminazene (another drug from the diamidine family) is widespread, resistance to pentamidine is rare and its clinical efficacy continues to be excellent with the cure rate of 93% up to 98% (Silva 1957; World Health 2013). The low resistance to pentamidine might be due to its uptake by three different transporters in the parasite (Barrett et al. 2007). But the fact that the parasites could become resistance to another drug from the same family can suggest that the resistance to pentamidine can also emerge. This drug extensively binds to lysosomes (Glaumann et al. 1994), leading to its accumulation in tissues including the spleen, the kidneys and the liver (Waalkes et al. 1970). The most adverse side effects include pain in the injection site, hypoglycemia, and hypotension.

Eflornithine in combination with nifurtimox: Eflornithine (DMFO) is used for severe cases of the infection. It is soluble in the water and can be administered orally. Administration of eflorithine is usually with nifurtimox to reduce the treatment time. This drug, unfortunately, is not effective on *T. b. rhodesiense*. Additionally, its associated cost is relatively high and is also hard to administer (e.g. requires nursing care). Resistance to this drug is also emerging.

Suramin: This drug is a polysulfonated naphthyl urea, has strong negative charge, is soluble in the water, and is one of the drugs with highest half-lives (44-54 days) in the human body (Broder

et al. 1985; Collins et al. 1986). Suramin inhibits numerous enzymes, including L-α-glycerophosphate oxidase (Fairlamb and Bowman 1977; Gutteridge 1985), glycerol-3-phosphate dehydrogenase (Fairlamb and Bowman 1980), RNA polymerase and kinases (Hawking 1978), hyaluronidase, urease, hexokinase, fumarase, trypsin (Pepin and Milord 1994) and reverse transcriptase (Cheson et al. 1987), and the receptor-mediated uptake of low-density lipoprotein by trypanosomes (Vansterkenburg et al. 1993). It is one of the main treatments for infections with *T. b. rhodesiense*. It can also be used for onchocerciasis treatment (Chijioke et al. 1998). Side effects include nausea and vomiting.

Characteristics of *T. brucei* gene regulatory pathways

Because of large environmental differences in mammalian host and insect vector, *T. brucei* undergoes extensive metabolic and morphological changes when shuffling between insect vector and mammalian host. It is estimated that 5-25% of genes are differentially expressed between two life-cycle stages (Jensen et al. 2009; Kabani et al. 2009; Queiroz et al. 2009; Kolev et al. 2010; Siegel et al. 2010; Veitch et al. 2010). As members of a highly divergent group of eukaryotes, trypanosomatids are unique in that non-related genes are constitutively co-transcribed into polycistronic units and processed into individual mRNAs by a combined trans-splicing and polyadenylation reaction (Kramer 2012). Thus, regulation of gene expression in trypanosomes occurs almost exclusively through post-transcriptional mechanisms. Although not the focus of this work, the emerging picture of regulatory pathways on trypanosomatids hints on the potentially important roles of post translational (e.g., proteasome in-/dependent degradation, protein phosphorylation) and epigenetics (such as histone and DNA modifications) mechanisms as well (Siegel et al. 2009; Anderson et al. 2013; Urbaniak et al. 2013; Hope et al. 2014; Maree and Patterton 2014; Reynolds et al. 2014). Dynamics of *T. brucei* gene expression patterns are mostly attributed to RNA binding proteins (RBPs), mediating RNA metabolism at multiple levels including polycistron RNA cleavage , mRNA trans-splicing, maturation (e.g., 5'capping and 3' poly-adenylation), localization, turnover, and translation (Milone et al. 2004; Horn 2008; Gunzl 2010; Nilsson et al. 2010; Kramer and Carrington 2011; Dostalova et al. 2013; Gupta et al. 2013; Jensen et al. 2014; Vasquez et al. 2014; Wilhelm et al. 2014; Buhlmann et al. 2015; Moura et al. 2015). The RNA regulon model suggests that a system analogous to prokaryotic DNA operons exists at the RNA level, whereby functionally related messages are regulated by

RBPs that bind short RNA sequence and/or structural patterns, known as RNA regulatory elements (RREs) (Keene 2007b; Jankowsky and Harris 2015).

RBPs are the key regulatory components in the RNA metabolism processes (Gazestani et al. 2014; Gerstberger et al. 2014b; Jankowsky and Harris 2015; Matia-Gonzalez et al. 2015; Castello et al. 2016; Li et al. 2016). Recent experimental data on other organisms like human, indicate the unexpected involvements of RBPs in pathways such as DNA damage response, human stem cell differentiation, and remodeling the protein interaction networks (Ellis et al. 2012; Gerstberger et al. 2014b; Hudson and Ortlund 2014; Gueroussov et al. 2015). Moreover, it is shown that many transcription factors can recognize and consequently bind to both RNA and DNA molecules, representing them as special type of RBPs (Hudson and Ortlund 2014). The RNA recognition is mediated by one or more protein domains present in RBPs. The most common domains are RNA Recognition Motif (RRM) and hnRNP K-homology (KH) domains (Lukong et al. 2008; Gerstberger et al. 2014a). However, recent studies suggested existence of many RBPs with no identifiable RNA binding domain, reflecting the high diversity of the RBPs and lack of knowledge in this area (Baltz et al. 2012; Castello et al. 2016). RRM domains are around 90 amino acids long, have a conserved structure composed of two consensus ribonucleoprotein motifs [RNA-binding protein 1 (RNP1) and RNP2] and bind to specific sequences in single stranded RNAs (Deo et al. 1999). KH domains are about 70 to 100 amino acids long, are highly conserved, and similar to RRM domains can recognize and bind to specific sequences in single stranded regions of transcripts (Lewis et al. 2000). The name of KH domain originates from its first discovery as a repeated region in the heteronuclear ribonucleoprotein particle (hnRNP) K (Lewis et al. 2000).

RREs are usually short (less than 10nt) and highly degenerate (Ray et al. 2013; Alipanahi et al. 2015). Hence, there exist potentially hundreds or even thousands instances of a motif in the transcriptome. Features other than sequence motif help RBPs to discern true RNA targets from random occurrences of a motif. Perhaps, the most important feature is the structural context that a motif resides in (Kligun and Mandel-Gutfreund 2015). The double stranded form of RNA forms a tighter structure compared to its counterpart in DNA, making nucleotides present in the duplex inaccessible for recognition and binding by RBPs (Draper 1999; Allers and Shamoo 2001). Therefore, RBPs that recognize their targets through a specific sequence patterns require the pattern to be in the single stranded region. However, it is not a requirement for the binding of

RBPs that recognize the structural elements (some are even specialized in the recognition of double stranded RNA) (Li et al. 2010; Li et al. 2014). The next important factor is the interactions that an RBP makes with other proteins. Such interactions can stabilize the binding of the RBP to its target (Keene 2007a; Hogan et al. 2008). Therefore, many RBPs operate in the context of ribonucleoprotein (RNPs) complexes. The other strategy for RBPs to discriminate the true RNA target is to have multiple RNA binding domains, each recognizing a specific RNA pattern. RBPs are also known to have both synergic as well as competition in binding with miRNAs, the other class of post transcriptional regulators (Jens and Rajewsky 2015). The combination of above mentioned factors forms the RNA operon of a cell (Keene 2007a; Keene 2007b).

Although the post-transcriptional gene regulatory programs of *T. brucei* occur in many different layers, they can be categorized in two distinct programs based on the target mRNAs, a) those regulating the fate of nuclear-encoded mRNAs and b) mechanisms that mediate the RNA editing process of mitochondrial-encoded mRNAs. Nevertheless, considering the cell as a complex and highly inter-connected biological system (Kitano 2002; Kitano 2004), these two regulatory programs are linked to each other and, therefore, affecting each of them can induce changes in the other one.

Post-transcriptional regulation of nuclear-encoded mRNAs

Trypanosomatid parasites have highly similar nuclear genomes. As an illustration, according to the OrthoMCL database, 90% and 78% of protein coding genes in *T. brucei* have orthologues in *T. cruzi* and *L. major Friedlin* organisms, respectively. Based on the latest annotations, *T. brucei* nuclear genome is composed of 11 chromosomes, ranging between 1Mb upto 5Mb in size (Aslett et al. 2010). These chromosomes encode for 12,094 genes (including likely pseudogenes), 66 tRNAs, and 106 rRNA molecules. Trypanosomatid subtelomers constitute the most repetitive regions of their genomes containing numerous high-copy number genes, as well as simple and complex sequence repeats. Variable surface glycoproteins (VSGs) are a group of genes highly enriched in subtelomeric regions of *T. brucei* genome. With estimated reservoir of more than 800 copies, this level of diversity in VSGs helps the parasite to effectively escape the mammalian

immune system. It is worth noting that the number of chromosomes and annotated genomic features varies among trypanosomatid parasites. *L. major*, for example, has 36 chromosomes. Moreover, in contrast to annotation of human genome, the numbering of chromosomes starts from the smallest ones in terms of size. So chromosomes with lower numbers (e.g. chromosome 1) have smaller size than the ones with bigger numbers (e.g., chromosome 11). In terms of gene annotations, 4114 *T. brucei* protein genes have annotated Gene ontology – biological process annotation (622 distinct terms) (Aslett et al. 2010). Moreover, 1221 genes are assigned to 93 different KEGG pathways (Kanehisa et al. 2014). However, most of these annotations are computationally inferred using sequence homology-based approaches. For example, out of 4114 genes with annotated Gene ontology biological process terms, only 741 genes have experimental evidence for their annotations (279 experimentally determined annotation terms). Finally, 4899 genes are predicted to have at least one functional domain according to the Interproscan webservice. Of these, 46 proteins have zinc finger-CCCH and Zinc finger-double stranded RNA binding domains, 22 proteins have zinc finger-C2H2 domain, 74 proteins have RRM domain, and nine proteins have KH and KH-like domains. There are also four proteins with alba functional domain that form heterodimers with each other and are associated with the polysomes. Moreover, ten proteins are identified to contain pumilio domain. However, the function of most of these proteins is not known yet. Puf9 is the well characterized member of this group and is involved in the cell cycle regulation of the parasite (Archer et al. 2009).

Nuclear genes in trypanosomatids are transcribed as polycistronic mRNAs that are further processed via trans-splicing, involving a polypyrimidine tract as the signal for splicing (Martinez-Calvillo et al. 2010; Kramer 2012). Because of the polycistronic transcription of often unrelated mRNAs, regulation of gene expression in trypanosomatids is mainly at the post-transcriptional level (Schwede et al. 2012). Regulation of nuclear-encoded mRNAs largely occurs in the nucleus and cytoplasm compartments. Nuclear post-transcriptional events play a major role in the mRNA maturation (e.g., trans-splicing) and mRNA trafficking (from nucleus to cytoplasm). Recent studies suggest that regulatory programs in the nucleus can be a major player for controlling the decay rate of nuclear-encoded mRNAs as well (Fadda et al. 2014). The regulatory events occurring in the cytoplasm are mostly related to the regulation of mRNA in terms of stability, localization, and translation. Evidence suggests that the regulation of nuclear-encoded mRNAs is the main strategy utilized by the parasite for the triggering or suppression of

developmental changes (Kolev et al. 2012; Wurst et al. 2012; Clayton 2014; Erben et al. 2014b; Gazestani et al. 2014; Mony et al. 2014). As an illustration, it has been shown that over-expression of cytoplasmic RBPs known as RBP6 and RBP4 helps the differentiation process in the tsetse fly and mammalian host life stages, respectively (Kolev et al. 2012; Mony et al. 2014). Conversely, it is reported that cells over-expressing RBP10 were unable to differentiate from BS to PS (Wurst et al. 2012). Trypanosomatid RBPs mediate other major cell processes as well, including cell cycle, cell membrane, heat shock, and cell viability (Archer et al. 2009; Fernandez-Moya et al. 2014; Jha et al. 2015). Consequently, small scale gene regulatory circuits for some specific pathways are emerging (Erben et al. 2014a; Lueong et al. 2016). However, RREs involved in the environmental response or the differentiation processes of the parasite are still mostly unknown. In the second chapter, we critically reviewed the current available knowledge on the trypanosmatid RBPs and their cognate RREs. We also discussed the advantages and disadvantages of computational and experimental strategies that have been employed so far to decode these regulatory pathways. In the third chapter, to expand our knowledge on the regulatory logics that dictates the fate of trypanosomatid transcripts, we developed a graph based strategy for reconstruction of an RRE-based gene regulatory network for *T. brucei* in a genome-wide unbiased setting to explain the observed changes in the gene expression levels of *T. brucei* genes during the developmental processes or environmental changes.

Post-transcriptional regulation of mitochondrial-encoded mRNAs

To respond and adapt to the highly variable environments encountered during its life cycle, *T. brucei* undergoes drastic morphological and molecular changes. Of these, changes in molecular pathways associated with energy metabolism in mitochondria are of the highest importance, making this organelle of interest as a therapeutic target. The mitochondrial genome consists of the 15% of the total cellular DNA. This genome is condensed in an organelle called “kinetoplast”. The kinetoplast DNA contains two different types of circular DNA of different sizes, the maxi- and the minicircles (Simpson 1987). The maxicircle is structurally and functionally equivalent to the mitochondrial genome in other eukaryotic organisms (Shapiro and Englund 1995; Shlomai 2004) and encodes 2 and 18 mitochondrial rRNAs and mRNAs, respectively. Similar to nuclear-encoded genes, post-transcriptional gene regulatory processes

play critical roles in controlling the gene expression of mitochondrial genes. In brief, mitochondrial gene expression starts with polycistron transcription. Cleaved transcripts are then subjected to a tailing process with impact on their stability (Etheridge et al. 2008; Aphasizheva et al. 2011). As explained below, tailed transcripts may also undergo editing-process in which translatable mRNAs are produced. Finally, a secondary tailing process marks the transcripts with correct open reading frames (ORFs) for the translation.

Most (78%) of the mitochondrial encoded transcripts are unusual and they need to be edited in order to be translated (Hashimi et al. 2013; Read et al. 2016). These cryptogenes are remodelled by using information of the 3'-uridylated guide RNAs, which are mainly encoded by minicircle genes. RNA editing is a process in which minicircle-encoded guide RNAs (gRNAs) dictate the insertion, and less frequently deletion, of a defined number of uridine residues at specific positions of precursor mRNAs turning them into the mature mRNAs. Editing creates initiation and/or termination codons, corrects frameshifts and, in the case of pan-edited mRNAs, creates an entire reading frame. The editing process is essential for the survival of both developmental stages, although the precise role of editing is not well-known in BS (Rusche et al. 2001; Schnaufer et al. 2001). Editing of mitochondrial RNA is developmentally regulated in a transcript-specific manner in *T. brucei* (Feagin et al. 1987; Koslowsky et al. 1990; Read et al. 1992). For example, edited CYb and cytochrome oxidase subunit II (COXII) are not detectable in BS (Feagin and Stuart 1988), while their edited mRNAs are highly abundant in PS or as another example, the abundance of fully edited NADH dehydrogenase subunit 7 and 8 mRNA is increased in BS compare to PS (Souza et al. 1992). This regulation occurs in coordination with the activities of the trypanosome mitochondrion during the developmental cycle in order to take advantage of the changing environmental conditions. The molecular basis for this developmental control is still unknown. In the second aim, we develop a high-throughput sequencing based approach to investigate the tail of trypanosomatid mitochondrial transcripts. In the third aim, using a large-scale RBP-centric strategy, we study the RBPs involved in the RNA editing process of *T. brucei*. However, the underlying mechanisms for the assembly as well as developmental regulation of RNA editing machinery is still elusive. Here, we chart a experimentally derived co-complex map of trypanosomatid proteins. The derived network was used to not only identify new components, but also investigate the composition of RNA editing machinery.

A post-transcriptional event impacting multiple other post-transcriptional processes is addition of 3' non-encoded tails to all RNA species in the mitochondria. rRNAs and regulatory RNAs involved in editing (guide RNAs) are oligouridylated, and mRNAs possess tails consisting of both adenosine (A) and uridine (U). Trypanosome mitochondrial mRNA tails fall into two distinct categories with differences in their general sizes and biological functions. Included in the first category are the fairly ubiquitous oligomer tails initially added to mRNAs called (in)itial tails, or "in-tails"; these consist of poly(A), poly(U), or a combination thereof, added by the poly(A) polymerase KPAP1 (Etheridge et al. 2008) and the terminal uridyltransferase RET1 (Aphasizhev et al. 2002; Aphasizhev et al. 2003). In-tails are recognized as stability elements (Ryan et al. 2003; Kao and Read 2005; Etheridge et al. 2008; Aphasizheva and Aphasizhev 2010), although the mechanisms by which in-tails regulate mRNA stability are not known. Sequencing of in-tails suggests transcript-specific variation in sequence composition and length (Souza et al. 1992; Kao and Read 2007; Aphasizheva and Aphasizhev 2010; Zimmer et al. 2012); also consistent with their regulatory roles.

The second tail category, "ex-tails", is longer tails with probable translational roles, generated by extensions appended to a subset of in-tails on translatable mRNAs only. The nucleotide extensions are fairly homogenous in A/U composition with KPAP1/RET1 addition of A and U (a 7:3 A/U ratio (Etheridge et al. 2008)) controlled by the pentatricopeptide protein complex KPAF1/KPAF2 (Aphasizheva et al. 2011). Extensions have a fairly consistent frequency of switching of addition from A to U and back. However, due to the complex nature of mitochondrial gene regulatory pathways in trypanosomatids, the impacts of tail variations on the fate of transcripts are still poorly understood. Here, we develop a novel deep sequencing based approach that can isolates the tail addition process form regulatory processes, allowing to not only investigate the direct impact of tail on transcripts, but also correlate the occurrence of the process with other regulatory processes such as RNA polycistron cleavage and RNA editing.

Chapter 2

Deciphering RNA regulatory elements in trypanosomatids: one piece at a time or genome-wide?

Morphological and metabolic changes in the life cycle of *Trypanosoma brucei* are accomplished by precise regulation of hundreds of genes. In the absence of transcriptional control, RNA-binding proteins (RBPs) shape the structure of gene regulatory maps in this organism, but current knowledge about their target RNAs, binding sites, and mechanisms of action is far from complete. Although recent technological advances have revolutionized the RBP-based approaches, the main framework for the RNA regulatory element (RRE)-based approaches has not changed over the last two decades in *T. brucei*. In this chapter that is published as an opinion in Trends in Parasitology journal (Gazestani et al. 2014), after highlighting the current challenges in RRE inference, we explain some genome-wide solutions that can significantly boost our current understanding about gene regulatory networks in *T. brucei*.

Approaches to study post transcriptional gene regulatory pathways

As members of a highly divergent group of eukaryotes, trypanosomatids are unique in that non-related genes are constitutively cotranscribed into polycistronic units and are processed into individual mRNAs by a combined trans-splicing and polyadenylation reaction (Kramer 2012). Thus, regulation of gene expression in trypanosomes occurs almost exclusively through post-transcriptional mechanisms. Such processes controlling mRNA localization, turnover, and translation are mediated in a large part by the dynamic interactions of RBPs with specific subpopulations of mRNAs (Kramer and Carrington 2011).

Recent studies in trypanosomes have characterized the role of several RBPs in various cellular processes, from cell cycle progression to differentiation between life stages (Clayton 2013). Pumilio/fem-3 binding factor 9 (PUF9), for example, is necessary for the function of the replicative processes in the early G2 phase of the cell cycle (Archer et al. 2009). RBP10 regulates bloodstream form-specific genes, and its depletion causes increases in mRNAs associated with the early stages of differentiation (Wurst et al. 2012). Overexpression of RBP6 induces transformation from insect-form procyclic cells to infective metacyclic forms (Kolev et al. 2012). Although an abundance of putative RBPs have been identified in trypanosomes, based on conserved RNA-binding domains, less is known about the RREs that are recognized in the RNA targets.

RBP-mediated regulatory maps in the genome can be studied using two main methods: RBP- and RRE-based approaches. In an RBP-based approach, the binding site of a specific RBP is defined using various experimental methods, such as individual nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) and RBP-immunoprecipitation coupled to high-throughput sequencing (RIP-seq) (Konig et al. 2010; Zhao et al. 2010). While these valuable methods have permitted detailed analyses of binding site information, the procedures are labor-intensive and limited to the study of functional elements for one RBP at a time. In this case, the RBP of choice must be selected based on preliminary knowledge about its requirement or functionality in a biological process of interest. However, in many cases, this type of information is not available; of the more than 150 putative RBPs in trypanosomes (De Gaudenzi et al. 2005; Caro et al. 2006; Kramer et al. 2010a), only a small proportion has been characterized. Consequently, many RREs are also uncharacterized. The RRE-based approach circumvents this issue by trying to identify biologically relevant RREs outside the context of a pre-specified RBP.

Recurrent challenges in studying RREs

Current RRE-based studies in trypanosomatids are mostly limited to gene-based approaches, which search for RREs in the 3′-untranslated region (UTR) of a limited number of genes (usually one single gene) at a time. By contrast, genome-wide approaches search for common regulatory sequences in sets of co-regulated genes. The most popular experimental RRE-based approaches in trypanosomatids are composed of serial deletions or mutational analyses on the 3′-UTRs of target genes to gradually narrow down regions containing regulatory elements (Jefferies et al. 1991; Hehl et al. 1994; Furger et al. 1997; Hotz et al. 1998; Mayho et al. 2006; MacGregor and Matthews 2012; Monk et al. 2013). The effect on the expression of the target gene or a reporter construct is measured following truncations or random deletions of 3′-UTR regions. Although deletion analyses can yield a stretch of nucleotides in which regulatory elements reside, they fall short of defining the exact consensus sequence of RREs. For example, MacGregor and Matthews narrowed down gene regulatory signals responsible for the gene expression of PAD1 to distal and proximal repression elements in the 3′-UTR, but were not able to identify the sequence or structure of the regulatory signals (MacGregor and Matthews 2012). In some cases, the same approach led to the identification of a relatively short regulatory element (between 16 and 34 nt) in the 3′-UTRs of *Trypanosoma brucei* (Hehl et al. 1994; Hotz et al. 1998; Monk et al. 2013). However, a large-scale study on RBPs in a wide range of organisms has indicated that RBPs usually bind to still shorter sequence patterns, around 7 nt in length, and show some level of degeneracy in their consensus sequence (Ray et al. 2013). In addition, crystal structures have confirmed that the actual lengths of RNA sequences recognized by individual RNA binding domains span only a few nucleotides: 4 nt for a CCCH zinc finger(Hudson et al. 2004), 1 nt for each pumilio repeat (Wang et al. 2002), and 2-8 nt for an RNA recognition motif (RRM) (Clery et al. 2008a). Evidence suggests that this might also be the case in trypanosomatids. For example, Archer et al. identified a 7 nt motif that maintains cell cycle-dependent expression of a PUF9 target gene (Archer et al. 2009). Another 8 nt RRE within the EP1 procyclin 26mer element putatively targets a wide range of *T. brucei* transcripts (Mayho et al. 2006).
Given that co-expressed genes may contain common regulatory motifs, gene-based approaches could uncover regulatory networks beyond the gene of interest. However, this can be complicated by the presence of higher order structures and multiple regulatory regions in the 3′-UTR. The search for single regulatory motifs is further hindered by the notion that RREs are

generally 10 nt or less and that trypanosomatid 3′-UTRs average 400 nt. This gives a virtually endless number of combinatorial possibilities for deletion experiments. In the case of a 34 nt regulatory region identified in the 3′-UTR of expression site associated gene ESAG9-EQ, which controls its expression in the bloodstream form, the regulatory element was not common to other developmentally co-regulated transcripts of the ESAG9 family (Monk et al. 2013). Without high resolution mapping of the elements present in each regulatory region or the possibility of extending a regulatory element found in one gene to other co-regulated genes of the same family or function, it is difficult for gene-based approaches to contribute to the understanding of regulatory networks.

Alternative genome-wide approaches

Genome-wide mapping of RREs is essential for a systems-level understanding of the *T. brucei* gene regulatory network. The availability of genome sequence for many organisms along with current advances in high-throughput technologies, especially deep sequencing, has revolutionized the field and led to the development of various accurate, genome-wide experimental strategies for finding functional elements in the genome (Zinzen et al. 2009; Lee et al. 2011; Arnold et al. 2013). It also contributed to the development of various powerful computational tools for comprehensive annotation of RREs (Xie et al. 2005; Elemento et al. 2007; Foat and Stormo 2009; Goodarzi et al. 2012). When accurate, computational approaches are extremely beneficial, mainly because they need minimal experimental requirements for the validation of predictions. A large scale study has shown that the results of these computational approaches are highly reliable and the newly predicted motifs can be recognized by RBPs in the genome (Hu et al. 2009). Additionally, several genome-wide studies have led to the successful prediction of RREs in *T. brucei*, reflecting the potential power of these approaches in this organism (Archer et al. 2011; Najafabadi et al. 2013). Here, we discuss two genome-wide computational methods for predicting RREs, and argue that they can be used as a reliable starting points for understanding the mechanisms that underlie gene regulation in *T. brucei*.

Comparative genomics

Several computational approaches seek conserved RREs in sets of orthologous genes by assuming that part of the regulatory circuits are conserved among closely related organisms

(Meireles-Filho and Stark 2009). These approaches can be categorized into two different classes: alignment-based and alignment-free approaches. Alignment-based approaches try to find conserved regions in the 3′-UTRs by performing whole genome multiple alignments. Studies have shown that these conserved regions are enriched for the RREs (Liu et al. 2004; Xie et al. 2005). To test the efficiency of this approach in trypanosomatids, we extracted whole genome multiple alignments from TriTrypdb v.5 (Aslett et al. 2010) considering 16 different trypanosomatid organisms. The results indicated that the 3′-UTRs in these organisms are highly diverged over the course of evolution (Fig 2.1). Repeating the same analysis considering only eight trypanosome organisms did not change the overall multiple alignment performance. The reliability of alignment-based approaches is highly dependent on the quality of the underlying multiple alignments. This is because the functional sites are very small (usually less than 10 nt) compared with the total 3′-UTR length (median length of 400 nt in *T. brucei* (Siegel et al. 2010)). Therefore, highly diverged 3′-UTRs can easily lead to erroneous multiple alignments and, consequently, lack of identification of functional elements. To alleviate this issue, alignment-free approaches have been developed (Chan et al. 2005).

These approaches examine whether orthologous genes in related organisms tend to have a specific motif, disregarding the location of the motif's occurrence (Fig 2.2). Based on this concept, a novel approach identified 222 linearly- and 166 structurally conserved RREs. Some of the predicted motifs (both structural and linear) overlap with previously known RREs in *T. brucei*, reflecting the conservation of some regulatory interactions among trypanosomatids (Najafabadi et al. 2013).

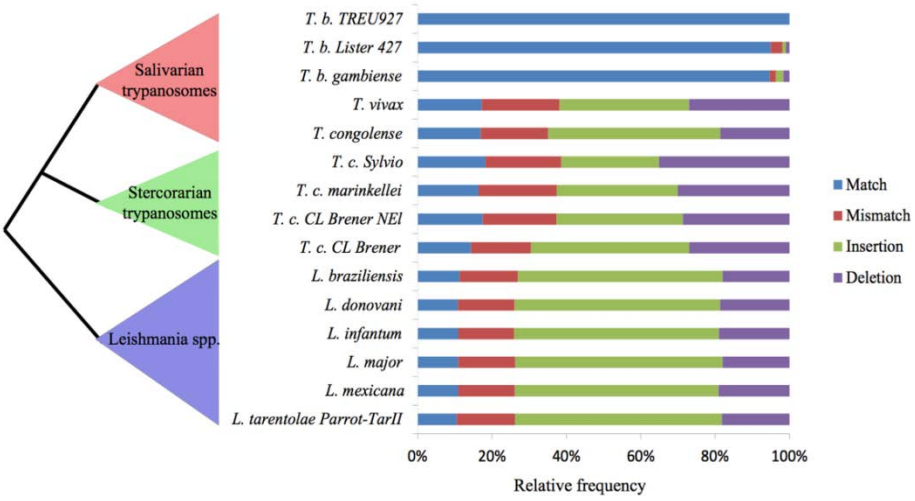


Fig 2.1. Conservation analysis of 3′-UTRs among 16 trypanosomatids. The 3′-UTRs of *T. brucei TREU927* were defined as the median length reported in (Siegel et al. 2010). For genes with no identified 3′-UTRs, 400 nt downstream of the translational stop codon were chosen as the 3′-UTRs. Considering *T. brucei TREU927* as the reference genome, whole genome multiple sequence alignments corresponding to the 3′-UTRs were extracted from TritrypDB v.5. To improve the accuracy of multiple sequence alignments, the extracted regions were realigned using the ClustalO program. Although 3′-UTR conservation among *T. brucei* sub-species (*T. brucei TREU927*, *T. brucei Lister 427*, *T. brucei gambianse*) is very high, alignment of the 3′-UTRs of *T. brucei TREU927* with other trypanosomatids are poorly conserved with many insertions and deletions in these regions. However, as expected, the 3′-UTRs of *T. brucei* shows higher similarity to other trypanosome organisms (*T. vivax*, *T. congolense*, and *T. cruzi*) compared with those of Leishmania. The phylogenetic relationships were extracted from (Hamilton et al. 2012a).

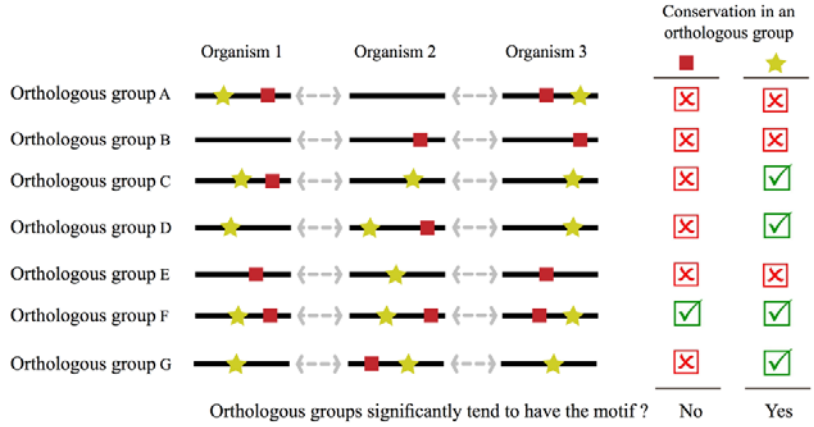


Fig 2.2. Conceptual representation of alignment-free comparative genomics approaches. After grouping orthologous genes together in a set of closely related organisms (grouped via two sided arrows in this figure), alignment free approaches search for conserved RREs within the 3′-UTRs, regardless of their position. The gene regulatory network conservation assumption implies that each orthologous group share a common set of RREs. Therefore, we expect to observe a significant number of orthologous groups that contain a functional motif (represented by yellow star). In contrast, the distribution for non-conserved sequences is expected to be random (represented by red square).

Expression-based approaches

The underlying assumption in expression-based approaches is that regulatory interactions are reflected in the whole genome transcriptome data. Based on this assumption, these approaches seek enriched motifs in sets of coordinately expressed genes (Elemento et al. 2007; Goodarzi et al. 2012) (Fig 2.3). The advantage of expression-based approaches over the comparative based approaches is that the former is organism specific and does not need to assume the conservation of gene regulatory networks in a set of organisms. Relaxation of this constraint is beneficial because biological studies indicate considerable differences among trypanosomatids in terms of host, lifestyle, and developmental stages. Supported by both principle (Carroll 2005) and experiment (Kunarso et al. 2010), gene regulatory network rewiring serves as a major source of evolutionary innovation (Weirauch and Hughes 2010). Interestingly, binding preference is remained conserved among orthologous transcription factors (Noyes et al. 2008) and RBPs (Ray et al. 2013). Therefore, the most likely source of divergence in gene regulatory networks is gain or loss of functional binding sites in the genome (Wittkopp and Kalay 2012). Thus, just looking at the conserved regulatory map utilizing comparative genomics approaches will lose a considerable number of RREs.

One variant of expression-based approaches was successfully applied to identify RREs involved in the cell cycle progression of *T. brucei* (Queiroz et al. 2009). However, there have been many more successful applications of these approaches in other organisms. One of the limitations of these approaches, which have hampered their wider applications to *T. brucei*, is their need for comprehensive transcriptome data for the inference. To address this issue, RREs were identified computationally by integrating three available transcriptome datasets from *T. brucei*. This analysis led to the prediction of 14 significant RREs (Shateri Najafabadi and Salavati 2010). Comparison of predicted RREs with current experimental data (Ray et al. 2013) indicated that three of the predicted RREs were significantly similar to the experimental motifs. Although limited, this comparison suggested that the integration procedure was successful to some extent.

The power of expression-based approaches is highly dependent on the expression data that is used for inference. Ideally, the expression data should

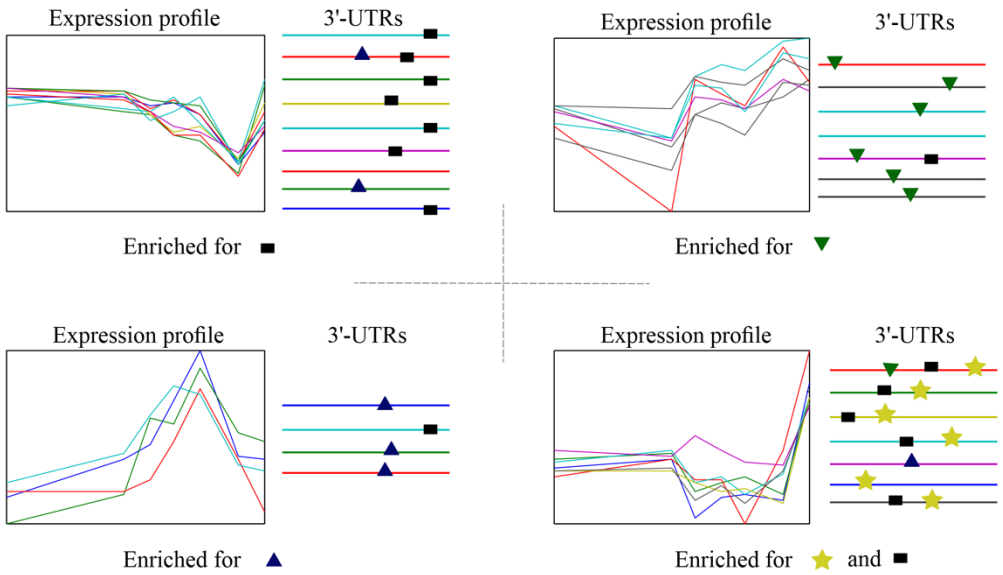


Fig 2.3. Schematic representation of the framework employed in most expression-based approaches. Although some expression-based approaches make predictions based on one single cell state (Foat et al. 2005; Foat and Stormo 2009), most of these approaches make predictions by specifying clusters of genes that are co-regulated with each other in a wide range of conditions (Elemento et al. 2007; Goodarzi et al. 2012). To find these clusters, the latter approaches benefit from the fact that the co-regulation of a set of genes will lead to their co-expression. Therefore, the regulatory regions of co-expressed genes are enrichment for the binding site(s) of regulator(s). Regardless of details, as illustrated in this figure, the co-expression based approaches search for motifs (represented as star, rectangular, and triangle) that are over-represented in sets (four clusters in this figure) of coherently expressed genes (The expression pattern and 3′-UTRs of genes are represented as colored

have sufficient resolution for discriminating various gene regulatory circuits from each other. Naturally, this amount of data can be obtained by extensive cell perturbation studies in different life stages and detailed temporal monitoring of gene fluctuations during the developmental processes.

STARR-seq

Recently, a genome-wide approach, termed self-transcribing active regulatory region sequencing (STARR-seq), has been developed to identify functional regions in the genome (Arnold et al. 2013). In this approach, which was originally developed to identify functional enhancers in the genome, a library of genomic sequences is created by random shearing of a genome into relatively small fragments of several hundreds of nt in length. The fragments are then inserted into the 3′-UTR of a reporter gene. The authors showed that, because the functions of enhancers are mostly independent of their location in the genome, the regulatory effect of each inserted region will be reflected in the expression level of the cognate reporter gene, which is monitored

by sequencing techniques. This approach has several advantages over the expression-based approaches. For example, because of its design, this approach can search the complete genome for finding RREs whereas expression-based approaches usually search only in the 3′-UTRs. Recent CLIP-seq data in *T. brucei* demonstrated that RREs can reside inside the coding sequence (Das et al. 2012). Besides, unlike the motifs predicted by expression-based approaches, there is no need to use a secondary tool to discriminate the functional instances of the motif from the non-functional fraction. Application of a similar but adapted approach in different life stages of *T. brucei* could help to create context-specific maps of functional RREs.

Profiling RNA secondary structure

Accumulating evidence highlights the importance of RNA secondary structure on its function and regulation (Wan et al. 2011). Some RBPs can recognize and bind to specific secondary structures in RNA targets (Oberstrass et al. 2006; Tadros et al. 2007). Additionally, a large-scale analysis of human RBPs has revealed that RBPs usually bind to single-stranded regions in the transcripts (Li et al. 2010). These findings are also supported in *T. brucei*. Although some RREs are associated with conserved secondary structures, the others are mostly functional in the single-stranded regions. Besides, fluctuation in temperature can lead to changes in the RNA secondary structure and, consequently, activation of some RREs in *Leishmania* (David et al. 2010). Advances in sequencing technology have allowed the development of several genome-wide approaches for profiling the RNA secondary structures (Kertesz et al. 2010; Underwood et al. 2010). These approaches benefit from the fact that some RNases are able to distinguish between single- and double-stranded RNA. For example, in one of these approaches, termed parallel analysis of RNA structure (PARS), the transcriptome library is treated separately with two distinct RNases that show preferential digestion toward either the single- or double-stranded RNA. After sequencing the treated samples, the structure for each nucleotide is determined by comparing its frequency in the single-stranded with the double-stranded sample. Although these approaches cannot capture the dynamic structure of RNA under in vivo conditions, they can enhance our knowledge of the structural properties of the *T. brucei* transcriptome.

Validating RRE predictions

Predicted RREs must be followed up with experimental and computational analyses not only to confirm the functionality but also to help to uncover the RBP related to each RRE.
To achieve this goal one may need to consider all possible sources of evidence for prediction of RREs. For example, application of a novel alignment-free approach, the Conserved Structural Motif Search Tool (COSMOS), identified 388 potentially conserved RREs (Najafabadi et al. 2013). Subsequent use of the available expression data led to the identification of those RREs that were more likely to contribute to mRNA abundance and stability. This procedure limited the number of candidates to 35 high-confidence RREs among the 388 elements. For experimental validation of newly found motifs, the authors selected a highly conserved adenylate/uridylate-rich element (ARE), AUUUAUU, which was predicted to be functional by both conservation and expression criteria (Najafabadi et al. 2013). Downstream bioinformatics analysis revealed that the predicted motif by COSMOS was generally associated with transcripts upregulated in late insect stages and downregulated in mammalian bloodstream stages, suggesting a regulatory role similar to the EP1 procyclin ARE but in contrast to another ARE involved in heat-shock response (Quijada et al. 2002; Haile et al. 2003; Droll et al. 2013). Further bioinformatics analyses uncovered three possible trans-acting factors that may be involved in the regulation of this ARE: RBP6 and the double-stranded RNA-binding domain proteins DRBD12 and DRBD13 (Najafabadi et al. 2013). Experiments that followed these predictions provided support for the functional interaction of these three proposed RBPs with the motif (Najafabadi et al. 2013). In particular, expression of each protein was inhibited by RNA interference (RNAi) and activated by overexpression. Consistent effects were observed in these two types of experiments; target transcripts that were downregulated after protein inhibition were upregulated after protein activation, and vice versa, suggesting specific effects of inhibition/activation. These experiments revealed a stabilizing role for DRBD13 and a destabilizing role for RBP6 and DRBD12. Additionally, RBP immunoprecipitation followed by sequencing (RIP-Seq) confirmed that one of the candidate proteins, DRBD13, is associated with trypanosomatid ARE-containing transcripts in vivo (Najafabadi et al. 2013). Alternatively, associated regulatory protein factors can be identified using RREs as ligands to pull down ribonucleoprotein complexes. Experimental approaches for isolating associated proteins involve the exogenous expression of transcripts consisting of small RNA tags, or aptamers, fused to the RNA sequence of interest (Windbichler

and Schroeder 2006; Walker et al. 2008). These RNA aptamers are then recognized and pulled down by a protein or small molecule that recognizes the aptamer. For example, the recently developed technique of RNA-binding protein purification and identification (RaPID) exploits the specific interaction between the E. coli bacteriophage coat protein (MS2-CP) and its cognate RNA. The MS2-CP is fused to a streptavidin-binding protein tag, which allows affinity purification using streptavidin-conjugated beads and subsequent analysis by mass spectrometry (Slobodin and Gerst 2010). A similar method successfully identified several known and novel ARE-binding proteins in mammalian cells using a streptavidin-binding aptamer fused to mRNA containing the ARE of mouse tumor necrosis factor α (Leppek and Stoecklin 2014).

Concluding remarks

The essential point in understanding the biology of trypanosomatids is uncovering the mechanisms by which the coordination of RREs and RBPs orchestrate the expression of target genes. Large-scale identification of RREs and their cognate protein factors would contribute to the construction of gene regulatory networks. Functional characterization of these networks and their role in cellular processes is essential for ultimately identifying points at which these networks can be manipulated pharmacologically to interfere with parasite development and transmission.

Chapter 3

Deciphering RNA Regulatory Elements Involved in the Developmental and Environmental Gene Regulation of *Trypanosoma brucei*

As elaborated before, *T. brucei* relies extensively on fine regulation of gene expression to respond and adapt to variable environments, with implications in transmission and infectivity. However, the involved regulatory elements and their mechanisms of actions are largely unknown. In the previous chapter, we highlighted the strengths of genome-wide transcriptome-based approaches for revealing the gene regulatory maps. In this chapter, published as a research paper in PLoS One (Gazestani and Salavati 2015), we present a novel approach based on the systematic integration of different transcriptome data sources, to predict RNA regulatory elements involved in the gene regulation of *T. brucei* parasite.

Application of our approach led to the prediction of 88 RNA regulatory elements for *T. brucei* with very low estimated false discovery rate. To date only a small fraction of RREs in *T. brucei* have been identified, yet eleven of our predicted motifs strikingly resemble experimentally-derived trypanosomatid regulatory elements. Our follow up analysis on these motifs demonstrated that not only consensus pattern of the predicted motifs match with the functional RREs, but also they show significant enrichment towards similar set of RNAs, demonstrating the high accuracy of the employed approach. Our results also suggested existence of an intricate and intertwined regulatory relationship between some of the regulatory elements and, consequently, their cognate RNA binding proteins. Moreover, the sequence characteristics of the predicted motifs highlighted the importance of A- and/or U-rich elements in the gene regulatory network of *T. brucei*. Importantly, these A- and/or U-rich motifs show distinct transcriptome and proteome responses to the life cycle changes of the parasite, suggesting their diverse regulatory roles. Lastly, comparison with previously predicted motifs on *T. brucei* suggested the superior performance of our approach based on the current limited knowledge of regulatory elements in *T. brucei*.

Background

Various computational approaches have been developed and applied for the genome-wide identification of RREs (reviewed in (Li et al. 2014)). In particular, approaches based on whole genome expression profiling have proved powerful to infer these elements, leading to the identification of many established, as well as new, RREs (Hughes et al. 2000; Bussemaker et al. 2001; Foat et al. 2005; Elemento et al. 2007; Foat and Stormo 2009; Goodarzi et al. 2012). Experimental results substantiate the view that many of the newly identified regulatory elements by these approaches are functional and can be recognized by the proteins on the genome (Hu et al. 2009).

Some expression-based computational approaches make predictions based on a single transcriptome experiment (Bussemaker et al. 2001; Foat et al. 2005; Foat and Stormo 2009), while others decipher RREs by seeking enriched or informative motifs in sets of genes with common regulators (Hughes et al. 2000; Elemento et al. 2007; Goodarzi et al. 2012). To find co-regulated genes, the latter approaches group genes according to their expression patterns based on a comprehensive transcriptome dataset that covers a wide range of diverse biological conditions. Although powerful, the lack of comprehensive transcriptome data has greatly hampered their application on non-model organisms including trypanosomatid parasites. In the case of *T. brucei*, there is several transcriptome datasets each with a relatively small numbers of samples gathered from different experimental conditions.

To tackle the problem of RRE inference in *T. brucei*, we have developed a novel graph-based approach, termed GRAFFER, that identifies RREs by systematic integration of different transcriptome data sources. Application of GRAFFER to *T. brucei* transcriptome data led to the discovery of 88 RREs, of which eleven motifs resemble the previously known regulatory elements for the parasite. We also demonstrate that the novel elements not only agree with expected characteristics of functional RREs, but also are responsive to both transcriptomic and proteomic changes of the parasite during its life cycle.

Results and discussion

Prediction of functional gene regulatory elements

To infer RREs involved in the developmental and/or environmental responses of *T. brucei*, we considered three independent genome-wide transcriptome studies on *T. brucei* that included different life stages (Jensen et al. 2009), developmental processes triggered by the addition of cis-aconitate and lowering the temperature (Queiroz et al. 2009), and responses to a variety of chemical perturbations (Najafabadi et al. 2013). It is suggested that around 5-25% of trypanosome genes are responsive to the environmental changes. Hence, in our analysis, we focused on 25% most variable genes in terms of expression patterns as determined by the transcriptome data. As elaborated in the method section, given the three microarray datasets (Jensen et al. 2009; Queiroz et al. 2009; Najafabadi et al. 2013), we first modeled each dataset as a co-expression graph, where vertices represented genes and edges represented co-expression over the dataset. Next, an integrated co-expression graph was constructed by considering edges that were present in all three initial co-expression graphs. Switching from expression profiles to co-expression graphs proved to be an efficient way to identify sets of co-expressed genes across multiple datasets (Yan et al. 2007). Preliminary topological analysis of the integrated co-expression graph revealed that it exhibits both striking characteristics of most biological networks, small-world behavior and the scale-free property (Fig 3.1).

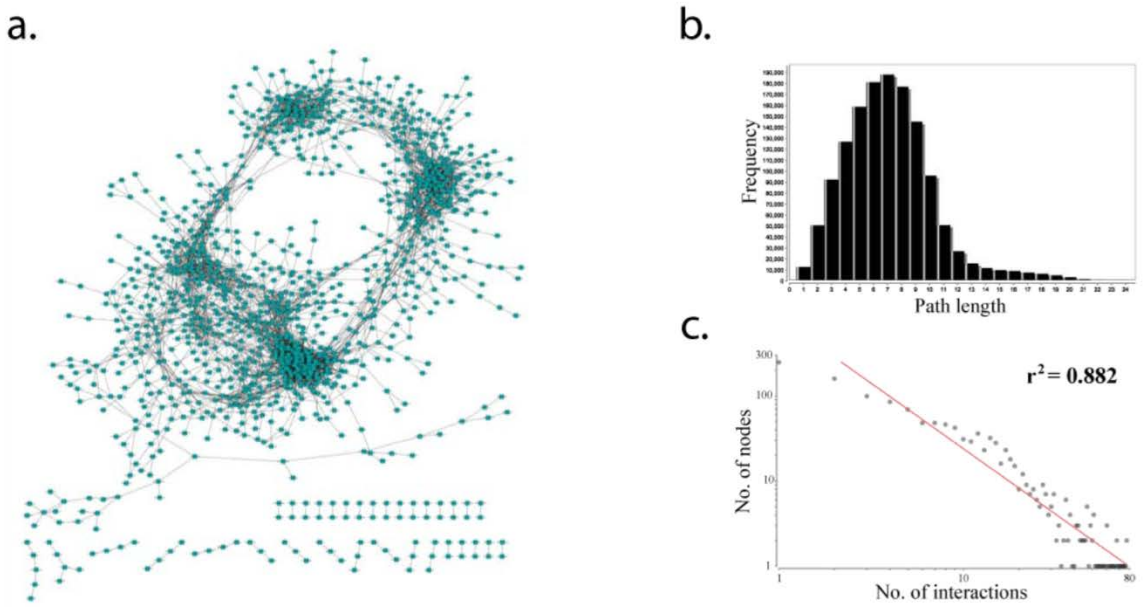


Fig 3.1. Constructed Co-expression graph based on three independent transcriptome datasets.

(a) Global view of the dichotomized co-expression graph for *T. brucei* genes, based on the integration of transcriptome data from three independent studies. The constructed graph is modular, i.e. there are highly connected regions in the graph that are separated from the other parts. The constructed co- expression graph has **(b)** scale-free and **(c)** small-world architecture.

Recent studies demonstrated that most RBPs recognize single stranded, linear RNA sequences and the structure around a binding site is mainly to support its single strandedness (Li et al. 2010; Ray et al. 2013). Studies on RREs in *T. brucei* have led to a similar idea; although there can be some structures associated with the functional regulatory sites (Walrad et al. 2009; Monk et al. 2013), these structures may not be conserved for the corresponding RREs (Monk et al. 2013). Therefore, in current work, we focused on linear sequence motifs for the identification of potentially functional RREs. Moreover, RREs tend to be enriched in the 3´-UTR region of trypanosomatid genes, although exceptions for some RREs (like prominent presence in the coding sequence) have been reported (Das et al. 2012). Here on, the terms “a gene harbors a motif” or “a gene targeted by a motif” were used, if the motif instance can be found in the 3´-UTR sequence of the gene (The employed approach for the selection of 3´-UTR regions are detailed in materials and methods). To discover linear RREs that target a set of coherently expressed genes, we developed a novel method, called GRAFFER, to search for linear motifs

whose targeted genes create a significantly dense module in the co-expression graph. To predict functional RREs, GRAFFER calculated the module density of more than $4 \times 10E6$ distinct linear motifs in the case of *T. brucei* integrated co-expression network. To assess the discriminative power of the defined score on the integrated co-expression graph of *T. brucei*, we compared the distribution of scores in this graph with a random graph, constructed by random permutation of gene labels in the integrated co-expression graph. As shown in Fig 3.2, the distribution of scores for motifs in the co-expression graph is right-skewed, while the distribution of scores for the same set of motifs in the random graph is randomly distributed. This figure clearly shows that the co-expression graph conveys information (based on the defined score) that is absent in the random graph.

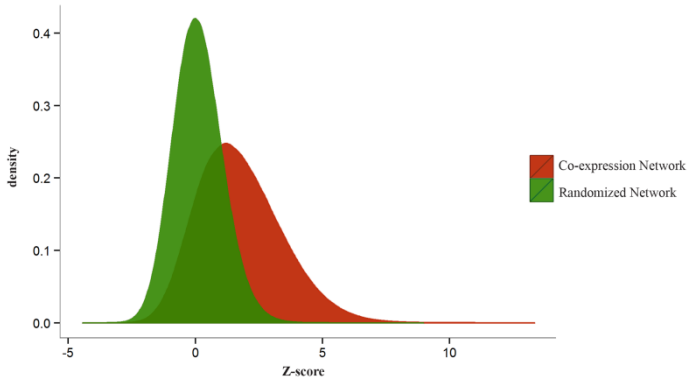


Fig 3.2. Distribution of motif modulation scores for the integrated co-expression graph and a random graph.
As illustrated, the distribution of motif scores for the integrated co-expression graph is right-skewed compared to that of the random graph.

Benchmarking GRAFFER on human

To systematically estimate the accuracy of GRAFFER predictions, we applied the approach on human, for which many RREs are experimentally identified, providing a rich context to systematically examine the accuracy rate of the approach. To search for RREs in human, we constructed a co-expression graph based on a compendium of 211 expression profiles across 38 distinct human hematopoietic cells, monitoring gene expression changes during the hematopoietic differentiation process (Novershtern et al. 2011). The interaction density of human co-expression graph was similar to the case of *T. brucei* integrated co-expression graph; and weights of edges were defined by Pearson correlation coefficient. 3′-UTRs of human genes was

defined the immediate 300nt down-stream of stop codon in the longest isoform of transcript, as described elsewhere (Elemento et al. 2007).

Application of GRAFFER led to the prediction of 49 significant non-redundant motifs whose targeted genes were significantly connected to each other in the co-expression graph of human, with Bonferroni corrected p-value less than 0.01. As expected for RREs, directionality analysis of GRAFFER motifs demonstrated that 47 motifs (~96%) show a strand bias and are significant only in the forward strand.

The predicted motifs target 49 densely connected modules in the co-expression graph. To assess the biological relevance of predicted modules, we first examined whether or not these modules were enriched for specific gene ontology (GO) biological process terms. This analysis revealed that 37 out of 49 predicted modules were enriched for at least one biological process, suggesting that although the modules were predicted solely based on characteristics of the 3′-UTR sequences and the co-expression graph, they have specific functions in the cell.

The recent large scale RNAcompete study has identified the binding preference of 205 distinct RBPs (Ray et al. 2013). This study also predicted a high confidence regulatory network for some of human's RBPs based on the integration of information on RREs and available transcriptome dataset (Ray et al. 2013). As illustrated in Fig 3.3 and detailed in RNAcompete section of supplementary text, comparison of predicted motifs with those of RNAcompete showed that 24 GRAFFER motifs are significantly similar to 62 RNAcompete experiments (some of the RNAcomplete experiments had replicates or identified the binding preference of several orthologous RBPs, leading to the matching of some GRAFFER motifs with multiple RNAcompete-derived motifs). Consistently, in cases that a GRAFFER motif matched with the binding site of an RBP with available predicted target RNAs, the predicted motif was significantly enriched in the 3′-UTR of the predicted targets as well.

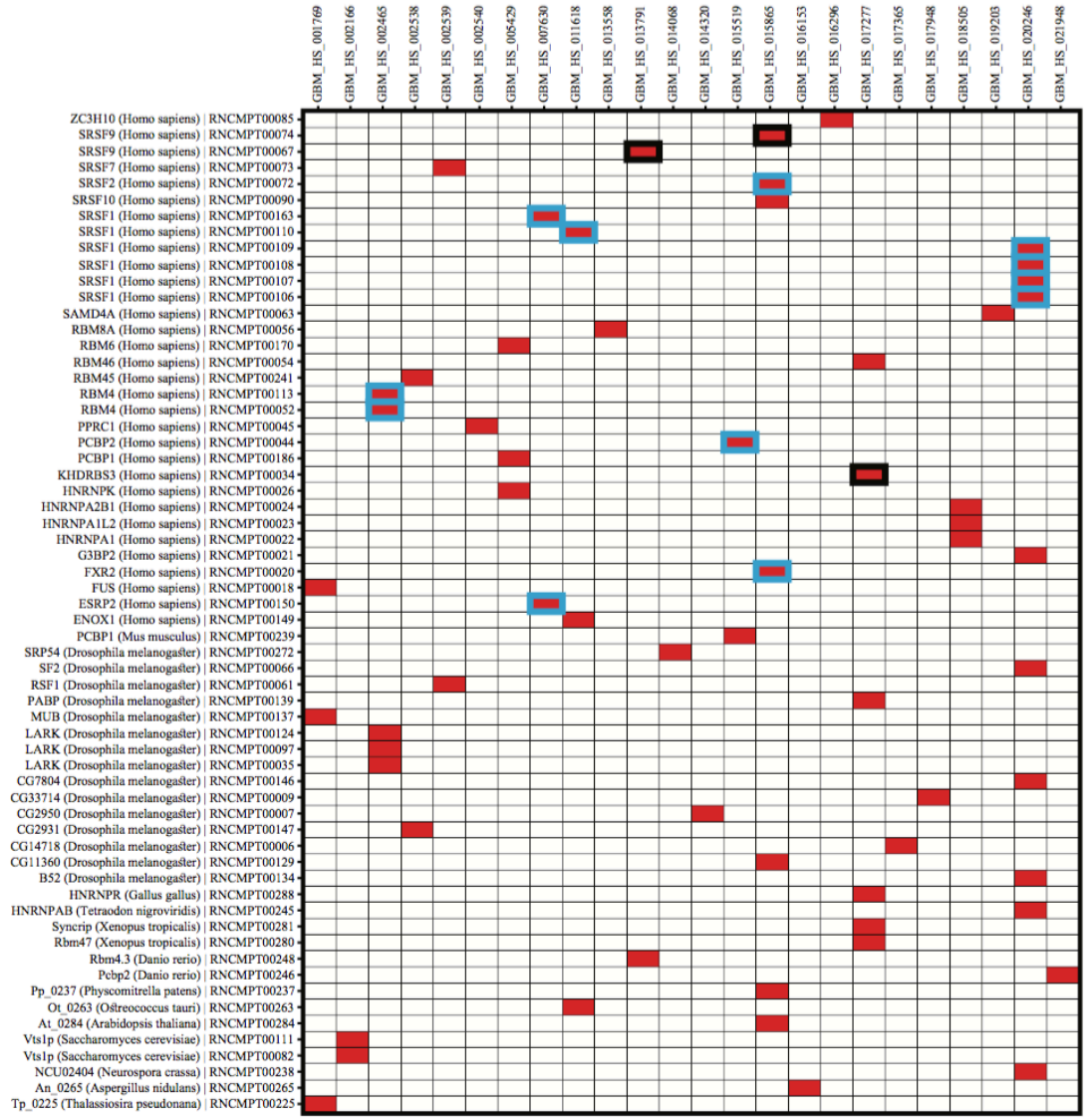


Fig 3.3. Comparison of predicted motifs for human with the identified RREs in a recent large-scale RNAcompete experiment.
24 out of 49 predicted motifs show significant similarity with identified RREs in 62 RNAcompete experiments (some of RNAcompete motifs are highly similar to each other because of the existence of experimental duplicates or conserved RNA binding domains). The bold blue frame indicates cases in which the GRAFFER motif was enriched (two tailed hypergeometric, p-value <0.01) among the RNA targets of the RBP as reported in (Ray et al. 2013); and the bold black frame indicates cases where GRAFFER motif was not enriched among the target RNAs.

To test whether the GRAFFER motifs can be related to miRNAs, we first examined if there is enrichment for the predicted targets of human miRNAs in the 49 found modules. This analysis showed that 42 modules (~86%) are enriched for the target RNAs of at least one human miRNA. Congruent with evidences about the complex interplay between RBPs and miRNAs (van

Kouwenhove et al. 2011; Ho and Marsden 2014), we found that many of modules that were predicted to be regulated by RBPs in the previous step can also be regulated with at least one miRNA. Moreover, we found that for 7 motifs, not only the cognate module is enriched for the target RNAs of a specific human miRNA, but also the motif match to the 5′- extremity of the miRNA (Fig 3.4; It should be noted that only human miRNAs were considered for matching with GRAFFER motifs). Interestingly, four of GRAFFER motifs that matched with human miRNA binding sites, showed significant similarity to the RBP binding sites as well which can be suggestive of potential competition for binding between RBPs and miRNAs.

The obtained results from this analysis demonstrated the power of our graph-based approach in identification of functional RREs based on co-expression graphs.

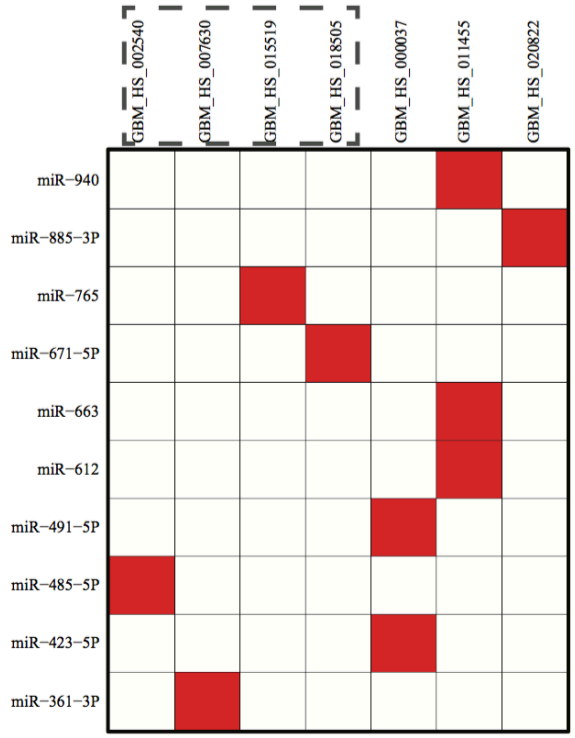


Fig 3.4. Comparison of GRAFFER motifs with the known human miRNAs
To examine whether or not the predicted motifs by GRAFFER can be the binding site of human miRNAs, we set two criteria: 1) The genes that harbor the motif should be enriched for the potential targets of a human miRNA. We used g:profiler web server for this analysis (Reimand et al. 2016); 2) The 5′-extermity of the miRNA should match to the reverse complement of the predicted motif sequence, allowing at most two nucleotide shifts in either miRNA or motif sequence. As illustrated, we found 7 motifs potentially represent the binding sites for 10 human miRNAs. Note that some of the predicted motifs not only can match with miRNAs, but also they can represent the binding site of RBPs (highlighted with the box).

Predicting RREs involved in the environmental responses of T. brucei using GRAFFER

Application of GRAFFER to *T. brucei* integrated co-expression graph led to the prediction of 88 non-redundant motifs whose targeted genes were significantly connected to each other in the graph (Bonferroni corrected p-value <0.01). However, applying GRAFFER with the same settings to 100 random networks, generated by random shuffling of gene labels in the co-expression graph, yielded 9.6 motifs on average (The employed randomization procedure changes the location of 3´-UTRs in the graph, while preserve the graph topological characteristics). The connection of a pair of genes in the integrated co-expression graph indicates their co-expression under various conditions; therefore, the significance of predicted motifs implies that the corresponding targeted genes by these motifs tend to be significantly co-expressed with each other under a wide range of conditions. As is expected from RREs, directional analysis of GRAFFER motifs showed that they mostly (more than 95%) have a strand bias and are significant only in the forward strand.

Characteristics of the predicted RREs

An experimentally deciphered regulatory network of 40 different RBPs in *Saccharomyces cerevisiae* revealed a complex combinatorial network among RBPs, such that different RBPs can target a similar set of RNAs (Hogan et al. 2008). The existence of such extensive regulatory networks is also suggested in *T. brucei* (Walrad et al. 2009; Clayton 2014). To explore the putative relationships among the predicted motifs, we examined the existence of significant patterns of co-occurrence for each pair of motifs (detailed in the method section). As shown in Fig 3.5, 29 pairs showed significant co-occurrence patterns with each other. This result supported the hypothesis that gene expression in *T. brucei* is regulated by a complex regulatory network. Lack of co-occurrence patterns for other motifs also indicated they target distinct sets of genes, suggesting diverse biological roles for them.

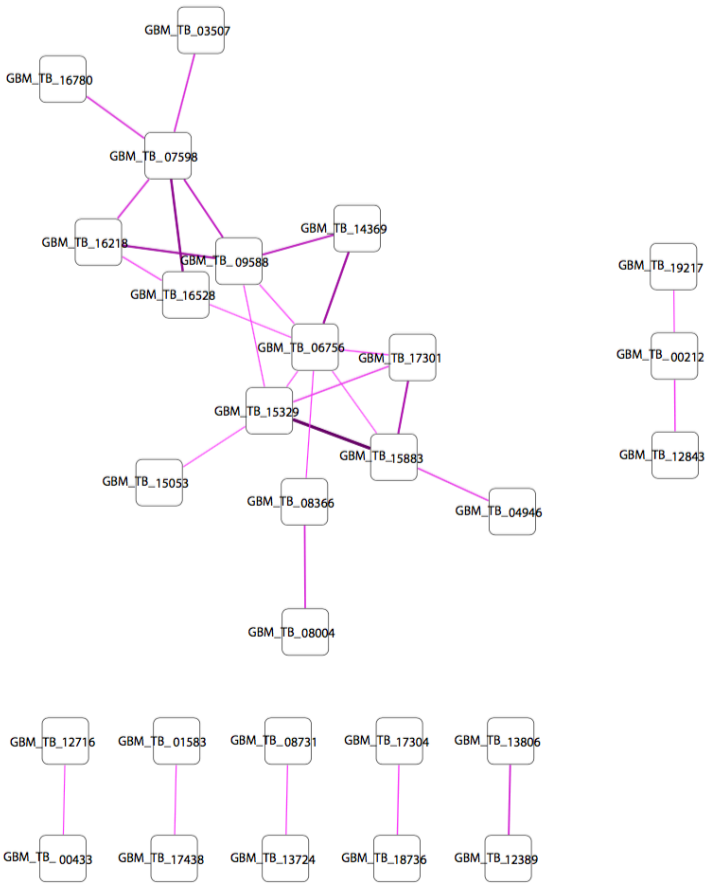


Fig 3.5. The motif co-occurrence network for 88 predicted motifs based on *T. brucei* co-expression graph
Motif co-occurrence profile represented as a network. Different RBPs can regulate the same set of transcripts. These combinatorial regulatory networks were captured by determining if the targeted genes by two different motifs significantly overlap with each other. The color density represents the calculated Z-scores for each interaction.

Additionally, we examined whether the predicted motifs showed specific expression patterns in the transcriptome data of each cell state. This analysis revealed that 84 out of 88 (95%) predicted motifs showed significant enrichment under at least one condition (Fig 3.6.a). We also considered the available proteomics data (Gunasekera et al. 2012; Urbaniak et al. 2012; Butter et al. 2013) to further demonstrate the functionality of predicted motifs at the proteome level. Following the same approach as the transcriptome data, we found that 19 motifs (22%) showed significant enrichment under at least one condition (Fig 3.6.b). Notably, the available proteomics data, compared with the transcriptome data, were from a limited set of conditions. Therefore, we

can expect this number to grow as more proteomic data become available. As discussed later, we show that enrichment results for several motifs are consistent with previous knowledge on the gene regulatory network of *T. brucei*.

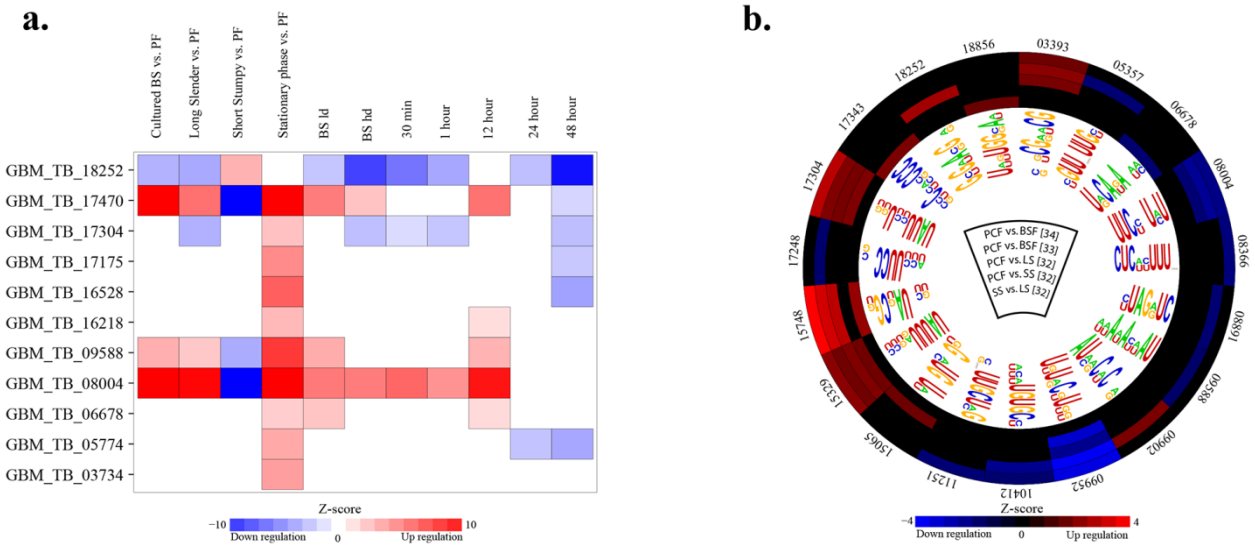


Fig 3.6. Motif enrichment analysis using available transcriptome and proteome data of *T. brucei*. For each GRAFFER motif, transcripts containing the motif in their 3′-UTR were selected and then tested for a statistically significant pattern in each cell state using standard Mann-Whitney rank sum statistic. (a) Developmental transcriptome responses of the eleven GRAFFER motifs that matched with the previously known RREs in *T. brucei*. For illustrative purposes, the full transcriptome enrichment analysis for the rest of the predicted motifs is not represented here. (b) Proteome enrichment analysis revealed that 19 motifs showed significant enrichment under at least one condition. The outer layer of circle indicates the motif id, excluding "GBM_TB" term due to illustration limitations. The intermediate circular layers indicate the up- or down-regulation of the motifs in a specific condition. The inner circular layer represents the consensus patterns of motifs. The center of circle shows the conditions that were tested for the enrichment with the reference for the study that data was extracted from. 1:(Gunasekera et al. 2012);2:(butter et al. 2013);3:(Urbaniak et al. 2012).

Comparison of predicted RREs with previously identified or predicted regulatory elements

To date, only a small fraction of RREs in *T. brucei* have been identified. Hence, we could not compare GRAFFER results with a large number of experimentally-derived regulatory elements. However, the fifth most significant motif (GBM_TB_17304) shows close similarity with one of the most intensely studied U-rich RREs in *T. brucei* (Hotz et al. 1998; Mayho et al. 2006). The predicted motif not only strikingly resembles the experimentally derived RRE, but also has highly overlapping RNA targets with it (Fig 3.7.a). It is suggested that the experimentally

determined RRE targets the 3´-UTR of many diverse sets of transcripts on a genome-wide scale (Mayho et al. 2006). From the functional perspective, it is experimentally verified that this RRE is involved in the developmental regulation of transcripts, with a destabilizing effect on target RNAs in the bloodstream form. Intriguingly, the GBM_TB_17304 motif also showed a similar effect on the targeted RNAs at both transcriptome and proteome level. More importantly, GBM_TB_17304 matches with the previously known functional instances of the experimentally derived regulatory elements (Fig 3.7.b). These lines of evidence indicate that potential follow up experimental works on GBM_TB_17304 can lead to the same biological knowledge as of the experimentally derived RRE.

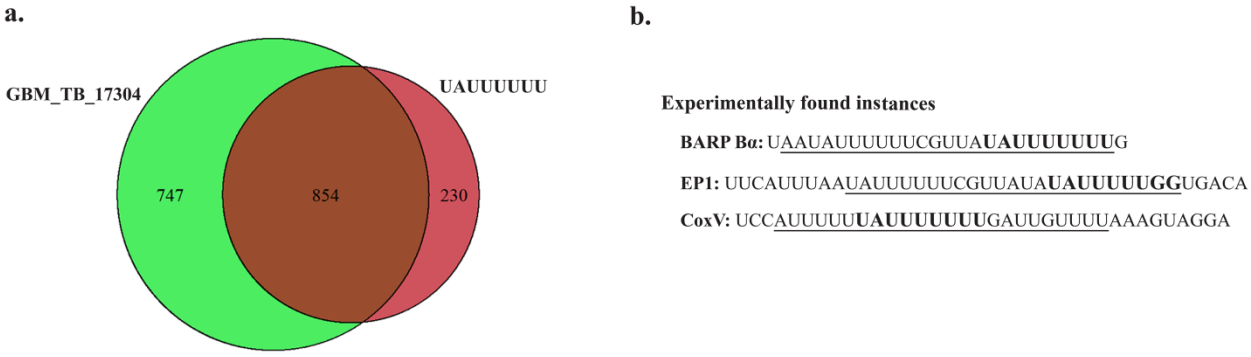


Fig 3.7. Developmentally regulated U-rich RRE in *T. brucei*.
Comparison of an experimentally established RRE (UAUUUUUU) that is involved in developmental regulation of *T. brucei* genes, with GRAFFER motif, GBM_TB_17304. (a) Venn-diagram of the transcripts that are targeted by UAUUUUUU and GBM_TB_17304 motifs. (b) Underlined regions show the U-rich regions with the experimentally-verified regulatory role that were used to infer UAUUUUUU regulatory element. The bold sequence in each U-rich region represents the part of region that matches with GBM_TB_17304 motif.

Another predicted motif (GBM_TB_16528) resembles a previously identified A/U-rich element that is involved in the heat-shock response of the parasite (Droll et al. 2013). It is experimentally verified that ZC3H11 zinc finger protein can recognize and bind to the A/U-rich element with stabilizing effect on the target RNAs during heat stress. Congruent with the experimental evidence, genes targeted by GBM_TB_16528 are gradually down-regulated during the differentiation of bloodstream cells to procyclic cells where along with the change of cell media, the temperature is reduced from 37°C to 27°C with most reduction in expression level observed after 48hrs of differentiation process (p-value <4E-05; Mann-Whitney rank sum test). Moreover, Genes targeted by GBM_TB_16528 also show moderately reduced expression patterns in

cultured bloodstream forms compared to the cultured procyclic cells (p-value <0.085; Mann-Whitney rank sum test). The bound transcripts to the ZC3H11 have been also reported using cross linking experiments coupled with deep sequencing (Droll et al. 2013). Importantly, GBM_TB_16528 is significantly enriched among the strongly bound transcripts to the ZC3H11 (p-value < 2E-13; two tailed hypergeometric test) and matches with the known instances of experimentally-determined A/U-rich element, as reported in (Droll et al. 2013).
Benefiting from an *in vitro*, Selex-based technique known as RNAcompete (Ray et al. 2009), a large-scale study has revealed the binding preference of 13 trypanosomatid RBPs (Ray et al. 2013). As detailed in materials and methods, comparison of the motifs revealed that 11 out of the 13 trypanosomatid Selex-based motifs showed significant similarity to nine out of 88 GRAFFER motifs. Matching the RNAcompete results with the GRAFFER predictions also led to new insights on the functional roles of the RBPs. As an illustration, the comparison suggested the recognition of GBM_TB_09588 motif by DRBD13 protein (Tb927.8.6650). A recent study has experimentally demonstrated that DRBD13 protein is essential for the procyclic life stage of the parasite and its tethering to RNA leads to the down regulation of the target in this life stage (Jha et al. 2015). Consistent with this, we found transcripts harboring GBM_TB_09588 are significantly down regulated in the procyclic stage compared to both long slender stage (p-value < 0.002; Mann-Whitney rank sum test) and cultured booldstream cells (p-value <0.0002; Mann-Whitney rank sum test). Additionally, re-analysis of available DRBD13 tandem affinity purification coupled with deep sequencing (RIP-seq) data (Najafabadi et al. 2013) indicated that GBM_TB_09588 is significantly enriched among the bound transcripts to the RBP (p-value <4E-32; Mann-Whitney rank sum test). Moreover, congruent with suggested role of DRBD13 protein on regulation of membrane associated proteins (Jha et al. 2015), genes targeted by GBM_TB_09588 are significantly enriched for genes involved in antigenic variation (p-value <8.30E-16; Fisher exact test). As another example, the comparison of RNAcompete results with GRAFFER predictions demonstrated the possible recognition of GBM_TB_16218 motif by DRBD12 protein (Tb927.7.5380). A previously published study have experimentally demonstrated the destabilization role of DRBD12 protein on its target (Erben et al. 2014b). Consistently, re-analsis of available microarray data on knock down of DRBD12 indicated that genes targeted by GBM_TB_16218 are significantly up-regulated in its knock down background (p-value < 0.0007; Mann-Whitney rank sum test). Moreover, re-analysis of available RNA-seq

data on the insect-stage life-cycle of *T. brucei* demonstrated about three fold up-regulation of DRBD12 in the proventriculus life stage (Kolev et al. 2012). Consistent with destabilization role of DRBD12 protein, transcripts containing GBM_TB_16218 motif are significantly down regulated in this life stage (p-value <0.02; Mann-Whitney rank sum test). This data suggests the possible role of DRBD12 RBP in the insect stage differentiation process of the *T. brucei*.
It worth noting that all four motifs discussed above have more than fifty percentages of A and/or U in their consensus sequence. However, as discussed above, they show diverse responses during the life cycle of the parasite. GBM_TB_17304 motif is significantly upregulated in the procyclic cells. In contrast, the two motifs of GBM_TB_09588 and GBM_TB_16528 are upregulated in the bloodstream life stage, while their mechanisms of actions are different than each other which is also supported by motif co-occurrence profiles. In fact, 64% (56 out of 88) of predicted motifs by GRAFFER have more than 50% A and/or T in their composition. However, these motifs mostly target different transcripts (as judged by motif co-occurrence profiles) and show different responses during the life cycle of *T. brucei*, suggesting a potentially distinct and diverse role for A- and/or U-rich motifs in the gene regulatory network of the parasite. Based on the available experimentally verified RREs, we compared the performance of GRAFFER with three other genome-wide computational studies of RREs on *T. brucei* (Mao et al. 2009; Shateri Najafabadi and Salavati 2010; Najafabadi et al. 2013). The comparison showed that our new approach outperformed all of them in terms of accuracy. Briefly, no experimental motifs with a developmental role were predicted by any of them. In addition, the comparison of RNAcompete-derived RREs with the predicted motifs from each article revealed that the RNAcompete RREs had better agreement with our new approach compared with the others (see materials and methods for the details). It is important to note that RREs are not extensively characterized in *T. brucei*; which led us to compare the predictions of each study with a limited set of previously known RREs. Therefore, some of the novel RREs predicted by these approaches may be valid, but not discovered yet.

Assessment of the dataset integration performance

To further evaluate the performance of our approach, we investigated the significance of predicted motifs in cases where we constructed a co-expression graph by considering only a subset of datasets (one or two datasets) instead of all three. It should be pointed out that

application of GRAFFER in these cases may lead to the prediction of some new motifs; however, because the number of samples in each dataset is relatively small, it is likely that regulatory circuits are not well separated from each other, leading to the connection of genes involved in parallel regulatory circuits in these graphs. Therefore, we only focused on the set of 88 motifs, ignoring the newly found instances in each case as they are not most likely reliable. As shown in Fig 3.8, each dataset on its own was not informative enough and few motifs remained significant in these networks (1% developmental, 16% life stages, and 20% chemical perturbations). However, most motifs became significant when the co-expression graph was constructed by integrating at least two datasets. This result was anticipated, because each dataset on its own had captured only a small set of cell states, but when the datasets were integrated with each other, the responses of the cell system became more evident, leading to the prediction of more significant motifs. Interestingly, we found that the integration of two contextually dependent datasets (developmental and life stages datasets) did not improve the performance of approach noticeably; however, the integration of two contextually independent datasets (life stages and chemical perturbations datasets or, alternatively, developmental and chemical perturbations datasets) boosted the inference power of the approach significantly. Previous studies suggested that although there is a significant gene expression remodeling in different life stages and during differentiation process, gene expression variations within each life stage are limited. This assumption has meant that almost all genome-wide studies (transcriptome and proteome) have focused on the developmental aspects of gene expression. In our analysis, we clearly observed that the integration of development-related datasets with a limited dataset (11 samples) from chemical perturbations increased the precision of co-expression graph dramatically. This analysis also provided insights into the functional regulatory roles of some of the predicted RREs. For example, developmentally regulated RRE (GBM_TB_17304) is significant in both the developmental and life stages datasets; however, it loses its significance in the chemical perturbations dataset.

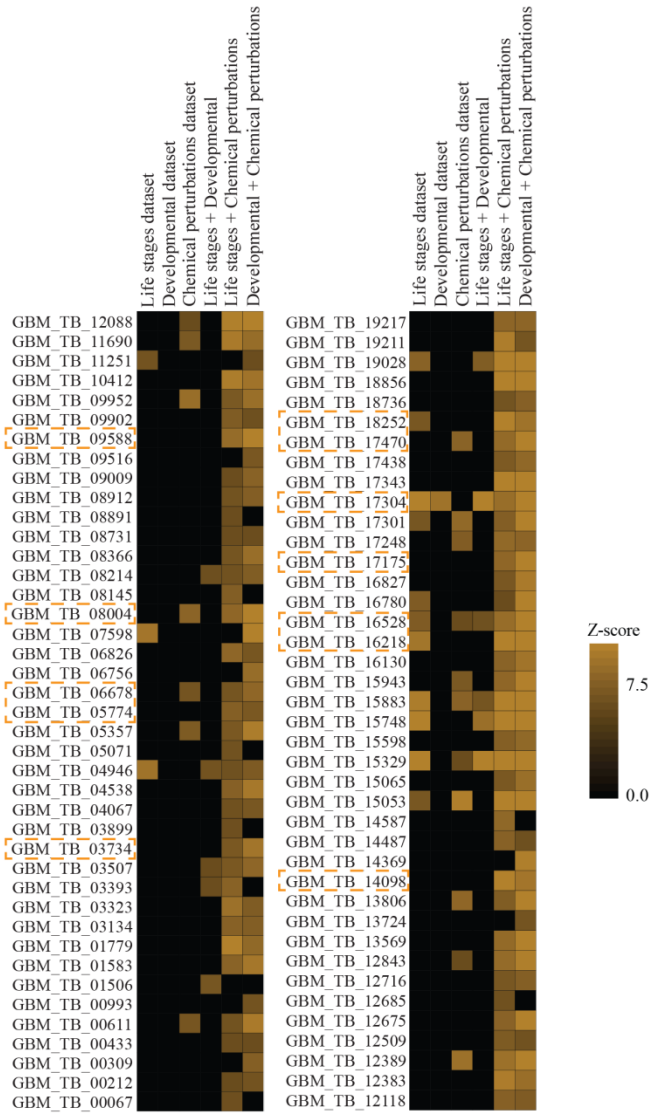


Fig 3.8. Performance of GRAFFER using only a subset of datasets.
The columns represent six different co-expression graphs that were constructed by considering one or two transcriptome datasets. The figure is pseudo-colored, with only conditions (i.e., co-expression graphs) that motif has significant Z-scores are only shown. The orange boxes around some of the motif names highlight the motifs that matched to the experimentally validated RREs.

Application of GRAFFER to Cell cycle transcriptome

Independent application of GRAFFER on each of the three datasets (Jensen et al. 2009; Queiroz et al. 2009; Najafabadi et al. 2013) indicated its power to find functional RREs from datasets with a relatively small number of samples, but limited relative to the case that datasets were integrated with each other. To further test this hypothesis, we considered the available cell cycle gene expression data in *T. brucei* (Archer et al. 2011), comprised of four cell states (Early G1, Late G1, S phase, and G2/M phase).

For the *T. brucei* cell cycle co-expression graph, we considered genes that showed at least a 1.5 folds change in one cell cycle stage compared with early G1 phase. Performing the same steps as

our previous attempt, we applied GRAFFER on the constructed co-expression graph from this dataset. In this case, our approach identified five significant motifs (Fig 3.9.a). The low number of significant motifs was anticipated because of the low number of samples in the dataset. Comparison of the predicted motifs with experimentally established motifs revealed that one of our motifs matched a well-studied RRE in trypanosomatids. This experimentally validated RRE is involved in cell cycle regulation in trypanosomatid organisms (Bhandari et al. 2011). Importantly, genes harboring each of these experimental and computational motifs were significantly upregulated in the late G1 cell cycle phase (Fig 3.9.b).

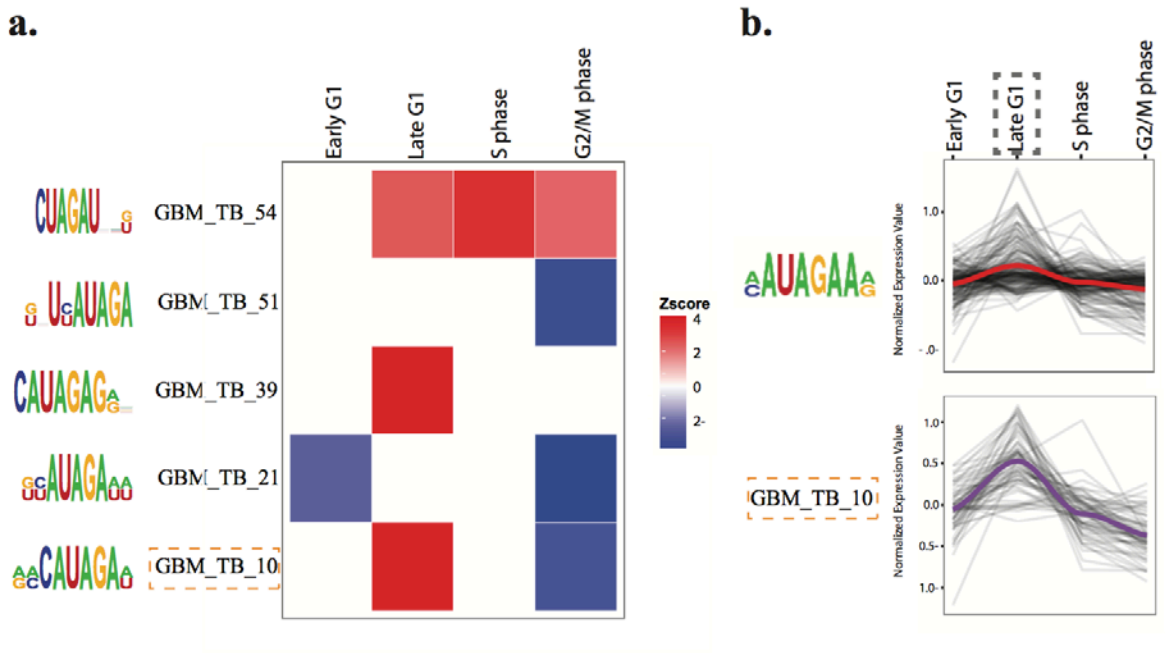


Fig 3.9. Transcriptome responses of GRAFFER motifs that were predicted based on the cell cycle transcriptome data of *T. brucei*.
(a) Predicted motifs are responsive to the transcriptome changes during cell cycle progression of *T. brucei*. **(b)** Comparison of an experimentally validated RRE, with a role in the cell cycle regulation, with GBM_TB_10. Both motifs are significantly upregulated in late G1 phase (Mann-Whitney rank sum statistic, p-value <0.05). The experimentally established RRE was extracted from (Bhandari et al. 2011).

Concluding remarks

In this chapter, we have introduced a graph-based solution to predict RREs by systematic integration of different transcriptome data sources. This property becomes particularly important in the study of non-model organisms with limited whole genome expression datasets. Application of our approach has led to the prediction of 88 RREs that function in the gene regulatory network of *T. brucei*. To date only a small fraction of RREs in *T. brucei* have been identified, yet eleven predicted motifs strikingly resemble experimentally-derived trypanosomatid regulatory elements. Further comparison of these eleven motifs with experimentally-derived RREs indicated that they not only target highly overlapping transcripts, but also show similar transcriptome and proteome responses to the environmental and developmental changes of *T. brucei*.

Application of GRAFFER on random graphs suggested false discovery rate of less than eleven percent for the predictions, suggesting a high accuracy rate for the predictions. Additionally, application of GRAFFER to human demonstrated that 55% of predictions match to previously known RREs. Moreover, our results indicated that 95% of the predicted motifs for *T. brucei* are responsive to the transcriptome and proteome changes in the life cycle of the parasite. In several cases, we have shown that these predictions match with previous knowledge on the gene regulatory network of *T. brucei*. Our results also led to the prediction of biological roles for several uncharacterized RREs and RBPs.

Consistent with experimental evidences [31], the motif co-occurrence patterns suggested intricate and intertwined regulatory relationship between some of the regulatory elements and, consequently, their cognate RBPs. However, these patterns also revealed that many motifs target distinct RNAs, suggesting regulation of a wide range of different trypanosomatid RNAs by RBPs. Moreover, the sequence characteristics of the predicted motifs highlight the importance of A- and/or U-rich elements in the gene regulatory network of *T. brucei*. Importantly, these motifs show distinct transcriptome and proteome responses to the life cycle changes of the parasite, suggesting their diverse regulatory roles.

Although GRAFFER is designed to allow inference of RNA regulatory elements based on limited transcriptome data sources by their systematic integration, it still relies on the concept of co-expression. The approach first models each dataset as a graph where edges represent co-expression over the dataset. The initially constructed co-expression graphs are then systematically integrated to gain a higher resolution picture of underlying regulatory circuits in the cells. Our analysis clearly demonstrated that while inference of regulatory elements based on a single dataset provided limited information, the integration step boosted the inference power significantly. However, due to reliance on the co-expression concept for the construction of the initial co-expression graphs, datasets with extremely limited number of sample (i.e. those monitoring expression changes in only two or three different conditions) cannot be used in our approach. Therefore, to infer RREs that are responsive to the developmental and/or environmental changes of the parasite, we have only considered transcriptome datasets that capture gene expressions in at least five biologically different conditions.

Experimental knowledge on mechanisms of actions of RBPs demonstrates that the recognized elements by these proteins are either single stranded or have particular secondary structures. Therefore, the secondary structure of RNAs play important role in the recognition of RREs by the RBPs. However, the current implementation of GRAFFER focuses only on the 3'-UTR sequences, ignoring the RNA secondary structures. Therefore, the 88 predicted are biased towards the linear motifs with no knowledge on the structural context of the motif instances present in the 3'-UTRs. Although recent studies demonstrated that most RBPs recognize single stranded RNA sequences and the structure around a binding site is mainly to support the single strandedness (Li et al. 2010; Ray et al. 2013), the employed motif searching procedure leads to systematic loss of structural dependent motifs, some of which are shown to play important roles in the gene regulatory of the parasite (Walrad et al. 2009; Monk et al. 2013; Fernandez-Moya et al. 2014).

Materials and methods

Construction of the integrated co-expression graph for T. brucei

We focused on three independent transcriptome studies (Jensen et al. 2009; Queiroz et al. 2009; Najafabadi et al. 2013) to construct an integrated co-expression graph of *T. brucei*. To select for 25% genes with most variable expression patterns, we observed the variation of each gene (i.e., its standard deviation) in each dataset independently and the top 30%, 32% and 37% variable genes from (Jensen et al. 2009), (Queiroz et al. 2009) and (Najafabadi et al. 2013) were selected, respectively. The common protein coding genes among all three, consisting of about 25% of *T. brucei* genes (~1900 gene) were chosen for further analysis.

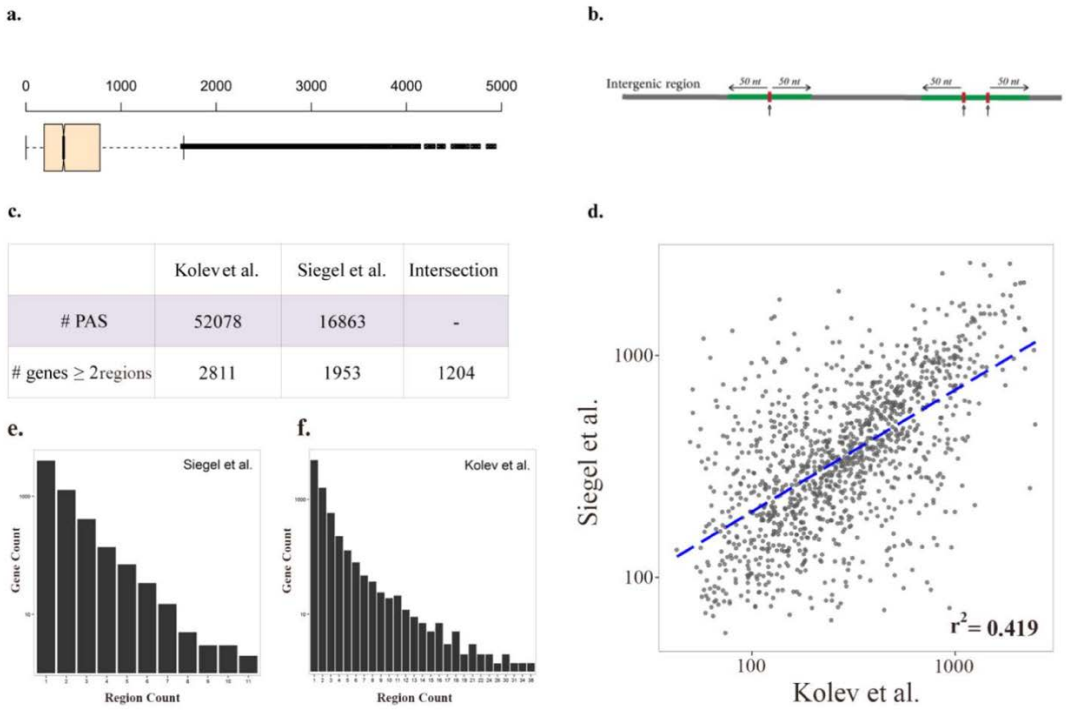
Initially, each of three microarray datasets were modeled as a weighted co-expression graph such that vertices represented genes, while their edges denoted the values of the pairwise Pearson correlation coefficient (PCC). The advantage of weighted co-expression graphs over un-weighted graphs is that the former preserve the underlying connectivity information. However, because the number of samples/conditions in each dataset is relatively small, weak correlations may not have biological relevance. To emphasize on strong correlations, we only considered those interactions which their squared values of the correlation coefficient (r^2) were equal or greater than 0.5. Moreover, negatively correlated pairs were excluded from each co-expression network as they do not support co-regulation. Next, an integrated co-expression graph was constructed by considering edges that are common in all three initial co-expression graphs. The edge weights of the integrated graph were defined as the average of weights for the corresponding edges in the three initial co-expression graphs.

Recently duplicated genes tend to have similar coding and 3´-UTR sequences. Thus, we would expect similar expression patterns for these genes in the microarray experiments because of cross-hybridization effects. Moreover, highly similar 3´-UTR sequences can cause bias in our motif scoring approach. To obviate these issues, from each two homologous genes that were present in the integrated co-expression graph, we randomly kept one and deleted the other one. Homologous genes in *T. brucei* genome were extracted from the MCL database v5.

T. brucei 3′-UTR sequences

The 3′-UTR sequences were downloaded from TriTrypDB v.5, considering lengths reported in (Siegel et al. 2010). In cases of alternative poly-adenylation, the median length was selected. In cases that gene did not have an identified 3′-UTR length, 400nt (the median 3′-UTR length of *T. brucei* genes) downstream of the translational stop codon was selected. Preliminary analysis of 3′-UTR lengths revealed that although the median length is 400nt, some transcripts can have very long 3′-UTRs (Fig 3.10.a). Recent discoveries suggested that alternative poly-adenylation site selection can have regulatory impact on the expression level of transcripts in different organisms (Elkon et al. 2012). For transcripts with alternative 3′-UTRs, the longer isoforms potentially have more binding sites for RNA-binding proteins and/or miRNAs. In general, the outcome of having more regulatory regions is that isoforms with shorter 3′-UTRs have elevated expression levels compare with the longer isoforms of the same transcript (Mayr and Bartel 2009). In support to the regulatory role of alternative poly-adenylation site selection, the 3′-UTR length of at least one transcript in *T. brucei* is reported to be developmentally regulated (Jager et al. 2007). Moreover, alternative trans-splicing (which can lead to variation in 3′-UTR lengths) plays a role in the developmental regulation of some *T. brucei* genes (Nilsson et al. 2010). Previous studies on *T. brucei* suggested that poly-adenylation site selection in this organism is linked to the selection of the downstream 3′-splice-acceptor site (Matthews et al. 1994). Considering both dependency on splice-acceptor-site selection and the error in sequencing that may occur because of the low complexity of 3′-UTR regions, the existence of minor variations on detected poly-adenylation sites was anticipated. To test the possibility that gene expression is regulated by alternative poly-adenylation site selection, we first examined the agreement between two published studies on poly-adenylation sites of *T. brucei* transcripts (Kolev et al. 2010; Siegel et al. 2010). Considering each study independently, we defined poly-adenylation regions by considering ±50nt around each detected poly-adenylation site. If two adjacent poly-adenylation sites had overlapping regions, relevant regions were merged and the new region was defined as the union of both. Thus, two poly-adenylation sites in different regions would be at least 100nt far from each other, shown schematically in Fig 3.10.b By applying this selection criterion, we tolerated false negative results to reduce false positives. This analysis revealed that for many genes in *T. brucei*, there are at least two poly-adenylation regions supported by two independent studies (Fig 3.10.c). Next, we examined the agreement of 3′-UTR length variation

for transcripts with at least two poly-adenylation regions in both studies. Considering standard deviation of 3′-UTR length variation obtained from each article, we observed a moderate but significant agreement for 3′-UTR length variation between the two studies (Fig 3.10.d). This result demonstrated that 3′-UTR length variation is replicable and two independent experiments with different coverage levels produced similar results. Intriguingly, we found that although transcripts with very long 3′-UTRs (length > 1000nt) are usually downregulated under most biological conditions (as expected); these genes are significantly upregulated in some specific stress conditions (Fig 3.11). Coherent upregulation of these genes under some stress conditions could occur by disruptions in 3′-UTR length regulation mechanisms under these stress conditions or by up- or downregulation of some specific RBPs that mediate 3′-UTR length variation in response to the stress. Considering 3′- UTR lengths according to (Siegel et al. 2010), statistical analysis of transcripts with long 3′-UTRs (length > 1000nt) showed that these transcripts have a significantly tendency to have more than one poly-adenylation region (Mann-Whitney rank sum, $p - value < 10E - 114$). Unfortunately, most poly-adenylation sites in *T. brucei* were detected in only one cell state (Procyclic form, log-phase). This restricted us to examining whether different isoforms of some transcripts are preferred in different cell states, but these data suggested that there may be other regulatory mechanisms in parallel to RREs, which regulate the expression levels of *T. brucei* genes, particularly for genes with long 3′-UTRs. Coherent up- or downregulation of these transcripts implies that they have predictable expression patterns, independent of their long 3′-UTR sequences. Besides, this coherency in expression patterns resulted in their significant connections to each other in the constructed co-expression network ($p - value < 10E - 34$). The significant connections of these transcripts to each other along with their long 3′-UTRs could compensate for the random distribution of some non-functional motifs, leading to a bias in our motif prediction approach. To take these issues into account, we restricted the maximum 3′-UTR length for each transcript to 1000nt (i.e., the first 1000nt of 3′-UTR regions were considered for motif prediction). We found that replacing considered 3′-UTR lengths with the defined lengths by Siegel et al. (Siegel et al. 2010) has no effect on the significance state of 88 predicted motifs, with only one exception. It is likely that by considering the whole 3′-UTR lengths instead of the truncated version, the approach will predict more motifs that may not be biologically relevant.



	Kolev et al.	Siegel et al.	Intersection
# PAS	52078	16863	-
# genes ≥ 2regions	2811	1953	1204

Fig 2.10. Characteristics of *T. brucei* 3′-UTRs
(a) 3′-UTR length variation of *T. brucei* genes according to Siegel et al. In cases where a gene has alternative poly-adenylation sites, the 3′-UTR length is defined as the median length; **(b)** schematic representation of the defined poly-adenylation sites. Upward arrows represent the location of detected poly-adenylation sites for a gene. Each region is defined as 50nt before and after the detected poly-adenylation site. If the distance between two poly-adenylation sites was less than 100nt, the two corresponding regions were merged together. **(c)** Number of poly-adenylation sites and regions determined in two independent studies. As shown, many genes have more than one determined poly-adenylation region. **(d)** Correlation of two studies for 3′-UTR length variation of genes with more than one poly-adenylation region. The Y-axis and X-axis indicate the standard division of 3′-UTR lengths according to (Siegel et al. 2010) and (Kolev et al. 2010), respectively. **(e)** Distribution of genes based on the number of poly-adenylation regions, according to (Siegel et al. 2010)**(f)** Distribution of genes based on the number of poly-adenylation regions, according to (Kolev et al. 2010).

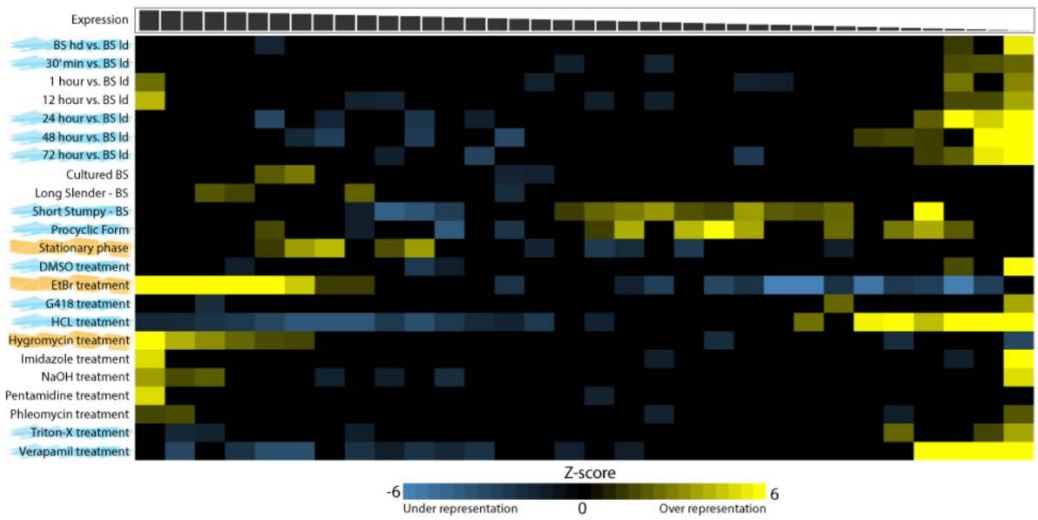


Fig 2.11. Patterns of up- and downregulation of genes with long 3′-UTRs under different experimental conditions.
For each condition, genes wee sorted according to their expression value. Sorted genes were then divided into 30 different bins. The enrichment of genes with long 3′-UTRs in each bin was examined using Fisher's exact test. Yellow color shows over-representation of genes in the corresponding bin. Similarly, blue represents under-representation of these genes in the cognate bin. The figure is pseudo-colored, only statistically significant bins are colored (Bonferroni corrected p-value < 0.05). Highlighted conditions on the left show overall significant up- or downregulation using Mann-Whitney rank sum statistics. Blue backgrounds indicate downregulation and orange backgrounds represent upregulation of genes with long 3′-UTRs.

Graph-based approach for finding functional elements in RNA (GRAFFER)

In current implementation of the GRAFFER algorithm, we considered only linear motifs generated over an alphabet of 11 characters (A, C, G, U, S=[CG], W=[AU], Y=[CU], R=[AG], M=[AC], K=[GU], N=[ACGU]). The terms "a gene harbors a motif" or "a gene targeted by a motif" were used, if the motif instance can be found in the 3′-UTR sequence of the gene. Accordingly, the module targeted by a motif is defined as the set of genes in the co-expression graph which are targeted by the motif.

Inspired by the cohesiveness concept in graph structure analysis (Nepusz et al. 2012; Khosravi et al. 2014), GRAFFER quantifies the extent of connections between genes harboring the same motif by defining a motif modulation score. The modulation score of a motif is defined as:

$$m = \frac{\sum_{intra-interactions} interaction\ weight}{\sum_{intra-interactions} interaction\ weight + \sum_{inter-interactions} interaction\ weight}$$

Where m represents the modulation score of the motif, intra-interactions are defined as the interactions of genes targeted by the same motif in the co-expression graph, and inter-interactions are defined as interactions of targeted genes with other genes (not targeted) in the graph.

To assess the statistical significance of observed motif modulation score for a motif, the corresponding Z-score was defined as:

$$Z - score = \frac{m - m_0}{sd}$$

Where m denotes the observed modulation score for the motif, and m_0 and sd represent the expected modulation score and standard deviation for the motif with that redundancy, respectively (motif redundancy in a graph is defined as the number of the genes in the graph that are targeted by the motif). The expected modulation score and standard deviation for a motif of particular redundancy were estimated by observing the distribution of modulation scores for 1000 randomly selected modules of the same redundancy as the motif in the graph. For a given motif, GRAFFER estimates the Z-scores by assuming a normal distribution for the modulation scores. The distributions of the modulation scores are dependent on the graph structure; however, our preliminary results based on Kolmogorov-Smirnov goodness-of-fit test showed that only extreme cases, i.e. motifs with very high or low redundancy in the graph, violate normal approximation. Therefore, a minimum and maximum acceptable number of occurrences for each motif are considered for the analysis. Each acceptable motif should target at least 20 and at max $(n - 20)$ genes in a co-expression graph with n nodes (e.g., genes). The lower limit will not cause a problem in our motif searching procedure because our goal is finding genome-wide conserved RREs.

A large scale experiment on a very diverse set of RBPs has demonstrated that these proteins tend to recognize and bind to short motifs with optimum predictive power at length of seven (Ray et al. 2013). The short length of binding site is also supported with crystallography data (Hudson et al. 2004; Clery et al. 2008b). The same binding characteristic is also supported for trypanosomatid RBPs (Gazestani et al. 2014). To search for motifs that target significantly dense modules in the co-expression network, GRAFFER starts by considering all possible 7-mer consensus patterns generated over an alphabet of 11 characters (as described above), with at least

4 non-degenerate bases and acceptable redundancy in the graph. The modulation score and the corresponding Z-score for each acceptable motif are then calculated. GRAFFER, next, selects motifs with significant modulation scores (Bonferroni corrected p-value <0.01) for the optimization process. The optimization process allows expansion in the motif consensus lengths. GRAFFER optimized each significant motif m by considering all possible, up to 9-mer, consensus patterns (constructed over the same alphabet) with the conserved consensus of m and chose the most significant one (the one with the highest Z-score) as the optimized motif. Finally, multiple optimized motifs can represent various derivative forms of a single RRE originated from different primary 7-mers. To avoid redundancy in the predicted motifs list, GRAFFER sorts optimized motifs based on their observed Z-scores. Starting from the most significant motif, it creates adjusted 3′-UTRs by masking all instances of the motif and then recalculates the Z-scores for the remaining motifs. Next, motifs are again sorted based on recalculated Z-scores and motifs that have lost their significant state after the masking procedure are discarded. This procedure is repeated for the next most significant motif. GRAFFER ends this cycle when no more significant motifs remained. As final report, GRAFFER reports back the motifs that remained significant in the above mentioned procedure. The employed procedure guarantees that each motif targets a significantly dense module in the co-expression graph and the significance state of each motif is independent of presence of the others.

Motif co-occurrence profile

To identify combinatorial interactions among predicted motifs, we compared the probability of co-occurrence of two motifs to the expected probability of co-occurrence by chance. To estimate the expected probability of co-occurrences for two motifs $\mathbf{m_1}$ and $\mathbf{m_2}$, random motif pools for each $\mathbf{m_1}$ and $\mathbf{m_2}$ were considered, each composed of 200 random motifs with the same length and redundancy (i.e., the number of the genes in the graph that are targeted by the motif) as $\mathbf{m_1}$ and $\mathbf{m_2}$, respectively. We observed the expected probability of co-occurrence for $\mathbf{m_1}$ and $\mathbf{m_2}$ by examining each possible combination of corresponding random motifs present in their pools. By assuming that the null model follows a binomial distribution, we reported the Z-score for the pair of $\mathbf{m_1}$ and $\mathbf{m_2}$ as

$$(\boldsymbol{K - NP_0})/\sqrt{\boldsymbol{NP_0(1 - P_0)}}$$

where **K** denotes the common targets between $\mathbf{m_1}$ and $\mathbf{m_2}$; **N** represents total number of unique targets for $\mathbf{m_1}$ and $\mathbf{m_2}$; and $\mathbf{P_0}$ represents the expected probability of co-occurrence for motifs $\mathbf{m_1}$ and $\mathbf{m_2}$.

Motif gene set enrichment analysis

To identify enrichment of a motif in a specific cell state in the proteome or transcriptome datasets, genes were ranked according to their normalized expression values. We then examined if the genes targeted by the motif showed statistically significant over-representation toward the top or bottom of the ranked list, using the standard Mann-Whitney rank sum statistic. The Benjamini-Hochberg false discovery rate of 0.05 was selected as the cut-off threshold.

RNAcompete

RNAcompete is a single-cycle competition based approach whereby 240,000 different sequences compete to bind to a single RBP (Ray et al. 2009). The RRE for the RBP is inferred by considering the affinity of every possible 7-mer for binding to the protein and calculating cognate E and Zscores. Recently, RNAcompete delineated the binding preference of 205 different genes from 24 diverse eukaryotes (Ray et al. 2013). This study also revealed that RBPs with similar RNA binding domains (more than 70% identity) typically have similar binding preferences. This observation suggested that binding site information for one RBP could be reliably transformed to other RBPs with a conserved RNA binding domain. However, because of the early-branching of Kinetoplastids in evolution from other eukaryotes, the binding preferences of their RBPs are slightly different from their homologs in other metazoans (Ray et al. 2013). Therefore, to validate the human results, we examined the similarity of each GRAFFER predicted motif to all RNAcompete motifs, excluding Kinetoplastids. In the same way, we compared GRAFFER motifs derived from kinetoplastids with the identified RREs of these organisms. To determine if a GRAFFER motif represents significant similarity with an RNAcompete motif, we set two criteria: 1) sequences containing the computationally predicted motif should be preferentially bound by the corresponding RBP; 2) both RNAcompete and GRAFFER motifs should show similarity at the sequence level, ensuring they both target a similar set of genes. To measure the preference for binding, RNAcompete probes containing the GRAFFER motif were identified and their preferences were examined using the Mann-Whitney sum of ranks test statistic (Benjamini–Hochberg corrected p-value cut-off threshold of 0.05). To

consider the similarity with the RNAcompete motifs, we extracted the consensus pattern of each RNAcompete motif, represented in the IUPAC-ambiguity codes. We then determined the enrichment of a predicted motif in an RNAcompete assay as valid, only if the well conserved region, i.e. the discriminative part, of the RNAcompete consensus pattern shared common sequences with the predicted motif. The well-conserved region of a consensus pattern is defined as the region comprising all one and two-degenerate positions (A,U,C,G, S=[CG], W=[AU], Y=[CU], R=[AG], M=[AC], K=[GU]). For example, the conserved region of RNCMPT00138 (from an RNAcompete assay) with consensus pattern of XXVUGAV is XXVUGA. However, the highly degenerate parts of computationally predicted motifs can match with many different conserved regions derived from different RNAcompete assays. For example, computational motifs that contain the fully degenerate sequence of length five (NNNNN), share common sequences with all well conserved regions of length five. To address this issue, we defined a degeneracy rate measure as the entropy of the part of a computational motif that matches with a well-conserved region divided by the entropy of a fully degenerate sequence with the same length. We only accepted matches with a degeneracy rate of below 50%. In cases where more than one GRAFFER motif matched to the RNAcompete assay, the motif with the highest enrichment was selected.

Comparison with previous studies

To evaluate the performance of our graph-based approach, we compared the GRAFFER results with three other genome-wide studies conducted on *T. brucei* (Mao et al. 2009; Shateri Najafabadi and Salavati 2010; Najafabadi et al. 2013). It is important to note that RREs are not extensively characterized in *T. brucei*; which forced us to compare the results of each study with a limited set of previously known RREs. Therefore, some of the novel RREs predicted by these approaches may be valid, but not discovered yet. Two of these studies applied the FIRE program (Elemento et al. 2007) in different contexts to predict RREs. FIRE is an information theory-based approach that seeks informative RREs from clusters of co-expressed genes. An independent experimental study showed that the predicted motifs for human are of high quality (Hu et al. 2009). The third study applied an alignment-free approach, which benefits from simultaneous consideration of four closely related Trypanosomatid species: *T. brucei*, *T. cruzi*, *T. vivax* and *T. congolence*. In the first genome wide analysis of *T. brucei* genes, the lack of

genome-wide experiments available at the time caused the authors to predict “function-specific” RREs by clustering genes according to their function (Mao et al. 2009). This analysis led to the identification of 21 RREs in the 3′-UTRs of *T. brucei* genes. Considering the same criteria as applied for the GRAFFER motifs, four out of the 21 predicted motifs showed significant similarity with only four different RNAcompete motifs. Predictions did not match with other experimentally-derived motifs. In the second genome-wide analysis of *T. brucei* genes, whole genome microarray data was available; therefore, the authors employed a sophisticated approach for direct integration of transcriptome measurements obtained from three independent studies (Shateri Najafabadi and Salavati 2010). Importantly, two of the transcriptome datasets used in the study are also used for predictions of RREs in our approach. Clustering of the co-expression network and application of FIRE algorithm in this case had led to the prediction of 14 RREs. Comparison with RNAcompete results revealed that three of the 14 predicted motifs showed significant similarity with only three different RNAcompete motifs. Predictions did not match with other experimentally-derived motifs.

In the third genome-wide analysis of *T. brucei* genes, a novel algorithm (COSMOS) was developed on the assumption that orthologous genes in close organisms tend to have a similar set of RREs (Najafabadi et al. 2013). Application of COSMOS on four closely related Trypanosomatid organisms revealed 222 linear and 166 structural motifs that are conserved among these four organisms.

Comparison with RNAcompete results revealed that nine of the 388 predicted motifs had significant similarity with nine different RNAcompete motifs. However, considering the GRAFFER and COSMOS motifs that matched to the same RNAcompete motif, in all cases the GRAFFER motifs showed higher selectivity (higher enrichment) than the COSMOS motifs.

It should be pointed that COSMOS was also able to identify three further well-studied motifs. One of them is a structural motif that could not be predicted in the current implementation of GRAFFER algorithm (GRAFFER only searches for linear motifs). The other two are cell cycle related motifs. GRAFFER successfully discovered one of these motifs from the transcriptome data of cell cycle progression (see above). However, the second motif is related to a set of transcripts with subtle variations in their expression, as mentioned in (Archer et al. 2011). In our motif prediction pipeline, we constructed co-expression graphs by focusing on highly variable

genes. Therefore, we most probably missed this motif because we did not have its cognate targets in the coexpression graph.

Chapter 4

circTAIL-seq, a targeted method for deep analysis of RNA 3′ tails, reveals transcript-specific differences in *T. brucei* mtRNAs

In the last two chapters, we focused on the post transcriptional gene regulatory programs of nuclear-encoded transcripts with emphasis on the regulatory pathways controlling the differentiation process of *T. brucei*. In addition to the nuclear-encoded transcripts, *T. brucei* finely tunes the expression level of mitochondrial transcripts during its life cycle to remodel the active metabolic pathways and thereby cope with changes in available energy sources. However, mechanisms by which mitochondrial gene expression changes are effected are poorly understood. Post-transcriptionally added 3' non-templated nucleotides, or tails, on mitochondrial mRNAs consisting of adenosine (A) and/or uridine (U) play roles in gene expression regulation and could potentially facilitate differing life stage gene expression. In the present chapter that is also presented as a method paper in RNA journal (Gazestani et al. 2016a), we introduce a novel, targeted approach to study mitochondrial polycistron processing and transcript tailing processes. Application of this novel approach to *T. brucei* mitochondrial transcripts provides new insights on the tail addition processes as well as RNA polycistron processing.

Background

The majority of eukaryotic transcripts acquire a nontemplated nucleotide addition or “tail” on their 3' termini after or nearly simultaneously with transcription. Although tail addition appears ubiquitous, tail length and composition are finely regulated in various cell compartments (e.g., nucleus, cytoplasm, mitochondria, and chloroplast) with implications for the mRNAs stability, transportation, and translation initiation (Zhang et al. 2010; Norbury 2013). Interestingly, the regulatory roles of these end structures can differ based on cellular needs. For example, while extension of tails in the early embryonic stages of zebrafish and xenopus enhances the translation rate of cognate transcripts, experimental data does not support the functionality of this regulatory process in the later life stages (Subtelny et al. 2014). Moreover, tails may have different states, each with a distinct regulatory role and biological impact on transcripts; e.g., while one tail state regulates the stability of a transcript, the other state with possibly differing tail length and/or composition regulates the translational rate of the transcript (Aphasizheva et al. 2011).

The average tail length of cytoplasmic transcripts varies between different organisms from relatively short lengths of 20-30 nt in yeast to less than one hundred for some metazoan organisms (Chang et al. 2014; Subtelny et al. 2014). The situation for the organellar mRNAs is also varied. Yeast mitochondrial transcripts (mtRNAs) entirely lack tails, adenine (A)-rich and uridine (U)-rich oligomers are part of the chloroplast and mitochondrial mRNA decay pathways in plants and algae, and multiple roles of poly(A) and other tails on human mtRNAs appear transcript-specific (Schuster and Stern 2009; Zimmer et al. 2009; Chang and Tong 2012; Rorbach and Minczuk 2012). However, the biological implications of tail variations within and across organisms are largely unknown.

The *Trypanosoma brucei* mitochondrion is an interesting system to study tailing mechanisms and their regulatory impacts on the transcripts, where an apparently convoluted tailing process is intertwined with other post-transcriptional events. Expression of *T. brucei* mitochondrial genes starts with their constitutive polycistronic transcription, followed by processing that is coupled with addition of fairly ubiquitous short tails. These are termed here as (in)itial tails or “in-tails” and are thought to mediate the transcript stability (Ryan et al. 2003; Kao and Read 2005; Etheridge et al. 2008; Aphasizheva and Aphasizhev 2010). Tail addition and modification are also linked to a unique type of editing that 12 trypanosome mtRNAs must undergo prior to translation (Stuart et al. 2005; Aphasizhev and Aphasizheva 2011; Hashimi et al. 2013;

Aphasizhev and Aphasizheva 2014). Finally, transcripts are potentially marked for translation by extensions appended to a subset of in-tails on only translatable mtRNAs. Extensions contain both A and U and are described as having a 7:3 A/U ratio (Etheridge et al. 2008), with a fairly consistent frequency of switching of addition from A to U and back. We are naming these latter, presumably translation-associated tails that possess these described extensions "ex-tails".

During the past decades, biochemical and other approaches have been developed for the identification of tail characteristics of populations of individual genes (Temperley et al. 2003; Beilharz and Preiss 2011; Slomovic and Schuster 2013). Some of these approaches provide only qualitative information on the tail length. Others such as circular RT-PCR or cloning of end-adapted 3' ends are labor-intensive and thus examine characteristics of a relatively small sample of tails from the transcript of interest. These limitations hamper their application to: 1) quantitative comparison of tail characteristics of the transcript in different biological conditions, 2) accurate description of multiple tail states in the tail population of the transcript; and 3) identification of tail characteristics for rare, but biologically interesting transcripts or degradation intermediates.

High throughput sequencing approaches to genome-wide tail inference of transcripts are now available and have profoundly expanded our understanding of the functions and regulatory potentials of tails (Chang et al. 2014; Slevin et al. 2014; Subtelny et al. 2014; Welch et al. 2015). However, due to the genome-wide design of these approaches, the number of sampled tails for most transcripts is still below 100 and can be even significantly less than that (or zero) for low abundance transcripts, a limitation that might explain, in some extent, the observed discrepancies between the results of some of these approaches (Lee et al. 2014; Zheng and Tian 2014). Therefore, genome-wide approaches are not universally suitable for in-depth analysis of tail characteristics for a focused subset of transcripts of interest.

As the second aim, we developed an approach to characterize tails on transcript populations and quantitatively compared tail populations of transcripts. For in depth and high-resolution analysis of tail population for a transcript of interest, we coupled conventional circular reverse transcription – polymerase chain reaction (cRT-PCR) to next-generation sequencing techniques and termed this "circTAIL-seq". As proof-of-principle, we applied the circTAIL-seq approach to two mtRNAs of *T. brucei*. The depth of circTAIL-seq allowed us to accurately detect and quantify different tail states and subtle differences of tail populations. Furthermore, we also

captured tail characteristics of rare but biologically intriguing variants of mtRNA ends that would likely be missed with other approaches. The diversity of tail lengths and compositions on trypanosome mtRNAs proved a tremendous asset for circTAIL-seq development. We describe a methodology that addresses both experimental and computational aspects to identify and characterize the tail population of target transcripts.

Results

Capturing tail census of a transcript by circTAIL-seq

The circTAIL-seq approach is composed of three major steps: library generation, next-generation sequencing, and the informatics workflow to extract tail information from the raw sequencing output (Fig 4.1.a). Library generation parallels the conventional 3' tail analysis of cRT-PCR used to investigate RNA 5' and 3' ends. (Perrin et al. 2004a; Perrin et al. 2004b; Slomovic and Schuster 2008; Aphasizheva and Aphasizhev 2010; Aphasizheva et al. 2011; Zimmer et al. 2012; Slomovic and Schuster 2013). Total RNA is first circularized using RNA ligase. Next, carefully positioned gene-specific primers containing adaptor sequence are used in reverse transcription and PCR to generate tail-containing amplicons bridging the 3'-5' junction that can be directly used as Illumina sequencing libraries. As detailed in Materials and Methods, the obtained reads are then preprocessed and subsequently aligned to the reference sequence for the gene of interest to identify the embedded tails and associated 3' and 5' termini sites (Fig 4.1.b).

We initially performed a pilot sequencing experiment on two *T. brucei* mitochondrial transcript tail populations acquired as described above. Results clearly indicated that sequential method optimization for circTAIL-seq was required. Table 1 describes read usability between the pilot experiment and two subsequent trials (total of three separate library preparation and sequencing trials). The initial trial was performed at the smallest possible sequencing scale on two single transcript amplicons. With it, we confirmed that amplicons could be subjected to deep sequencing, and also obtained reads with which to develop an informatics workflow. Studies of 5' and 3' RNA ends often used end-ligation of adaptors to RNA, eliminating the need for gene-specific reverse primers by semi-specific amplification of the target gene. In this first trial, we also sequenced amplicons of 3' end-adapted rather than circularized RNAs generated as

described for miRNA end-sequencing (Diebel et al. 2010) facilitated by pre-adenylation of the adaptor (Hafner et al. 2008). As we obtained only a few hundred gene-specific reads within the entire Illumina read file from end-adapted amplicons (not shown), we proceeded with optimization of circTAIL-seq.

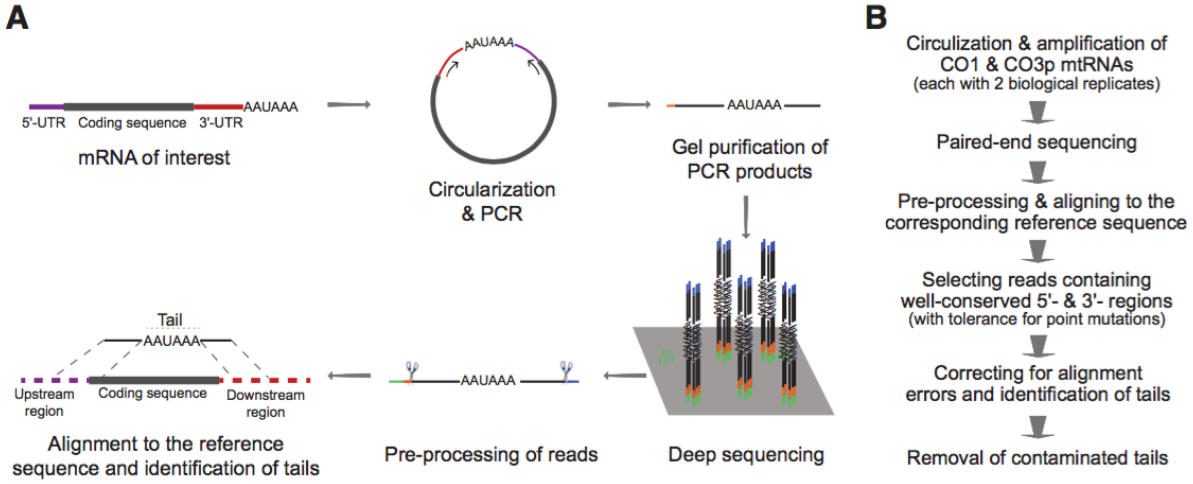


Fig 4.1. Experimental and computational steps of circTail-seq
(*a*) Schematic illustrating steps of circTAIL-seq. Thick gray line indicates a coding region of a generic mRNA that is not amplified in the process. The generic RNAs 5′ UTR and 3′ UTRs are represented in violet and red, respectively. "AAUAAA" represents all potential nonencoded tails on the ends of the RNAs. The orange region on PCR products is the bar code introduced during PCR. (*b*) Informatics workflow for circTAIL-seq data analysis.

The second trial was performed on a larger set of individual libraries that more realistically mirrored the number of samples we anticipate from the application of circTAIL-seq approach. We included transcripts in two biological replicates in order to gain insight into reproducibility. Reads for the first two trial experiments confirmed that primers annealed in locations resulting in generated amplicons containing enough transcript 5' and 3' end sequence for the subsequent tail extraction program. Unfortunately, very few tails that met the traditional definition of an ex-tail (Etheridge et al. 2008; Aphasizheva et al. 2011; Zimmer et al. 2012) were observed in either of the first two trials; we found instead that most of the longest reads were artifacts resulting from aberrant PCR amplification. Thus, optimization of the amplicon-generating PCR reaction was necessary to reduce or eliminate artifacts. We developed a PCR optimization protocol to optimize PCR reactions for each transcript. In addition, we changed our thermostable polymerase to one requiring short annealing and extension times that seems to reduce artifact abundance

(data not shown). In conclusion, our first two optimization experiments suggest that for any circTAIL-seq experiments, especially those involving transcripts of unknown 3' and 5' UTR length, a trial amplicon sequencing at nano scale is prudent to verify identity of generated amplicons and also to select proper primer residing location.

The third sequencing trial performed on the same targets as trial 2 with amplicons generated using optimized PCR provided us ample reads and very few artifacts. There were no clear influences of read yield on analysis in this trial where analyzable tails varied from ~150,000 to ~1.5 million per sample. Nor was there an influence attributable to having a particular barcode in the primer sequence. High proportions of the returned reads contained primer annealing sequence and were deemed usable. Encouragingly, some of the sample files contained reads in which ex-tails were observed by manual perusal. Importantly, the fraction of reads appearing only once in each sequencing run (ranging from 4.5% to 38%) indicated that abundance of circularized mitochondrial templates in the initial RNA pool is not limiting, suggesting that even by devoting high throughput sequencing to a pre-specified transcript, we may not fully capture the tail diversity of the transcript (Fig 4.2). Finally, we note that sample PCR product abundances are quantitated individually and multiplexed with a goal of equal proportions of each sample the multiplexed sequencing reaction. Therefore, differences between tail populations to be compared will typically be less than an order of magnitude, sometimes differing no more than two-fold. In conclusion, we have developed amplicon generation and sequencing protocols for circTAIL-seq that provide adequate sample sizes, high percentages of usable reads, and the potential for capturing longer tail sequences.

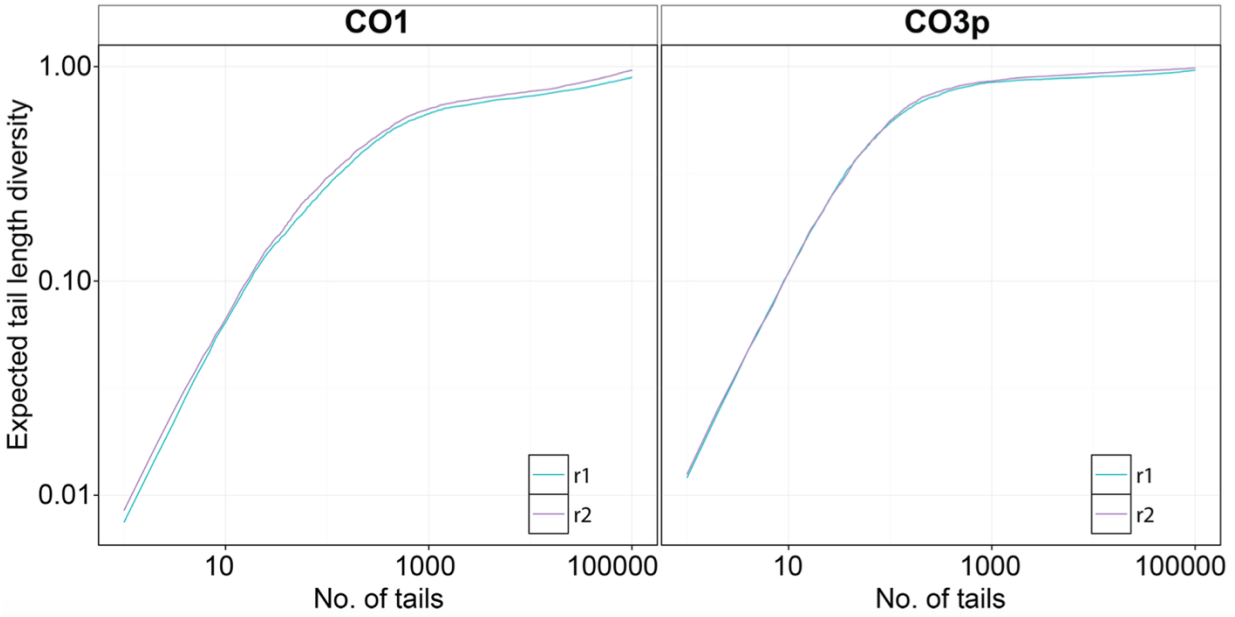


Fig 4.2.Captured tail diversity relative to the sequencing depth.
For each sample, the graph represents the expected tail length diversity across different tail population sizes, relative to the total observed diversity in the sample. As illustrated, in most cases a deep coverage is reached only in sample sizes of greater than 1000 tails. For each sample, the expected tail length diversity at each population size is calculated by averaging the observed tail length diversity of 100 random population drawn from the sample. Only tail length diversity is represented by the graphs and tail composition variation that might exist for each length is ignored.

Two T. brucei mitochondrial transcripts were selected to demonstrate the analytical potential of circTAIL-seq

We next developed methodologies that could eventually be used to compare mtRNA tails between transcripts and changes in tails in response to internal and/or environmental stimuli. Here, we focus on the methodology development and proof-of-principle for circTAIL-seq using selected replicate tail populations of only two *T. brucei* mitochondrial transcripts, CO1 (cytochrome oxidase subunit I) and pre-edited CO3 (CO3p; encoding cytochrome oxidase subunit III). The tailing process of *T. brucei* mitochondrial transcripts is regulated and is significantly coupled with the RNA editing status of transcripts, i.e., although most mitochondrial transcripts acquire in-tails, only the transcripts with correct open reading frames (ORFs) can undergo the ex-tailing process that mark them for translation. CO1 gene encodes the transcripts with correct ORF, so can be ex-tailed and, subsequently, translated upon cleavage to the monocistronic form with no need for editing. It is known from conventional RNA blots that a

sub-population of CO1 is ex-tailed, although details of the length differential are not possible to garner with that method. In contrast, CO3p transcript does not possess a translatable ORF. Thus it cannot be associated with the ribosome and we would not expect it to be ex-tailed (Aphasizheva et al. 2011). Additionally, limited published tail sequences suggest that the typical length and composition of CO1 and CO3 in-tails would likely be different (Decker and Sollner-Webb 1990; Kao and Read 2007). Because of these differences, biological replicates of tail populations on CO1 and CO3p transcripts were used to illustrate the consistency of circTAIL-seq results.

Analysis of circTAIL-seq data demonstrates complexity in tail populations that may not be captured in low-resolution settings

Large-scale analysis of tail populations has been performed on tails of cytosolic mRNAs (Chang et al. 2014; Subtelny et al. 2014; Welch et al. 2015), histone mRNA degradation products (Slevin et al., 2014; Welch et al., 2015), and miRNAs (Wyman et al. 2011), yet trypanosome tails are far more complex than these. Analyzing complementary aspect of tails with multiple metrics proved useful because it provided multiple lines of evidence to draw conclusions. Moreover, to ascertain that our reported results are not overwhelmingly affected by potential PCR amplification bias favoring short length sequences, we verified that reported results are supported by the analysis of unique reads (not shown) as well as total reads. This was possible as a benefit of the extremely high heterogeneity of read sequences.

Tail Length: The simplest tail characteristic to describe is its length. We constructed probability density functions from the collected tails in each sample, with the area under each density curve set to 1. Profiles for replicate samples for CO1 and CO3p are displayed on single graphs in Fig 4.3.a, allowing us to analyze reproducibility and transcript-specific differences. The density curves demonstrated overall good reproducibility of tail length data. Comparison of density curves indicated that CO3p transcripts have significantly longer tails than CO1 (p-value <2.2E-16, Wilcoxon-Mann-Whitney rank sum test with pooling the replicates). Additionally, CO3p tail length distribution showed wide variation with two major peaks, while the majority of CO1 tails were narrowly distributed in length with a second minor population in a longer length category of 50 or more nucleotides.

Tail-less reads

Tail length density curves also demonstrated the presence of tail-less reads in circTAIL-seq results (Fig 4.3.a). The percentage of CO1 tails lacking reads ranged from 0.02% to 0.84%, barely detectable, while CO3p contained higher percentages of tail-less reads ranging from 1.4% to 8.8%. We found that tail-less reads reproducibly have significantly shorter 3' untranslated regions (UTRs) than the tailed reads in all biological replicates of both transcripts (p-value < 2.2E-10, Wilcoxon-Mann-Whitney test), Shorter 3' ends may in fact be decay intermediates, suggesting circTAIL-seq approach has been able to capture this difficult-to-detect population. Another interesting observation was the moderate enrichment of tail-less reads with unusually long embedded 3' UTRs (Fig 4.3.b). The polycistronic nature of mitochondrial transcription suggests that these reads might be RNAs currently undergoing 3' exonucleolytic cleavage to generate the mature, typical 3' end from a larger precursor. Evidence such as this would be supportive of the idea that 3' exoribonucleases act during polycistronic mtRNA processing (Mattiacio and Read 2008; Aphasizhev and Aphasizheva 2011). Hence, circTAIL-seq analysis of all mtRNAs could yield valuable insights into the process of polycistron processing, of which little is known.

Tail Composition

The other important characteristic of 3' RNA tails is their overall nucleotide composition. Fig 4.3c presents density curves for the percentage of A in total composition from none (value of 0) to 100% (value of 1) *per tail* for each sample's population. We found that CO3p incorporates far more Us than CO1. This was surprising as CO3p transcripts are expected to be exclusively in an in-tailed state that previous studies commonly described as primarily poly(A) (Bhat et al. 1992; Etheridge et al. 2008; Aphasizhev and Aphasizheva 2011; Aphasizheva et al. 2011). To probe this distinction, we examined the nucleotide composition in each tail population as a function of nucleotide position in the tail (Fig 4.4.a) by first merging the two biological replicates of each transcript. The analysis was performed out to 60 nt, after which no additional differences were observed (not shown).

Results for tails on the two transcript populations analyzed indicated A is the most common nucleotide in all positions but the first three nucleotides of CO3p tails. However, U abundance not only differed between transcripts, but showed transcript-specific variability in positioning. In

CO1 tails, Us are rare until about position 17, where they increase in frequency to reach to the relative frequency of about 1/3 in position ~28, and remain near this frequency afterwards.

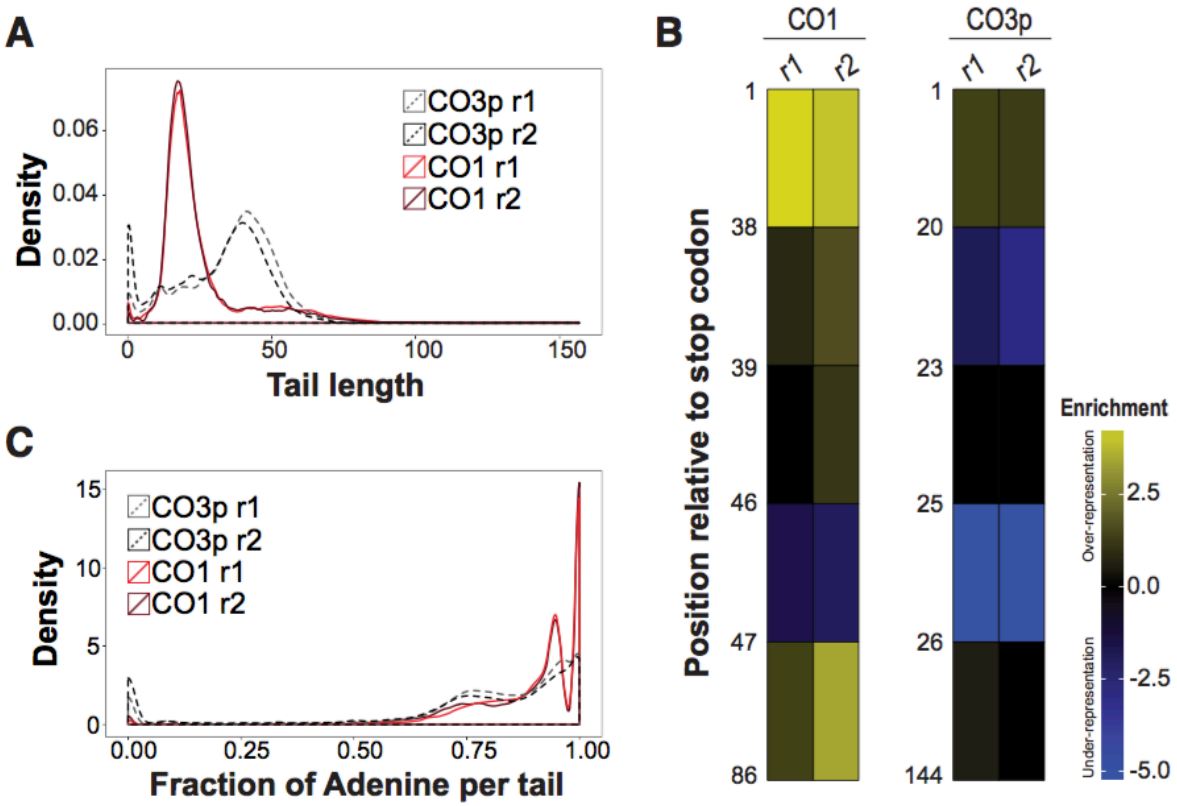


Fig 4.3. Characteristics of tails inferred by circTail-seq
(*a*) Density curves comparing lengths of tail populations from CO1 or CO3p transcripts. CO1 tails >100 nt are present but are not abundant enough to be observed on a chart of this scale. (*b*) Enrichment analysis of tail-less reads in five roughly equally populated bins. The figure is pseudo-colored, showing only significant enrichments (*P*-values <0.01 and fold-enrichment >1.5), with blue and yellow colors indicating underrepresentation and overrepresentation of tail-less reads, respectively. (*c*) Density curves comparing fraction nucleotides that are "A" in each tail [0 = tails that are oligo(U), 1 = tails that are oligo(A)] from CO1 and CO3p transcripts. Distributions were inferred using R statistical package. Area *under* each curve is 1. "r1" and "r2" are biological replicates.

Assuming ex-tail sequence extensions to start somewhere between tail position 20 and 40 nt (Aphasizheva et al. 2011; Zimmer et al. 2012), this observation on the CO1 transcript is consistent with a transition from an in-tail (mainly As) to an ex-tail (extension of the in-tail with A/U composition of approximately 7:3 ratio). In contrast, Us on CO3p tails are at positions consistent with belonging to in-tails, demonstrating that U can be a common component of in-tails for some mitochondrial transcripts. These observations indicate that analyzing tail

composition by position (possible only with high numbers of tail sequences) can provide important insights not possible by analyzing overall composition alone. To summarize our entire analysis of tail composition, CO1 and CO3p appear to have overall nucleotide compositional differences (Fig 4.3.c) as well as positional signatures (Fig 4.4.a).

Other differences in nucleotide patterns

While the above metrics are highly informative, further details can be elucidated from the tails data that shed light on potential differences in in-tail and ex-tail A/U addition patterns. For instance, a tail comprised of 50% A and 50% U can consist of alternations of single As and Us, or of alternating homopolymer stretches of As and Us. The initial regions of tails in Fig 4.4.a represent the sum of the entire population, and therefore could consist of sub-populations of A and U homopolymers, or a single population that is fairly heteropolymeric. Fig 4.4.b presents deviation from expected probability of each listed nucleotide tetramer at every position along the first 60 nt of all tails. These heat maps are arranged so that poly(A) tetramer is the top listing, combinations of single Us within A stretches are next, then alternating A/U combinations, next U polymer stretches interspersed with A, and finally a U tetramer. Therefore, vertical locations in the heat map are more or less associated with multiple versus single additions of a certain nucleotide.

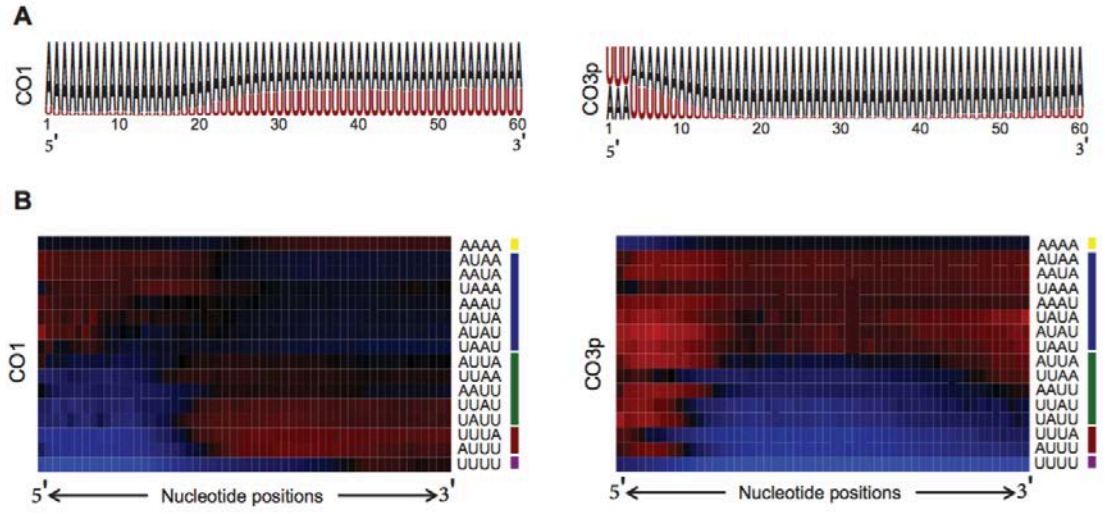


Fig 4.4. Compositional characteristics of the inferred tails
(*a*) Relative abundance of each nucleotide at each position from tail positions 1–60. All tails possessing a nucleotide at the analyzed position in the total population were considered. (*b*) Heat map describing the occurrence of the indicated tetramer (nucleotide position 1 of the tetramer at the position indicated at the*bottom*), relative to the likelihood of that tetramer given the average nucleotide compositions at those positions under an independently distributed model. "Nucleotide positions" are tail positions 1–60 from 5′ to 3′ as in *A*. Both plots are colored on the same scale; the lowest (most intense red) value was 0.0217 (occurring 1/46th as often as expected under an independently distributed model), and the highest (most intense blue) value was 3356 (occurring 3356 times as often as expected under an independently distributed model). At each position, the entire population of tails possessing a full tetramer starting at that position is analyzed.

Fig 4.4.b demonstrates that with tail numbers obtained from circTAIL-seq, we can observe variable tetramer probability frequencies, with different tail nucleotide heat map patterns observed for CO1 and CO3p. U polymers of two nucleotides or more were over-represented in CO3p tails along the entire analyzed region of 60 nt with the exception of the first 10 nt, where only U tetramers were over-represented. In contrast, single Us within the sequence were either under-represented or occurred approximately as expected. As CO3p is a pre-edited transcript that should not associate with the ribosome, we do not expect to see a transition to ex-tails with frequent A/U alterations. This result is consistent with such an expectation.

While U polymer-containing tetramers are also over-represented in CO1 tails, this over-representation only occurs within the first 20-30 nt. This likely relates to the fact that for CO1, nucleotide addition beyond the first 30 or 40 is in ex-tail state and thus switching of A/U addition is far more frequent. Analyzing the deviation from expected probability of these different tetramers provided important observations, especially when we concentrate on differences specifically between tetramers consisting of U stretches of two or more (homopolymer) compared to tetramers containing single Us. Overall, U homopolymers are overrepresented in tails from a transcript where we expect to find no ex-tails (CO3p) as well asthe initial tail regions of tail populations from a transcript where we expect to find ex-tails (CO1). In contrast, we see tetramers containing single Us become overrepresented in the 3' end of CO1 where we expect to find sequence consistent with ex-tail additions. Until now, demonstrating the existence of an ex-tail among tail sequences required publishing the tail sequence and indicating the site of transition to frequent nucleotide switching compared to the beginning of the tail (Etheridge et al. 2008; Aphasizheva et al. 2011). This tetramer analysis therefore represents a way to demonstrate an increase or change in frequency of nucleotide switching in an entire population. We note, however, that although our sequencing strategy detects both types of tails, because of possible PCR amplification bias, the relative abundance of ex-tails to in-tails may not be a true representative of the underlying populations. Because of this, transitions in deviations from expected tetramer frequencies may be even more dramatic than suggested by our plots.

circTAIL-seq can capture differences in both 5' and 3' UTR lengths

Aligning reads to a reference sequence also provided 3' and 5' termini information for each transcript population. Sequenced amplicons can thus reveal UTR lengths and degree of termini

homogeneity. For instance, analysis of the 5' termini derived from CO3p reads is shown in Fig 4.5, and the most frequently encountered 5' terminus (the largest “U” in 5' UTR sequence) was the terminus previously identified by primer extension (A. Estevez and L. Simpson, unpublished). Therefore, transcript termini can also be defined by circTAIL-seq.
This is particularly useful when UTRs are heterogeneous in length, such as the 3' UTR of CO3p, which has much higher length heterogeneity than what we observed for CO1. We hypothesize that transcripts with high UTR length heterogeneity undergo more exonuclease processing after a downstream cleavage than transcripts such as CO1. In contrast, CO1 possesses a 3' UTR that could have been generated by a tightly controlled endonuclease cleavage event, although proving this is well beyond this study’s scope.

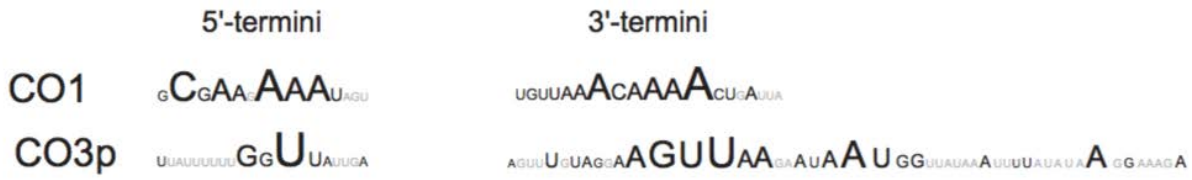


Fig 4.5. Identified 3′ and 5′ termini of transcripts
3′ and 5′ UTR termini derived from populations of circularized molecules. Replicate experiments have been pooled by first normalizing for tail counts, so average density for each nucleotide is shown. Black represents the termini with occurrence probability >0.01%, with the size of the nucleotide corresponding to its frequency as a terminus. The gray nucleotides represent positions with probabilities between 0.1% and zero that exist between nucleotide positions that are more frequently termini. Other nucleotides that are termini with a probability <0.1% are not shown.

DISCUSSION

Here we have developed experimental and computational methodologies that allow coupling of circular RT-PCR tail analysis with high throughput sequencing technology. Development required optimization of tail-containing amplicon generation, Illumina sequencing modifications to adjust for amplicons containing problematic regions of identical sequence, and development of a stringent informatics workflow to separate true tails from contaminating sequences. Using this rigorous process we obtain 100,000 – 1,000,000 tails per transcript amplicon.

circTAIL-seq has three major advantages over conventional circular RT-PCR approach. First, it eliminates the cloning step that is the most labor intensive, and thus limiting step of the circular RT-PCR approach. Second, the bias of circularized RT-PCR approach toward collecting a population's shortest tails, introduced in both the PCR and cloning steps, is reduced (albeit not removed) by eliminating the cloning step. Benefiting from the extremely high heterogeneity of read sequences that stem from variations in tail length and composition as well as 5' and 3' termini, we were able to show that the circTAIL-seq data are minimally affected by PCR bias as considering either total reads or unique reads for the analysis led mainly to the same results. For future experiments, a spike-in addition of RNAs of known length and frequency could be added to remove the bias completely if it is deemed necessary.

Third and most importantly, although augmentation of tails from low-abundance cytosolic reads is possible (Welch et al. 2015), circTAIL-seq to the best of our knowledge provides highest depth of tail analysis compared to other techniques developed thus far. It provides a high-resolution picture on tail population for specific transcripts of interest, and is possibly the only current method that is efficient for analyzing 5' as well as 3' ends of organellar RNAs. This critical characteristic empowers the user to go beyond overall average tail characteristics to examine interesting sub-populations of tails within a dataset and identify specific nucleotide patterns that appear in populations. Applications of this approach include determining changes in tail qualities in response to environmental or internal stimuli, or upon silencing of genes of interest in mRNA processing pathways. circTAIL-seq will prove invaluable especially in cases that target transcript tail populations prove low abundance relative to other tails in cells, or tail populations are highly heterogeneous such as in *Chlamydomonas* chloroplasts and mitochondria (Zimmer et al. 2009).

We have demonstrated the range of characteristics that can be compared in large tail datasets using this methodology, and provided evidence that previously unknown differences between tail populations exist on trypanosome mtRNAs, many of which may not be captured in lower resolution settings. We were able to define transcript-specific differences in tail populations and concurrently define population-wide 3' and 5' termini of the transcripts. These sequencing and analysis methods can potentially be used to describe and compare 3' and 5' ends of any sort of RNA when specific transcripts are the study's focus.

Finally, the high throughput sequencing setting of circTAIL-seq approach can be adjusted based on the expected tail characteristics for the transcript of interest. Here, the employed setting, 150 bp paired-end, was selected based on the previous biological knowledge on tail characteristics of *T. brucei* mtRNAs which are highly heterogeneous (composed of As and Us) with almost no A homopolymer stretches of longer than 30 nt. Therefore, inaccurate quantitation of length of homopolymers problematic in highthroughput sequencing technologies (Chang et al. 2014) is not a problem in our study, but could be addressed in a context where circTAIL-seq was used to analyze highly homopolymeric tail populations. In summary, there are multiple possibilities for adaptation of circTAIL-seq to answer outstanding questions about the roles of non-templated tails on RNAs.

Materials and methods related to the analysis parts

Read processing

Raw reads (deposited in Sequence Read Archive SRP064265) were sorted by barcode and Illumina primer ends removed. Downstream read processing was performed on a Galaxy platform maintained by the Minnesota Supercomputing Institute. Variable sequences (4, 5, or 6 nt long) that are part of the PCR primers positioned between the Illumina primer sequence and gene-specific primer sequence were removed from both R1 and R2 reads using Trimmomatic HEADCROP task (Bolger et al. 2014), specifying number of nucleotides to remove. R1 and R2 reads were then merged into single consensus reads using PEAR ((Zhang et al. 2014); default settings). After conversion by FASTQ groomer, consensus reads were reverse complemented so reads would possess the proper directionality. This file was verified for per base sequence quality of 30 or better. The sequence was then subjected to a search for gene-specific primer annealing region of reverse amplification primer for the 5' end at 80% similarity. Reads fulfilling this criterion were selected for the follow up analysis.

Tail inference

We developed a software package written in C#, called circTAIL-seq Analyzer (available at http://trypsNetDB.org/circTAILseqAnalyzer.zip), to systematically extract and analyze tail

sequences from the pre-processed reads generated by circTAIL-seq deep sequencing. The circTAIL-seq Analyzer first aligns the reads to the reference sequence for the transcript of interest (including DNA sequence downstream and upstream of CDS) using Needleman-Wunsch pairwise global alignment algorithm (Needleman and Wunsch 1970). To have reliable identification of tail sequences, circTAIL-seq analyzer only considers reads as valid that contain the well conserved 5' and 3' regions of the reference sequence, excluding the binding regions of the primers ("well conserved" reference sequence regions were defined as those regions where 90% or more reads aligned). The program permits a limited number of point mutations in the conserved region (max 2 mutations, different settings for each gene based on the observed diversity and the length of the conserved region) to account for diversity present in the population. The well-conserved regions and allowable number of point mutations can be selected by the program as default values or adjusted based on the users' needs. In the case of CO3p, primers were specifically designed to cover the initial editing region of the mtRNA immediately 5' to the final editing site, thus permitting for selection of tails from RNAs that have not initiated editing as judged by the alignment.

circTAIL-seq Analyzer next infers the embedded tails in the selected reads based on the alignment results. However, visual inspection of results demonstrated small fractions of tails (less than 0.2% of tails in each sample) are contaminated with genomic/transcriptomic sequences (mostly rRNA) that can arise due to fragment incorporation during circularization. Therefore, the program filters out those tails that match (using NCBI BLAST, e-value < 0.001) to a masked version of *T. brucei* reference genome in which interspersed repeats and low complexity parts of the genome were masked out by dustMasker program (Morgulis et al. 2006). The program reports back primary alignment results, reads lacking the well-conserved 5' and 3' regions, tails contaminated with other genomic/transcriptomic sequences as judged by BLAST, the inferred tails for the reads passed filtration criteria with inferred 3' and 5' termini, overall tail counts, tail length distribution, and nucleotide composition distributions.

Chapter 5

Identification of proteins involved in the mitochondrial post transcriptional gene regulatory network of *T. brucei*

In the last three chapters, we focused on RNA-centric approaches to gain insights on the post transcriptional gene regulatory programs in *T. brucei*. While in the last two decades much effort has been devoted to identify and characterize individual RBPs, the interactions that interplay among trypanosomatid RBPs are mostly neglected. As the third objective, we aimed to construct a genome-wide protein interaction network to investigate the interactions of trypanosomatid RBPs in cellular context. To this end, reasoning physically interacting proteins would show similar fractionation patterns, we charted a global interaction map by performing four deep fractionation experiments on whole-cell, mitochondrial-, and cytosolic-enriched cell extracts using two orthogonal techniques of glycerol gradient and ion exchange fractionations. The constructed network is highly informative on proteins involved in processes such as RNA editing, RNA trafficking, and protein translation. Importantly, results suggested involvement of several novel proteins associated with RNA editing machinery that was experimentally confirmed. This chapter is presented as a research paper in PLoS Neglected Tropical Diseases journal (Gazestani et al. 2016b).

Background

Protein interaction maps offer an invaluable resource for functional annotation of proteins (Sharan et al. 2007). Current methodological/instrumental advances have led to the development of several *ex vivo*, *in vivo*, and *in silico* approaches to systematically chart protein interactions and complexes (Yousefi et al. 2012). Optimized yeast two-hybrid (Y2H) approaches have been employed to infer pairwise interactions among proteins (Parrish et al. 2006; Rajagopala et al. 2014). Immunoprecipitation (Malovannaya et al. 2011), biochemical fractionation (Andersen et al. 2003; Havugimana et al. 2012; Kirkwood et al. 2013; Wan et al. 2015), and affinity purification (AP)-based approaches (Krogan et al. 2006; Guruharsha et al. 2011; Wilhelm et al. 2014) are widely used for the identification of protein complexes in a specific cell context. Additionally, functional association of proteins can be predicted computationally using data types such as transcriptomics data (Stuart et al. 2003), synthetic lethality (Zhang et al. 2005), and chemical sensitivity (Giaever et al. 1999). However, each of these approaches has limitations and is inherently associated, in varying degrees, with false positive and negative results. In AP-based approaches, for example, tagging the protein may affect the binding partners of the tagged protein by inactivation, capping the binding site, or changing the localization of the protein. Highly expressed proteins are also often co-purified with the tagged protein as a false positive contaminant. Moreover, transient interactions are likely to be lost if stringent conditions are used for the purification of the tagged protein. In biochemical fractionation strategies, fortuitous interactions can arise because confounding protein complexes may still be present in the same fraction regardless of in-depth fractionation (Havugimana et al. 2012). In addition to non-negligible false positive rates, the Y2H system is relatively weak at detection of co-complex associations, although it works well at capturing binary, particularly transient, interactions (Yu et al. 2008). Therefore, the integration of data from different approaches has been shown to improve the precision of protein maps (Stelzl 2014).

To explore the protein complexes underlying the survival and pathogenesis of *T. brucei*, we performed four high resolution fractionation experiments benefiting from two orthogonal, complementary biochemical approaches. High-resolution mass spectrometry analysis of the fractions led to the construction of a global co-fractionation network for *T. brucei* in procyclic life stage. Evaluation of the constructed network demonstrated that it has topological and

biological characteristics that are similar to those observed in the sampled networks of model organisms from previous large-scale studies. Importantly, our results demonstrated significantly higher precision for those interactions that were supported by the two orthogonal fractionation approaches compared to those that co-fractionated only in one approach. To extract a high-confidence core network, we combined the fractionation-derived network with other orthogonal resources of protein-protein interaction data. This high-confidence network predicts the network context of 866 protein groups, including many hypothetical and experimentally unannotated proteins. Clustering of this high-confidence network led to the assignment of 716 protein groups to 128 complexes. To our knowledge, this study presents the largest proteomics-based interaction map of trypanosomatid parasites to date, providing a useful resource for formulating new biological hypothesises and further experimental leads. To showcase the utility of this protein complex map, we used it to reveal novel factors involved in the mitochondrial post-transcriptional regulation of *T. brucei*, and validated them by several independent experiments.

Results and Discussion

Construction of the co-fractionation networks

Biochemical fractionation techniques have been widely applied to trypanosomatid organisms to test the possibility of physical associations among a set of pre-specified proteins. Fractionation approaches allow dissection of protein complexes based on different biochemical properties. In the glycerol gradient (GG) fractionation approach, protein complexes become separated according to their shape/density. Alternatively, complexes were fractionated on the basis of their overall charge in ion exchange high performance liquid chromatography (IEX) experiments. As summarized in Fig 5.1.a, by coupling GG and IEX deep fractionation techniques with semi-quantitative, ultra-sensitive, mass spectrometry, we generalized the approach to chart a proteome-scale *T. brucei* interaction network. We were able to observe the fractionation pattern of 3354 protein groups (paralogous proteins with nearly identical sequences were grouped together) across total of 133 separate fractions from whole cell-GG (48 fractions), mitochondrial-GG (46 fractions), cytosolic-IEX (19 fractions), and mitochondrial-IEX (20 fractions) experiments on *T. brucei* procyclic form cells (Fig 5.2.a). Due to complementary design of experiments, we were able to observe the fractionation patterns of 1398(42% of total) proteins in

both GG and IEX fractionation approaches, providing a global picture of protein complexes present in *T. brucei* procyclic cells (Fig 5.2.b).

Comparison of our results from mitochondrial experiments with a previously reported repository of mitochondrial proteins (Panigrahi et al. 2009) revealed that 79% of the detected proteins in our experiments were independently supported to be present in the mitochondria. Comparing the mitochondrial signal probability of the proteins identified in our mitochondrial samples with the above mentioned list and also a list of transcripts that are expressed in procyclic life stage of *T. brucei* (Kolev et al. 2010) suggested that the identified proteins were significantly enriched (comparing to proteins identified in (Panigrahi et al. 2009); p-value < 1.4E-08, Wilcoxon-Mann-Whitney rank sum test) for the mitochondrial proteins (Fig 5.2.c). The same analysis on the proteins identified in the cytosolic experiment demonstrated a significant depletion (comparing to total procyclic transcripts; p-value <1.4E-07, Wilcoxon-Mann-Whitney rank sum test) of mitochondrial proteins in the sample. These results demonstrate the accuracy of the employed compartment enrichment procedures.

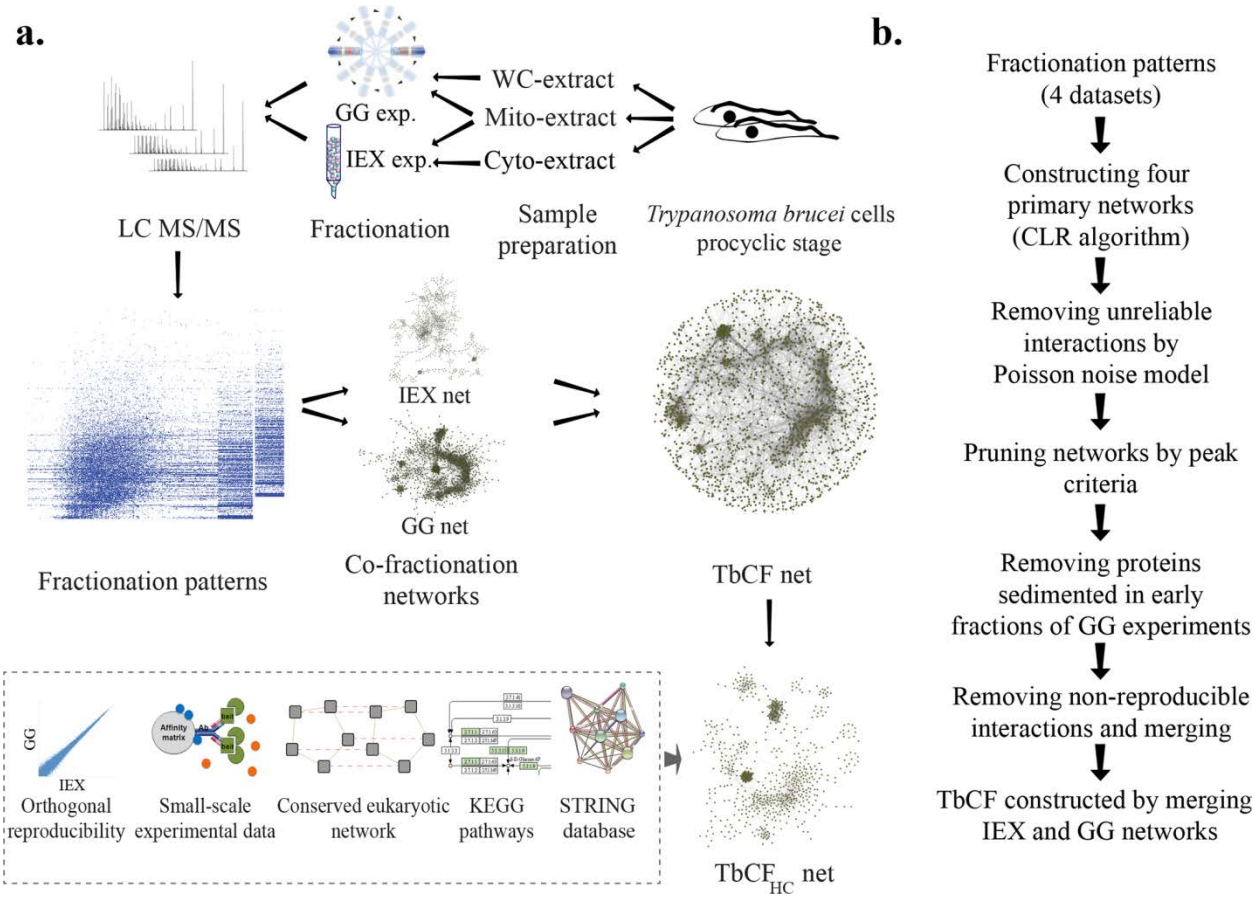


Fig 5.1. The strategy used for construction of a high-confidence protein interaction network.
a) Preparations of cell extracts (whole-cell, and enriched extracts from mitochondria and cytosol) from *T. brucei* procyclic cells were subjected to in-depth fractionation using two different fractionation techniques. In total, 133 fractions were analyzed by mass spectrometry to generate the fractionation patterns for a total of 3354 *T. brucei* protein groups in four separate fractionation experiments. Next, a co-fractionation network for each of four experiments was constructed based on the observed fractionation patterns. After stringent filtration, these four networks were merged based on the employed separation technique to form GG and IEX networks. TbCF net was constructed by those interactions that were present either in IEX or GG networks. TbCF$_{HC}$ net, the high-confidence subset of TbCF net, was generated by identification of interactions that were supported by at least one orthogonal resource. WC: Whole Cell; Mito: Mitochondrial; Cyto: Cytosolic; GG: Glycerol Gradient; IEX: Ion Exchange Chromatography. b) The computational pipeline used for the inference of TbCF net. Starting from the fractionation patterns, four preliminary co-fractionation networks were constructed and then refined with four additional filtration criteria to eliminate spurious interactions. First, those interactions that were sensitive to the addition of noise were eliminated. Second, the unshared-peak interactions were discarded from each co-fractionation network. Next, early co-sedimenting proteins (Proteins that are expected to not be involved stably in protein complexes) were eliminated from the networks. Finally, the non-reproducible interactions were removed and networks related to each fractionation approach were merged to generate GG and IEX networks. The GG co-sedimentation and IEX co-elution networks were merged together to form TbCF net.

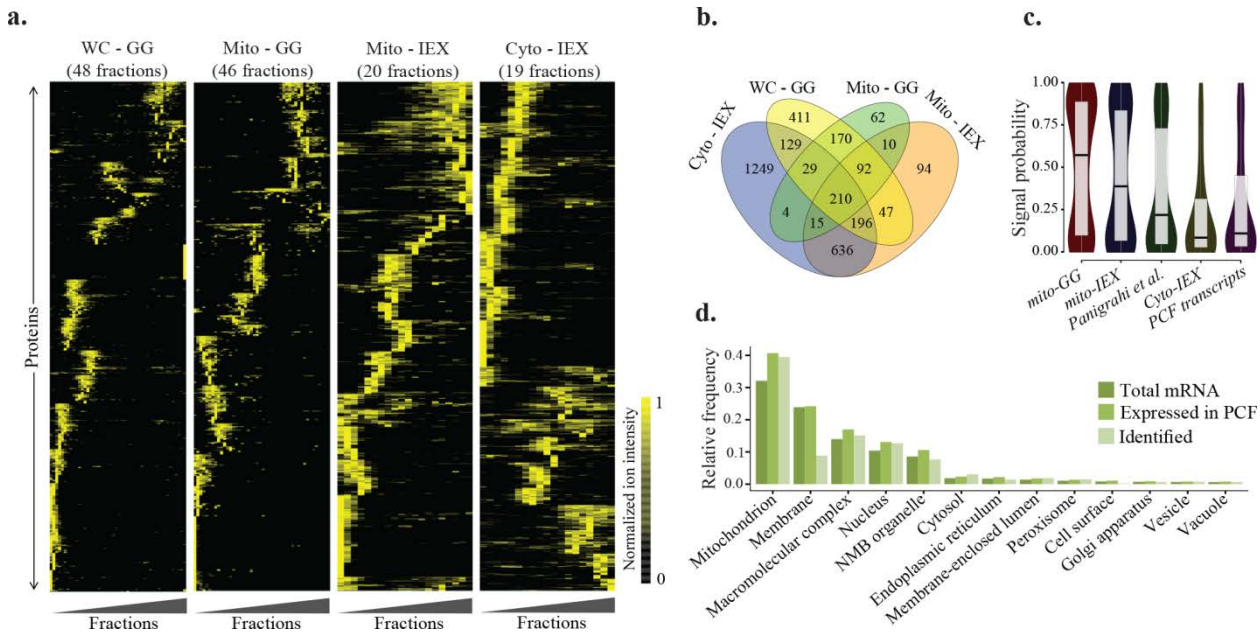


Fig 5.2. High resolution fractionation experiments on the procyclic stage of *T. brucei*.
a) Hierarchical clustering of fractionation patterns for proteins identified in each of four experiments on *T. brucei* procyclic cells. Each row represents a protein and each column a fraction. It should be noted that the number of fractions as well as number of identified proteins varies among datasets (i.e., fractionation experiments). Moreover, datasets were analyzed independently and, therefore, the positions of proteins are not preserved in the graphs. b) Venn diagram representing proteins identified in each of four experiments. c) Comparison of the mitochondrial signal probability distribution for proteins identified in our mitochondrial and cytosolic enriched experiments with those identified by (Panigrahi et al. 2009) and total identified mRNAs in procyclic life stage of *T. brucei* (Kolev et al. 2010). The signal probability for each protein was calculated using the mitoProt web service (https://ihg.gsf.de/ihg/mitoprot.html). As illustrated, the signal probability for our mitochondrial enriched samples were significantly higher (comparing to proteins identified in (Panigrahi et al. 2009); p-value < 1.4E-08, Wilcoxon-Mann-Whitney rank sum test) and for the cytosolic experiment significantly lower (comparing to total procyclic transcripts; p-value <1.4E-07, Wilcoxon-Mann-Whitney rank sum test) than that of the other two groups. d) Gene ontology–cellular component (GO-CC) analysis of all identified proteins in this study compared with the mRNAs that are expected to be expressed in the procyclic stage (Kolev et al. 2010) and total predicted mRNAs in *T. brucei*. NMB Organelle: Non-membrane-bounded Organelle.

In total, proteins identified in our fractionation experiments cover 43% of all expressed protein coding genes in *T. brucei* procyclic stage (Kolev et al. 2010) and 77–127% of number of proteins reported in previous proteome-wide SILAC studies (Gunasekera et al. 2012; Urbaniak et al. 2012; Butter et al. 2013). To test for existence of bias in the data, we examined the distribution of detected proteins in terms of cellular component and number of putative transmembrane domains. As expected, soluble proteins were preferentially detected by the employed fractionation approaches, while the membrane associated proteins were under-represented (Figs 5.2.d and 5.3.a). The detected proteins did not show bias in terms of protein length (Fig 5.3.b).

Moreover, observing the transcriptome responses of the identified proteins in different life stages (Jensen et al. 2009) and over the differentiation process (Kabani et al. 2009; Queiroz et al. 2009) indicated that they show over-expression trend in the procyclic life stage, but biased towards more abundant transcripts in the cell (Fig 5.4). Because paralogous proteins with nearly identical sequences are expected to play similar functions in the cell, we randomly selected one representative protein from each protein group to construct a protein interaction map based on the observed fractionation patterns.

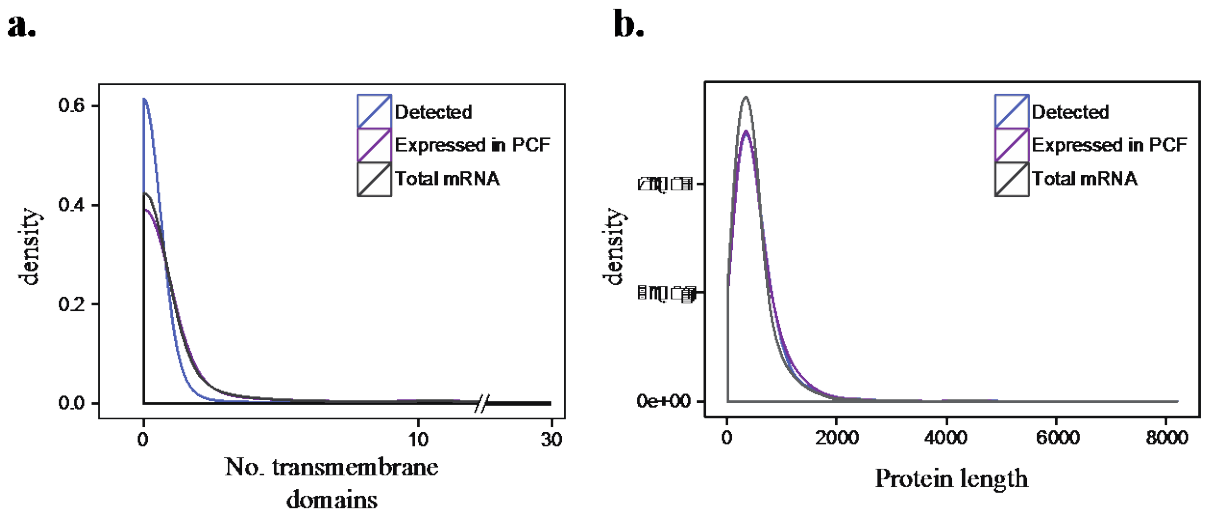


Fig 5.3. Biological properties of the identified proteins in this study.

Comparison of number of transmembrane domains (**a**) and protein length (**b**) distributions for proteins identified in this study with those that are expected to be present in procyclic stage (Kolev et al. 2010) and total predicted proteins in *T. brucei* (TriTrypDB v5).

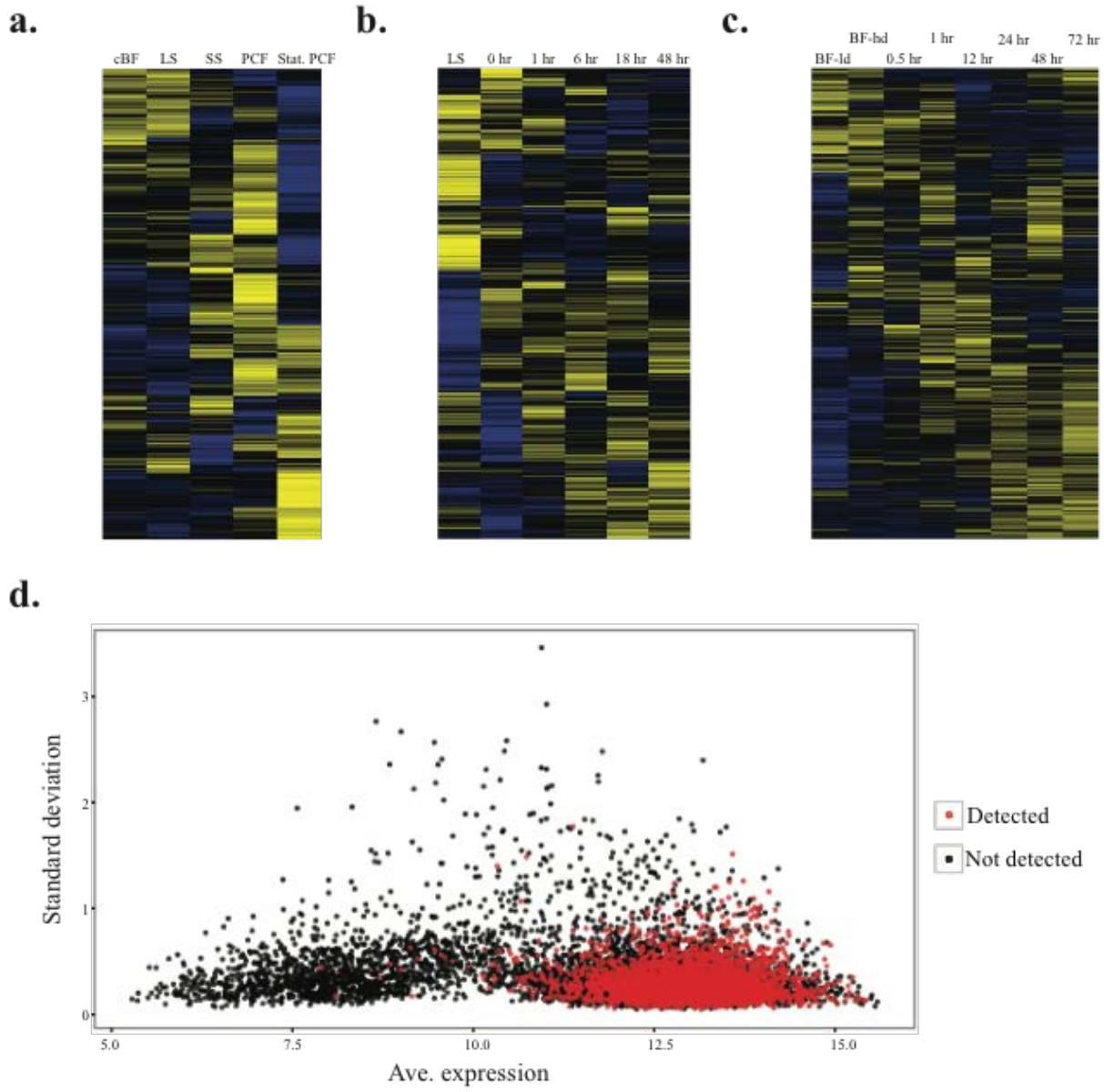


Fig 5.4. Transcriptome characteristics of the identified proteins in different life stages and during the differentiation.

Heatmaps represent expression patterns of the identified proteins in different life stages (**a**, extracted from (*Jensen et al*. 2009)) and during the differentiation process from the bloodstream to the procyclic form (**b**, extracted from (*Kabani et al*. 2009)) and (**c**, extracted from (Queiroz et al. 2009)). For each study, the expression data of each gene was normalized to have mean zero and standard deviation equal to one. The yellow color represents up-regulation and blue indicates the down-regulation. **d)** Plot represents average expression of transcripts against their standard deviations in five different life stages of *T. brucei* (Jensen et al. 2009). As shown, proteins identified in this proteomic-based study are biased towards more abundant transcripts in the cell.

The relative abundance of proteins in each fraction were measured based on label-free MS2 ion intensity-based approaches (Detailed in the method section). To decipher physical associations from the data, we reasoned physically interacting proteins would show similar patterns across the biochemical fractions. The employed bioinformatics approach for the inference of co-fractionation networks based on the observed patterns is summarized in Fig 5.1.b. As described below, for stringent analysis of data, physically interacting protein pairs were predicted based on five criteria: 1) showing significant similarity in their fractionation patterns; 2) being robust to the addition of noise in their observed co-fractionation similarity; 3) showing shared-peak in at least one fraction; 4) being stably interacting with protein complexes as judge by GG-derived patterns; 5) reproducibility of co-fractionation when the same fractionation technique was used. Initially, we constructed four separate co-fractionation networks by application of the context likelihood relatedness (CLR) algorithm (Faith et al. 2007) on each of four datasets, separately. As described in the methods section, the CLR algorithm is an information theoretic-based approach that can predict the association of two random variables (proteins) based on their observed values (fractionation patterns). It predicts two proteins as interacting only if the similarity of their patterns is significantly higher than what expected based on the background (estimated Benjamini-corrected FDR of 5%). Low abundance proteins might not have reliable patterns in the dataset and therefore show random similarities to other proteins in their fractionation patterns. To confirm that observed similarity is not due to presence of noise in the dataset, co-fractionation similarities were recalculated after the generation of Poisson noise models from each dataset. In each network, we discarded those interactions that lost their significant similarity due to the addition of noise (Detailed in the method section). To further reduce the possibility of chance co-fractionation, we only kept those interactions for which both interacting proteins had a peak in at least one shared fraction. A protein was deemed to have a peak in a fraction if its ion intensity for that fraction was at least 80% of its second-highest intensity. We observed that filtering the interactions by this more stringent criterion led to an increased reproducibility rate of results and, most likely, the elimination of putative false positive interactions from the networks (Fig 5.5.a). In GG fractionation experiments, early fractions are highly enriched for monomeric proteins in the cell (e.g., many enzymes) or those that are not stably involved in the complexes. Since interacting partners for these groups of proteins cannot be reliably identified by our approach, proteins that had a peak only in the top two fractions of

GG experiments were discarded from the networks. Cytosolic-IEX and mitochondrial-IEX networks were significantly depleted for proteins of one another (p-value < 6E-35), most likely because of the enrichment procedures used for these two experiments. To check the technical reproducibility of the results, we focused on interactions occurring among proteins that were detected in both GG experiments. As illustrated in Fig 5.5.b, we found that more than half of co-sedimented protein pairs (940 reproducible interactions) in one GG experiment were also co-sedimented in the other (FDR ≤0.05 in one experiment and p-value ≤0.05 in the other). Moreover, comparing the number of common interactions between whole cell-GG and mitochondrial-GG networks with random networks with the same structural characteristics revealed that these two networks are significantly enriched for reproducible interactions (Fig 5.6). These results indicated that significantly co-sedimented protein pairs in the whole cell-GG experiment are typically co-sedimented in the mitochondrial-GG experiment and vice versa. To increase the accuracy, we removed the interactions among those proteins that were not consistently co-fractionated with each other in both experiments (i.e., despite the detection of proteins in both experiments, they co-fractionated in only one experiment). Overall, merging GG networks led to a network composed of 12,196 interactions among 1,417 proteins. The IEX experiments led to the identification of 1,708 interactions among 1,261 proteins. The smaller size of IEX network compared to those of GG is because former experiments had smaller number of fractions and, therefore, were less informative on protein complexes compared to GG experiments. We next assessed the agreements between IEX-derived and GG-derived networks. As shown in Fig 5.5.c, comparison of cytosolic-IEX network with whole cell-GG network indicated that interactions have peak in the reproducible region (median correlation of 0.47 and 0.83 for cytosolic-IEX and whole cell-GG, respectively), but skewed toward the region with negative IEX correlations. Importantly, none of significantly co-eluted protein pairs in cytosolic-IEX experiment had correlations less than 0.47 in the whole cell-GG dataset. Comparison of mitochondrial-IEX network with mitochondrial-GG network led to a similar result (Fig 5.5.d); i.e., interactions had peak in the reproducible region (median correlation of 0.53 and 0.73 in mitochondrial-IEX and mitochondrial-GG datasets, respectively). Moreover, comparison with random graphs demonstrated that the IEX-derived and GG-derived networks were significantly enriched for the reproducible interactions. Due to separation of protein complexes in IEX and GG experiments based on different biochemical properties, the confounding complexes in one

approach have lower chance to co-fractionate in the other approach as well. However, the observed discrepancies in co-fractionation patterns are not only because of confounding elements, but rather reflect the differences in the nature of the two fractionation experiments as well. As an illustration, we observed that members of translation initiation complex co-sedimented in the whole cell-GG network with the median correlation of 0.85. However, their median correlation in the cytosolic-IEX experiments reduced to 0.53. Likewise, comparison of the co-fractionation patterns among a set of proteins enriched for RNA-dependent interactions demonstrated their strong co-sedimentation in the GG experiments, but separation in the IEX experiment. This discrepancy can be due to the use of salt gradient in the latter experiment and disruption of the less stable ionic interaction complexes. However, these effects were minimal on complexes that are known to form more stable complexes such as ribosome, proteasome, F0F1 ATPase, and core editosome (discussed more on the validation part). These results suggested that our GG fractionation experiments better preserved the less stable interactions (i.e. ionic-based interactions), but our IEX experiments favored more stable interactions. For the follow-up analysis, we merged the GG and IEX networks together. This global physical map, named TbCF net (*T. brucei* co-fractionation network), was constructed by distinguishing two types of interactions: 1) Those protein pairs that were reproducibly co-fractionated in both GG and IEX networks (FDR $\leq$0.05 in one experiment and p-value $\leq$0.05 in the other); and 2) Those protein pairs that were co-fractionated only in one experiment. The TbCF net connects 2,151 proteins with 13,865 interactions. The orthogonal reproducible part of TbCF network, termed $TbCF_{OR}$ net, was composed of 2,601 (19% of total) interactions among 828 (38% of total) proteins.

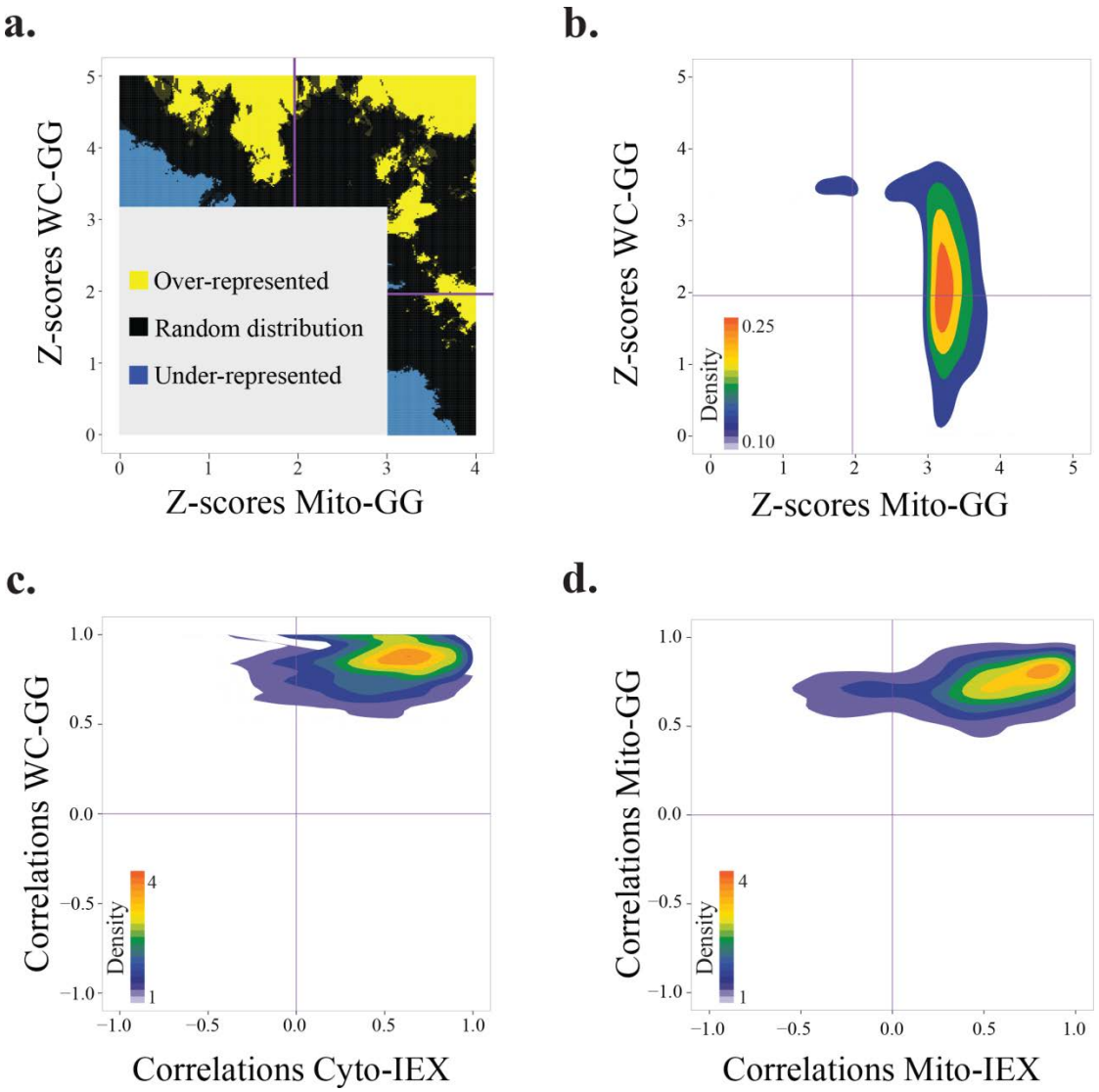


Fig 5.5. Validation of filtration steps used for stringent analysis of fractionation patterns.
a) Interactions present in each of whole cell-GG and mitochondrial-GG networks were categorized as either shared-peak interactions (interacting proteins show a peak in at least one shared fraction) or unshared-peak interactions (interacting proteins show completely distinct peaks). Gray area represents the co-sedimentation space that was not significant in either fractionation experiment. Blue color represents the region depleted for the shared-peak interactions and the yellow region demonstrates the area for over-represented shared-peak interactions. As shown, the shared-peak interactions are highly depleted in non-reproducible regions (upper-left and lower-right regions in the graph). Enrichment at each point on the graph was calculated using a two-tailed hypergeometric test by focusing on the closest 110 interactions to that point (p-value ≤0.05). b) Distribution of z-scores for significant interactions identified in whole cell-GG and/or mitochondrial-GG experiments. The horizontal and vertical purple lines intersect with the XY-axes at points corresponding to the p-value equal to 0.05. As illustrated, more than half of interactions fall in the region that is significant in both experiments (upper-right region). c) Distribution of correlations for significant interactions identified in either whole cell-GG (WC-GG) or cytosolic-IEX (Cyto-IEX) datasets. As shown, interactions have clear peak (color coded in red) in the highly reproducible region (upper-right region). d) Distribution of correlations for significant interactions identified in either mitochondrial-GG (Mito-GG) or mitochondrial-IEX (Mito-IEX) datasets. As shown, interactions have clear peak (color coded in red) in the highly reproducible region (upper-right region).

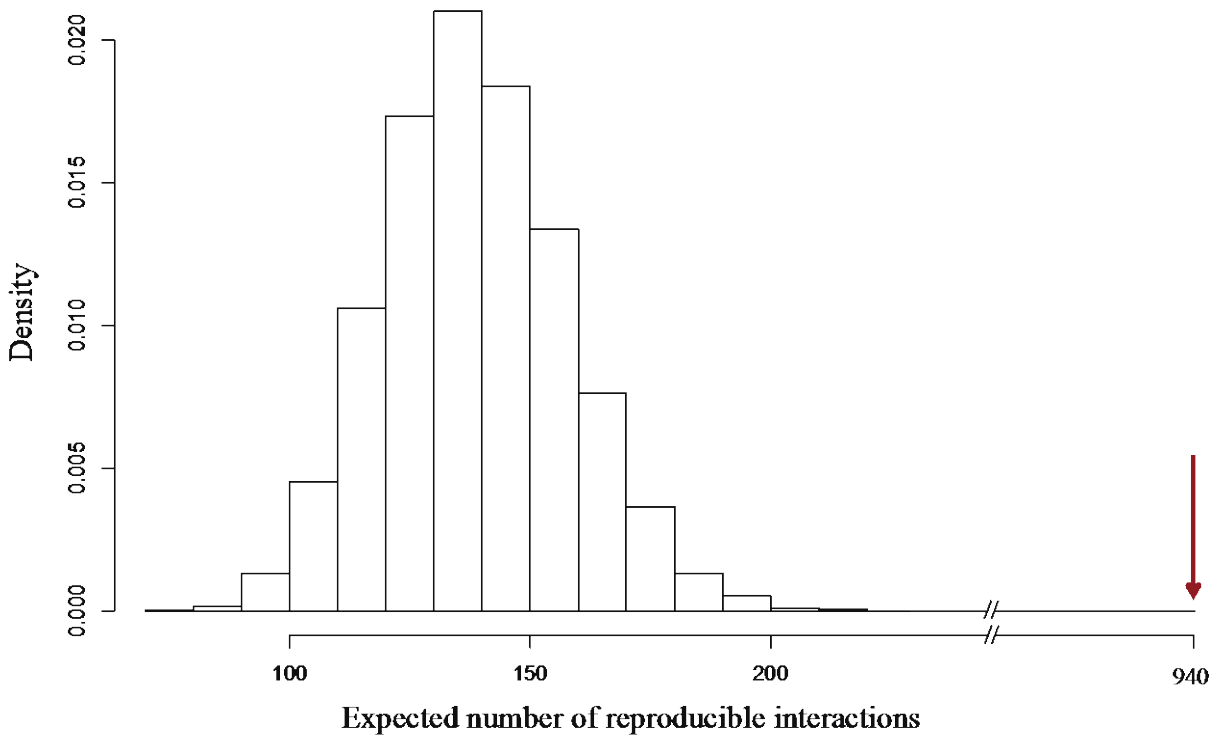


Fig 5.6. Whole cell-GG and mitochondrial-GG networks are highly enriched for common interactions as judged by random graphs.
One hundred different random datasets were generated by shuffling protein labels from each of the whole cell-GG and mitochondrial-GG datasets. Applying the same analysis pipeline as TbCF net, the distribution corresponding to the number of reproducible interactions (FDR ≤0.05 in one dataset and p-value ≤0.05 in the other) between each possible combination of random datasets (10,000 combinations in total) were observed. As illustrated, the expected number of reproducible interactions by chance is 137. However, the whole cell-GG and mitochondrial-GG networks share 940 reproducible interactions (the red arrow) with each other.

Consistency of the predicted network with previous findings

To test the validity of TbCF net, we first examined the topological properties of the constructed network (Fig 5.7.a). In agreement with protein interaction networks of model organisms, TbCF net has scales-free architecture (Barabasi 2009); i.e., while most proteins interact with a small number of proteins, some of them (known as hubs) are highly linked to the other proteins (Fig 5.7.b). Moreover, as shown in Fig 5.7.c, each protein typically can be reached from every other protein by a small traverse in the network (i.e., small path length), as expected from a network with a small-world property (Lizier et al. 2011). Additionally, we observed that except for ribosomal proteins which make a large, densely intra-connected module in the TbCF net, the

topological coefficient also decreases with the number of neighbors (Fig 5.7.d), reflecting that the number of common neighbors for hub proteins compared with the other proteins in the network is relatively low. This feature indicates that highly interacting proteins are not sporadically connected to each other.

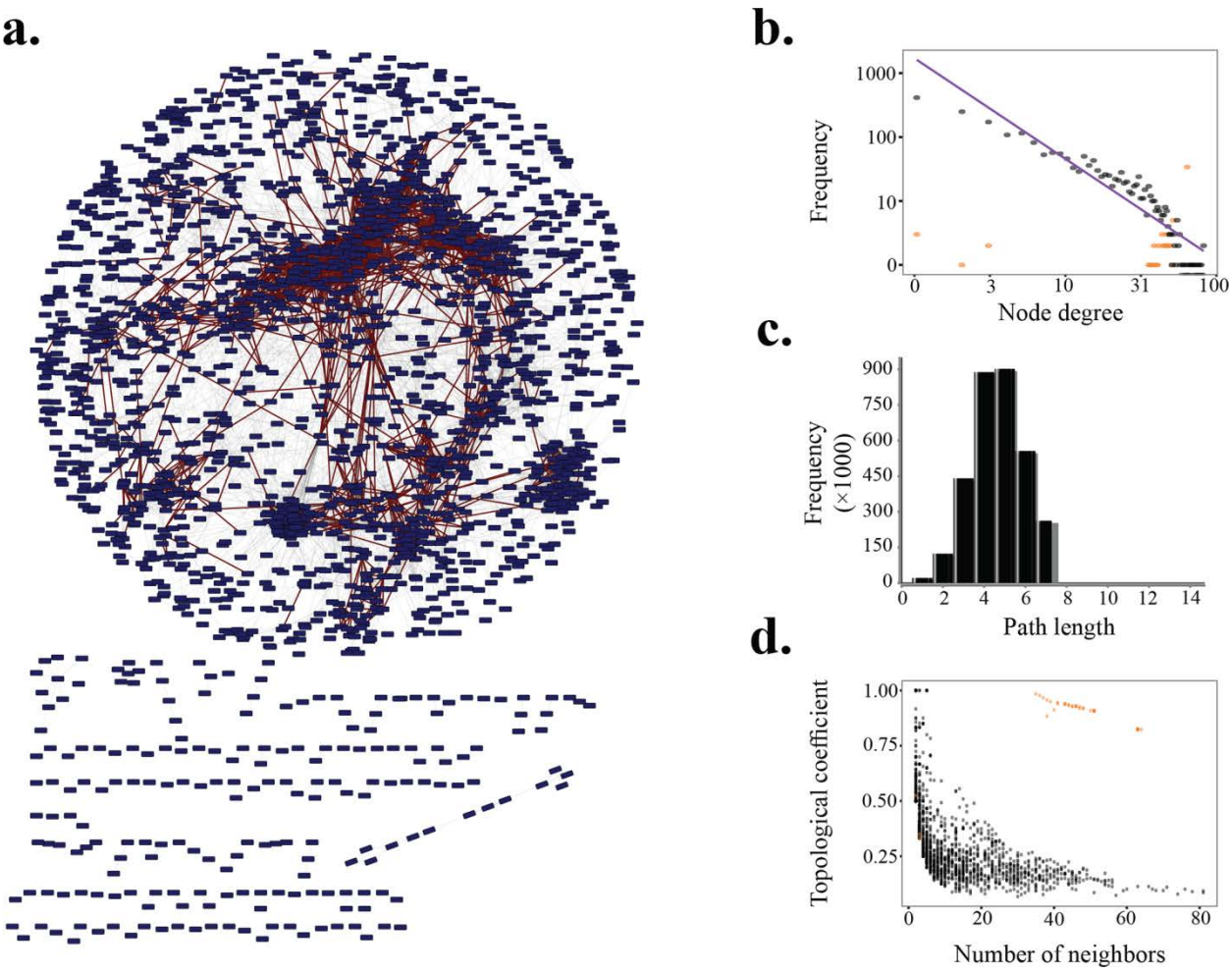


Fig 5.7. Topological characteristics of TbCF net.
a) Representation of TbCF net. Red interactions correspond to those that were reproducibly identified in both GG and IEX fractionation experiments. b) Distribution of number of nodes as a function of number of node neighbors (i.e., node degree) in logarithmic scale. Interactions between ribosomal and non-ribosomal proteins were analyzed separately and were distinguished in the graph by yellow and black colors, respectively. c) Distribution of shortest path lengths in the TbCF net. In accordance with the small world property of biological networks, the path length is usually small with a unimodal peak at five. To avoid bias, ribosomal proteins were discarded for this analysis. d) Distribution of the topological coefficient of nodes as a function of node degree. Similar to hub degree analysis, interactions between ribosomal and non-ribosomal proteins were analyzed separately and were distinguished in the graph by yellow and black colors, respectively.

We next assessed different biological features that can be expected from a protein interaction network. To this end, we first examined whether interacting proteins in the network tend to be involved in the same biological process. As shown in Fig 5.8.a, we observed that higher similarity scores of fractionation patterns consistently led to an increased probability that two proteins are involved in the same biological process, as judged by gene ontology-biological process (GO-BP) analysis. Importantly, this analysis suggested the higher precision of $TbCF_{OR}$ net compared to the TbCF net. KEGG pathway (Kanehisa et al. 2014) analysis also led to similar results, i.e., the chance of two proteins participating in the same KEGG pathway increases by having a higher similarity score (data not shown). Although proteins in the same KEGG pathway do not necessarily interact with each other as they can have functional rather than physical associations, protein complexes are involved in some KEGG pathways such as ribosome and proteasome. Importantly, we found that KEGG pathway interactions that were also supported by the TbCF net had significantly higher (p-value <3.5E-33, Wilcoxon-Mann-Whitney rank sum test) co-localization scores compared with the other KEGG interactions (Fig 5.8.b), suggesting the physical nature of associations for the captured interactions. In a more stringent analysis, we next investigated whether or not proteins involved in a same KEGG pathway tend to form a significant module in the TbCF net. To this end, we used a score, termed modulation score (Khosravi et al. 2014), that assess the density of connections among a set of pre-specified proteins; e.g., proteins with a shared KEGG attribute (See methods for details). This analysis revealed that members of 15 KEGG pathways form densely connected modules (p-value<0.05) in the TbCF net (Fig 5.8.c). Interestingly, we observed a significant modulation score for many pathways that are well-known to involve protein complexes (ribosomes, proteasome, RNA transport, oxidative phosphorylation, etc.). The same analysis on the $TbCF_{OR}$ net indicated that this network is not biased towards a specific process and most KEGG pathways remain or even become significantly connected in the $TbCF_{OR}$ net. However, as discussed earlier, $TbCF_{OR}$ net were depleted for those pathways that involved less stable and salt sensitive interactions (Fig 5.8.c). We also analyzed the TbCF net in terms of gene expression responses. In model organisms, it has been shown that transcriptional co-regulation of proteins plays a major role in the efficient control of the cell homeostasis in different environments (Stuart et al. 2003). It has also been shown in *T. brucei* that functionally-related proteins tend to be co-expressed with each other [56]. Indeed, gene expression analysis indicated that interacting proteins in the TbCF net

show significantly high (p-value < 2.2E-16, Wilcoxon-Mann-Whitney rank sum test) co-expression trends in different life stages and also during the differentiation process (Fig 5.8.d). Consistent with GO-BP semantic similarity analysis, co-expression analysis also indicated a significantly higher (p-value < 2E-07, Wilcoxon-Mann-Whitney rank sum test) co-expression trend for interactions of $TbCF_{OR}$ net compared to those of TbCF net, suggesting a higher precision of the former network.

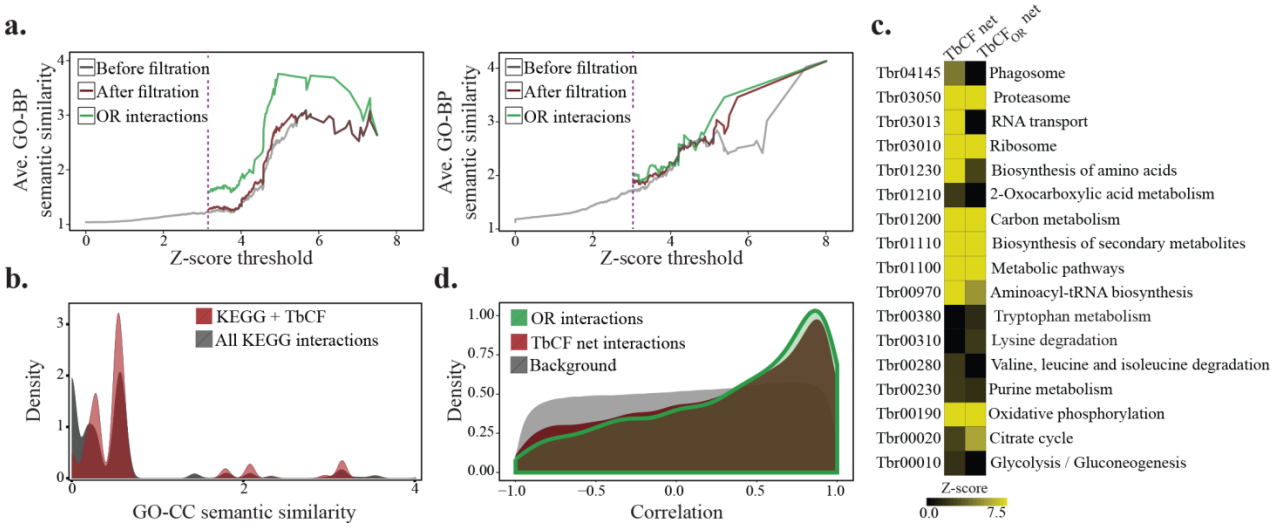


Fig 5.8. Biological validation of the constructed co-fractionation network.
a) The average GO-BP semantic similarity was calculated across different z-score cut-off thresholds for whole cell-GG and mitochondrial-GG experiments, separately. The purple line highlights the co-sedimentation cut-off threshold corresponding to a false discovery rate of 0.05. To examine the applied filtration steps (elimination of noise sensitive, unshared-peak, early sedimenting, and non-reproducible interactions), we applied the same analysis to whole cell-GG and mitochondrial-GG networks before and after the filtration steps. As shown, the employed filtration steps have led to an increase in precision in both networks. However, the reproducible interactions constantly had higher similarity compared to the non-reproducible interactions. b) GO-CC semantic similarity distribution of all possible protein pairs with the same KEGG pathway was compared with the distribution of those KEGG interactions that were also supported by TbCF net. As illustrated, protein pairs in the latter group have significantly higher (p-value <3.5e-33, Wilcoxon-Mann-Whitney rank sum test) co-localization scores compared to the all functional interactions supported by KEGG. c) KEGG pathway modulation analysis of TbCF and $TbCF_{OR}$ networks. The graph is pseudo-colored where the yellow color indicates significant connections (p-value <0.05) among proteins in the corresponding KEGG pathway. d) Distribution of observed co-expression similarities among all proteins detected in our study (background), protein pairs interacting in TbCF net, and those in $TbCF_{OR}$ net. Gene expression data were extracted from (Jensen et al. 2009). OR interactions: Orthogonally reproducible interactions; KEGG + TbCF: KEGG interactions supported by TbCF network.

Extracting the high-confidence subset of TbCF net

To estimate precision of TbCF net, we focused on twenty complexes that have been identified experimentally in *T. brucei*, composed of more than 200 proteins. This analysis demonstrated a precision of 34% in TbCF net with a recall of 80% on proteins and 17% on interactions, which is comparable to previous high-throughput studies of protein interactions based on biochemical fractionation (Havugimana et al. 2012; Wan et al. 2015) (see methods for details). However, the estimated precision can be an underestimate as in many cases, subunits of the complexes had been partially identified because of application of stringent conditions. Importantly, consistent with the expected higher precision of $TbCF_{OR}$ net as judged by GO-BP and co-expression analysis, the same analysis on $TbCF_{OR}$ net with literature interactions suggested a precision of 59% (~2-fold increase in precision). As mentioned in the introduction, the integration of different data sources is highly recommended to reduce false positive results from an interactome. However, due to the lack of a global, unbiased protein interaction map for *T. brucei*, we were not able to utilize the state-of-art machine learning approaches to systematically integrate additional data sources. To experimentally examine the extent to which the precision of TbCF network can be further improved by other orthogonal information, we performed TAP-TAG TEV-elution of RNA editing ligase 1 (REL1) protein. We selected this protein because it is a subunit of a well-studied *T. brucei* complex, the core editosome. TEV-elution is a low stringency condition in which transient interactors along with many contaminants can co-purify with the tagged protein. TEV-elution of REL1 protein led to the co-purification of 83 proteins as putative interacting partners. We extracted the sub-network from TbCF net that was restricted to the co-purified proteins. Consistent with that, REL1 was directly connected to its known interacting partners, while well separated from other contaminant proteins in the network (Fig 5.9.a). It should be noted that since only a small subset (i.e., five proteins) of previously known interacting partners of REL1 protein were co-purified in the pull down experiment, and the constructed sub-network was restricted to the identified proteins in that experiment, the other previously known interacting proteins that were also connected to REL1 in TbCF net are not represented in the sub-network, indicating the importance of queried proteins on obtaining a comprehensive sub-network. To further assess the TbCF net, we considered six separate pull down experiments of aminoacyl-tRNA synthetase (aaRS) proteins in different *T. brucei* life stages (Cestari et al. 2013). In these pull down experiments, 262 proteins were co-purified as

candidate proteins involved in *T. brucei* tRNA-synthesis with a minimum overall mass spectral count of two. Following the same procedure as that used for REL1, the sub-network of these proteins suggested the existence of several distinct complexes among the proteins co-purified with aaRS proteins (Fig 5.9.b). Interestingly, one of these complexes was highly enriched for proteins involved in tRNA-synthesis including the recently identified members of MARS complex (Cestari et al. 2013) and two additional hypothetical proteins. Therefore, the TbCF net was able to successfully distinguish contaminant proteins from direct interactors in both cases. Integration of various TAP-TAG experiments with the TbCF net also suggested that the chance of retrieving a high false positive rate with a targeted search, i.e., a list of putatively interacting proteins, is very low. In our next attempt, we considered the functional protein interaction network of *T. brucei* deposited in the STRING database (Franceschini et al. 2013). In the STRING database, most of available functional association data for *T. brucei* is inferred based on the indirect approaches (such as text mining and co-expression) rather than direct experimental evidence. We extracted the STRING network using a medium confidence level threshold (the default threshold set by STRING) for the proteins present in the TbCF net. The retained STRING network was densely connected with 19,119 interactions among 1,402 proteins. Next, we derived a secondary network composed of the interactions that were supported by both STRING and TbCF net, termed $TbCF_{STRING}$ net. The $TbCF_{STRING}$ net was composed of 2,413 interactions among 449 proteins (that included 13% of interactions and 32% of proteins in the primary STRING network). To assess the validity of the network, the $TbCF_{STRING}$ net was clustered using the clusterOne algorithm (Nepusz et al. 2012). As shown in Fig 5.9.c, many previously known protein complexes were recovered by this approach, indicating the high accuracy of the $TbCF_{STRING}$ net. This result indicated that integration of our co-fractionation network with other independent sources (like AP and STRING in this case) leads to the elimination of false positive interactions from both sources.

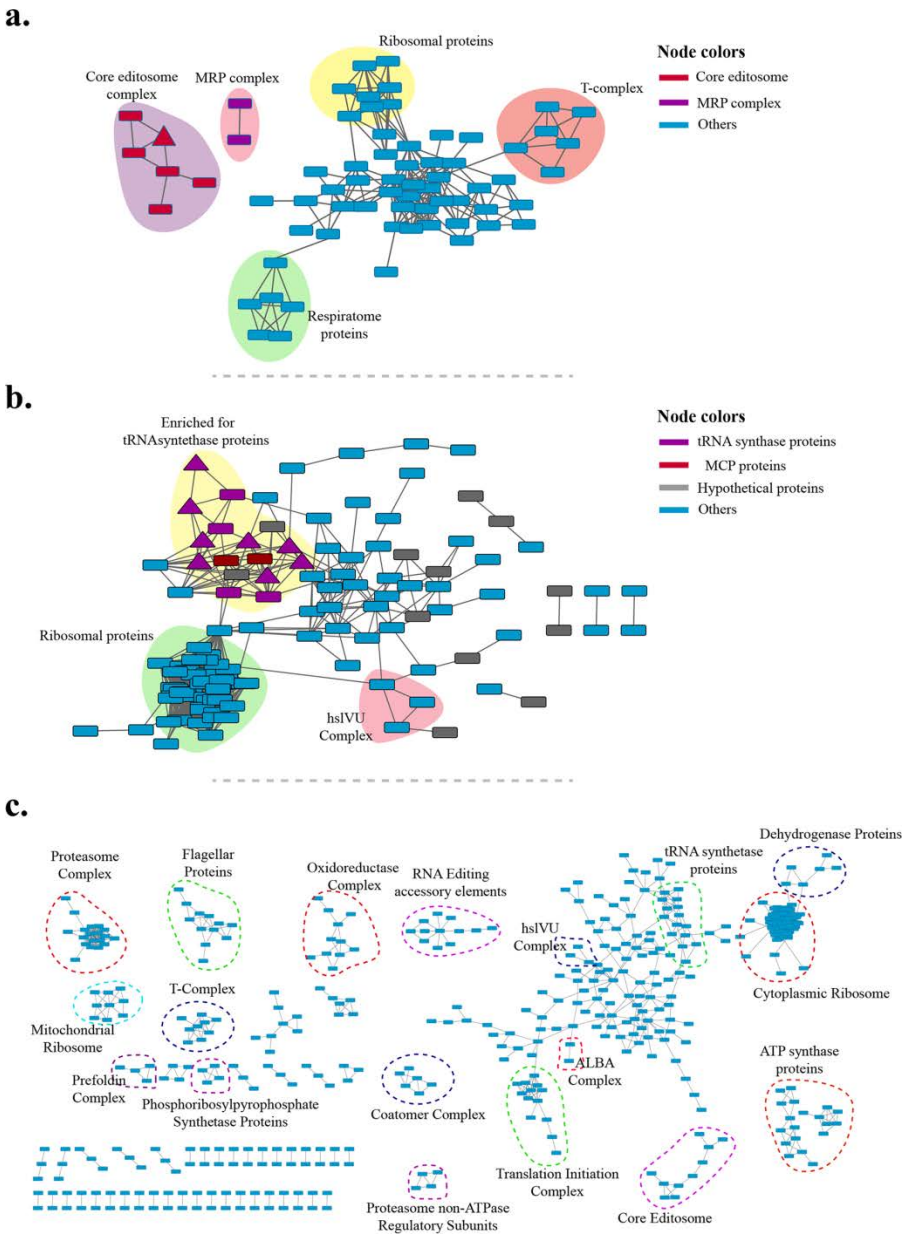


Fig 5.9. Integration of TbCF net with other highly contaminated resources.
a) The sub-network of TbCF net that was restricted to 83 proteins co-purified with REL1 protein in the TEV-elution experiment. The REL1 protein is represented as a triangular node. b) The tRNA synthase sub-network, which was constructed by restricting TbCF net to 262 proteins that co-purified with tRNA synthase proteins in six independent pull-down experiments. The proteins identified in these pull-down experiments are represented as rectangular nodes. Interestingly, TbCF net suggested that some of the other tRNA synthase proteins, although not detected in these pull-down experiments, were significantly co-fractionated with the subunits of MARS complex. These additional tRNA synthase proteins are represented as triangular nodes. The list of *T. brucei* tRNA synthase proteins was extracted from (Berriman et al. 2005). c) Representation of the $TbCF_{STRING}$ net which was constructed by considering the common sub-network between TbCF net and the STRING network.

Based on the findings in the previous step, we partitioned the interactions in the TbCF net into two parts; those with high confidence, $TbCF_{HC}$, and those with no external evidence, $TbCF_{NE}$. As schematically shown in Fig 5.1.a, the high confidence group was composed of interactions that were supported by at least one of the orthogonal resources including KEGG pathways, the STRING database, interlog-mapping, extensive literature searches, and orthogonal reproducibility (see methods for details). As discussed earlier, interaction data in each of these orthogonal resources (e.g. KEGG, STRING, AP, etc.) suffers from false positives and also, for some sources, does not imply physical interaction among protein pairs. However, we expected to observe elimination of the false positive interactions by integration of these data with our fractionation-derived network. $TbCF_{HC}$ network was composed of 4,726 interactions among 866 proteins, encompassing 34% of the total interactions in TbCF net. As represented in Fig 5.10, analysis of $TbCF_{HC}$ net indicated its improvement over TbCF net in terms of GO-BP (p-value <2E-08), GO-CC (p-value <6E-07), and co-expression (p-value <5E-117). As judged by the analysis of the same twenty protein complexes discussed above, the estimated precision for $TbCF_{HC}$ net was 80% and 68% without and with the exclusion of literature-derived data from the network, respectively. The estimated precisions for both $TbCF_{OR}$ and $TbCF_{HC}$ networks suggest that integration of TbCF net with other orthogonal resources leads to the overall false discovery rate of less than 40%.

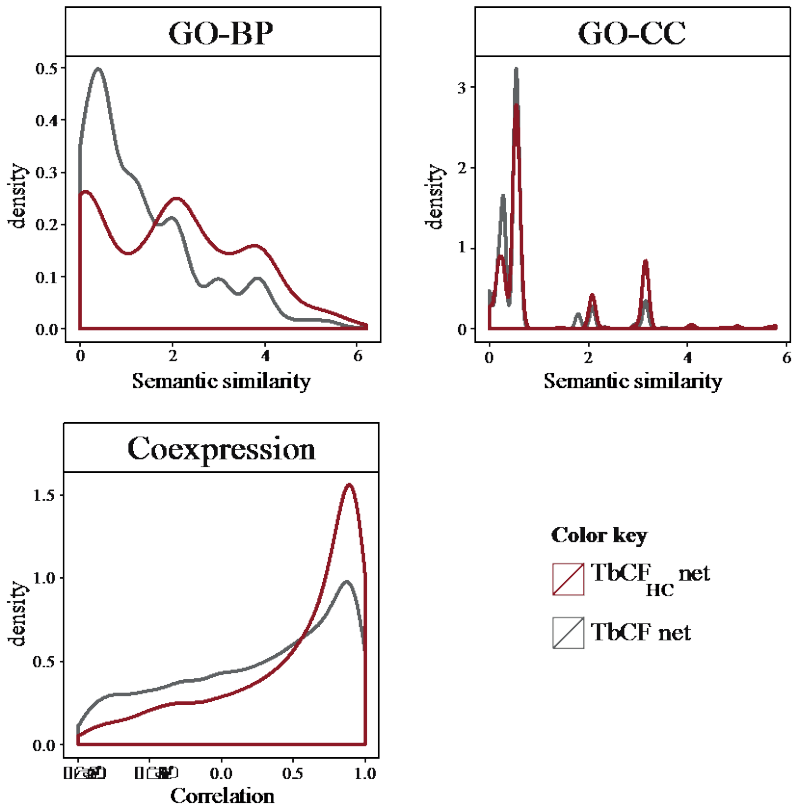


Fig 5.10. Comparing the biological characteristics of $TbCF_{HC}$ net with those of TbCF net.
Comparison of GO-BP, GO-CC, and Co-expression distributions indicates $TbCF_{HC}$ is significantly improved over TbCF net. For co-expression analysis, data from (Jensen et al. 2009) were used.

Clustering $TbCF_{HC}$ net using ClusterOne algorithm led to the prediction of 128 protein complexes among 716 proteins. The predicted complexes varied in size between three and 70 with the median size of four proteins per complex that is similar to reports from other organisms (Havugimana et al. 2012). Many of the predicted complexes were significantly enriched for subunits of previously known *T. brucei* complexes. For example, as illustrated in Fig 5.11, we were able to successfully identify complexes related to cytoplasmic and mitochondrial ribosomes, proteasome, T-Complex, intraflagellar transport, translation initiation, and mitochondrial RNA editing. To further assess the quality of predicted complexes, we associated them with the available large-scale RNAi screening data for *T. brucei* (Alsford et al. 2011). Because each protein complex act as a functional unit in the cell, it is expected that the essential proteins to be over- or under-represented in a protein complex, depending on the function of the

complex (Hart et al. 2007). Indeed, enrichment analysis of predicted complexes indicated over representation of complexes with enriched or depleted fraction of essential proteins, while complexes with random distribution of essential proteins were under represented (Fig 5.12.a). The predicted essential complexes recapitulated the previous findings for *T. brucei*. As an illustration, the results indicated that the cytoplasmic ribosomal complex (Complex 1), proteasome (Complex 3), and T-complex (Complexes 10 and 11) are essential in all life stages of the parasite (Fig 5.12.b). Conversely, the complex 7 which is highly enriched for intraflagellar proteins (Fig 5.12.b), was not essential in the procyclic life stage of *T. brucei* (Broadhead et al. 2006). Additionally, complexes related to the mitochondrial ribosome (Complex 2 and 8) were depleted for the essential proteins in the bloodstream life stages (Fig 5.12.b).

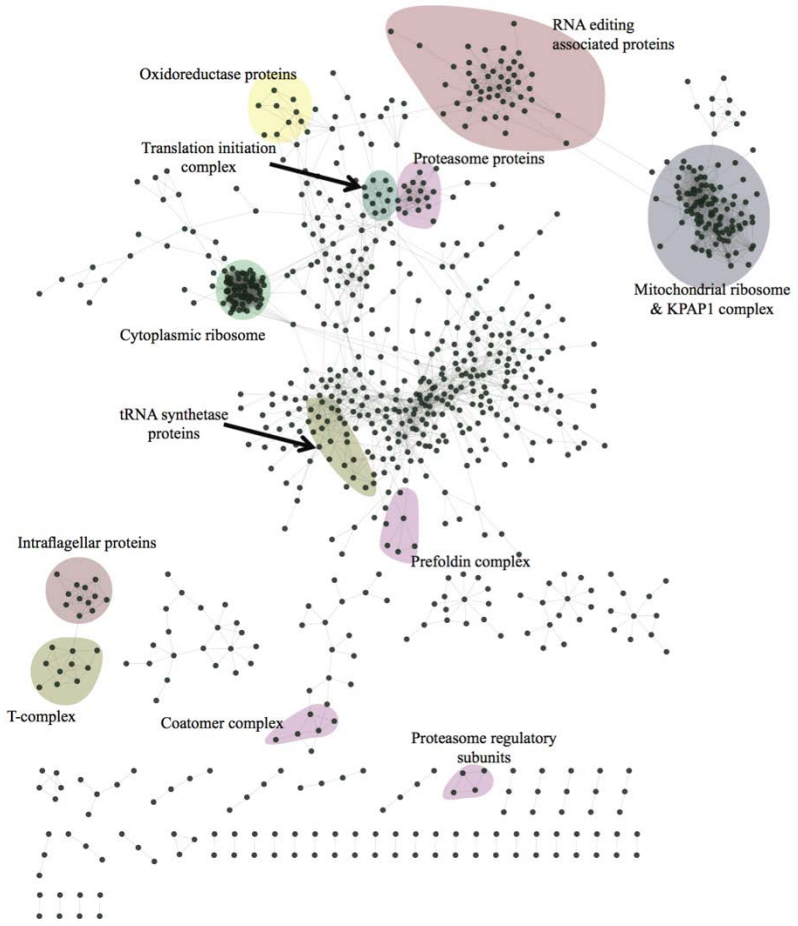


Fig 5.11. Graphical representation of $TbCF_{HC}$ net.
As illustrated, clustering of TbCFHC net led to the recovery of many of previously identified complexes in *T. brucei*. Clustering also predicted some new complexes and assigned new members to the previously characterized complexes.

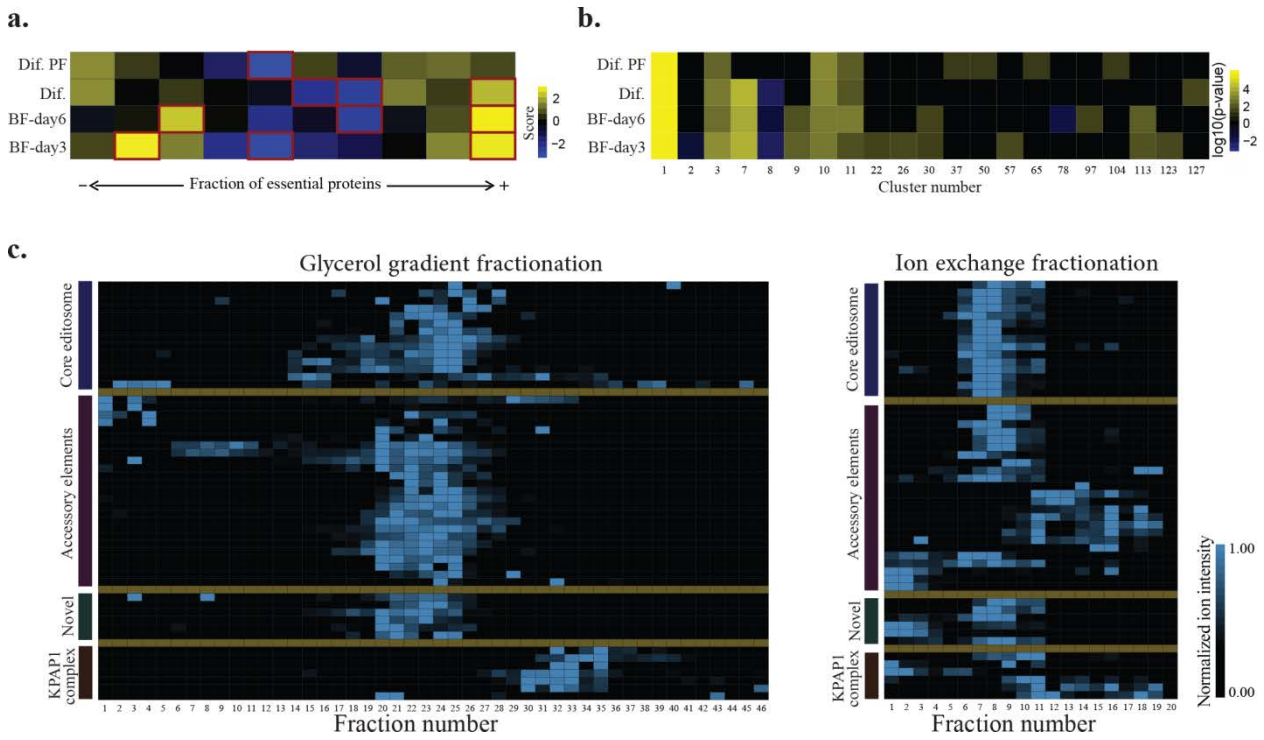


Fig 5.12. Evaluation of predicted complexes based on $TbCF_{HC}$ net.
a) The 128 predicted complexes were classified to ten uniformly spaced bins based on the fraction of essential proteins present in each complex. The number of complexes present in each bin was compared to the expected background distribution using MATLAB kernel smoothing function. The expected background distribution was estimated by applying the same clustering analysis (i.e., ClusterOne algorithm) to 10,000 random networks from TbCF net, generated by shuffling the protein labels, while preserving structural properties. Figure represents the calculated kernel score for each bin, with blue and yellow colors indicating under-representation and over-representation, respectively. Significantly enriched or depleted bins are shown by red borders. As shown, we observe over-representation of $TbCF_{HC}$ net-derived complexes with enriched or depleted fraction of essential genes (corresponding to right and left bins, respectively). However, $TbCF_{HC}$ net-derived complexes with random distribution of essential genes (corresponding to the bins in the middle) were under-represented. The list of essential genes in 4 different life stages of *T. brucei* was extracted from (Alsford et al. 2011). (Dif. PF: differentiated procyclic cells; Dif.: Differentiating cells from bloodstream to procyclic; BF-day6: Representing bloodstream stumpy form; BF-day3: Representing bloodstream short slender form). b) The 128 predicted complexes were examined for possible enrichment of essential proteins in different life stages and during the differentiation process from bloodstream to the procyclic form (see above for abbreviations). The enrichment analysis was performed using Fisher's exact test. At the p-value cut-off threshold of 0.05, the yellow color indicates a significant over-expression and the blue color represents a significant under-representation of essential proteins in a complex. It should be noted that most of predicted complexes had small sizes (median of four subunits per complex). Therefore, the statistical test did not have enough power to detect significant over-/under- representations of essential genes in these small complexes. Correspondence of previously known protein complexes to the reported complexes in this figure: cytoplasmic ribosomal complex (Complex 1), proteasome (Complex 3), and T-complex (Complexes 10 and 11), intraflagellar proteins (Complex 7), mitochondrial ribosome (Complex 2 and 8). c) GG and IEX fractionation patterns for the proteins known/predicted to be involved in the RNA editing and KPAP1 complex. Proteins are categorized in four groups of core editosome, accessory elements, novel proteins, and KPAP1 complex.

Validation of $TbCF_{HC}$ net

Clustering of $TbCF_{HC}$ net also led to the prediction of complex membership for 188 protein groups currently annotated as hypothetical, 350 protein groups with the annotated name as putative, and 635 protein groups lacked experimental GO-BP annotation (evidence codes of EXP, IDA, IPI, IMP, IGI, and IEP according to TriTrypDB v5). To experimentally verify the quality of predictions, we focused on the overlapping complexes that were highly enriched for proteins involved in the Kinetoplastid RNA editing process.

Mitochondrial gene regulation is a highly interesting, yet not fully understood, process in *T. brucei*. In this process, mitochondrial genes are transcribed as polycistronic units. After cleavage, mitochondrial transcripts (mtRNAs) become stabilized by the addition of short tails at their 3' ends (Etheridge et al. 2008). Intriguingly, most of the produced mtRNAs originally do not possess correct open reading frames (ORFs) and require to be edited before the translation. In the editing process, small RNAs, known as guide RNAs (gRNAs), dictate the insertion, and less frequently deletion, of a defined number of uridine nucleotides at pre-specified positions in mtRNAs (Stuart et al. 2005). The edited mtRNAs become marked for the translation by the addition of long A/U tails (Aphasizheva et al. 2011). This highly complicated and intertwined process provides the parasite multiple post-transcriptional regulatory layers over mtRNAs. Mitochondrial post-transcriptional gene regulation is essential for the survival of the parasite in both bloodstream and procyclic life stages, although the precise role of this process is not well understood in the bloodstream form (Rusche et al. 2001; Schnaufer et al. 2001). Experimental evidence confirms differential regulation of some mtRNAs in at least editing step during the parasite's life cycle (Feagin et al. 1987; Read et al. 1992). For example, while the Cytochrome b (Cyb; a subunit of complex III) and cytochrome oxidase subunit II (COII) transcripts are preferentially edited in procyclic form, some subunits of complex I such as NADH-ubiquinone oxidoreductase subunit 8 (ND8) and NADH-ubiquinone oxidoreductase subunit 7 (ND7) are mostly edited in the bloodstream stage. This developmental regulation in RNA editing is coordinated with the activities of the trypanosome mitochondria during its life cycle to allow the adaptation and survival of the organism in changing environmental conditions (Schnaufer et al. 2002). Consistent with the developmental regulation of the RNA editing process, comparative sedimentation analysis of the RNA editing machinery demonstrated that the complexes

associated with the RNA editing machinery of bloodstream and procyclic forms are not identical (Halbig et al. 2004).

The $TbCF_{HC}$ net suggested the physical association of 49 protein groups (50 proteins) with the RNA editing machinery of *T. brucei* in the procyclic life stage. Many of these proteins are well-known to be involved in the RNA editing process. For example, among the 50 predicted proteins, 17 are known to be involved in the core editosome complex (Schnaufer et al. 2010), while 21 function in the MRB1 complex (Ammerman et al. 2012). For detailed analysis of their interactions, we distinguished predicted interactions originated from IEX experiments to those of GG experiment. This analysis indicated that members of core editosome and some of the accessory elements were reproducibly co-fractionated in both approaches. However, about 70% of significantly co-sedimenting protein pairs in the GG experiments were not significantly co-eluted in the IEX experiments. Visual inspection of the fractionation patterns of these 50 proteins confirmed that they mainly co-sedimented together in mitochondrial-GG experiment, but dissociated into different protein clusters in the mitochondrial-IEX experiment (Fig 5.12.c). It should be pointed out that many proteins that are functionally associated with the RNA editing machinery, mediate their functions through binding to the RNA and GG and IEX mitochondrial fractionation experiments were not RNase treated. However, technical differences between the IEX and GG experiments provided a high resolution picture of the distinct complexes involved in the mitochondrial post-transcriptional regulation of *T. brucei*. For example, while members of core editosome were reproducibly co-fractionated in both approaches, comparison of IEX and GG data suggested existence of at least three distinct groups of proteins among the accessory elements. These three groups were co-sedimented with each other and with core editosome in the mitochondrial-GG experiment. However, they failed to co-fractionate in the mitochondrial-IEX experiment possibly due to the increased salt concentration, suggesting RNA-dependent or less stable nature of interactions between these groups. Consistent with previous reports (Ammerman et al. 2012), observed fractionation patterns supported a direct interaction between Tb927.8.8170 and Tb927.11.16860, while suggested RNA-dependent interaction of Tb927.2.6070 with MRB core proteins (i.e., Tb927.5.3010, Tb927.10.11870, and Tb927.10.10130). However, comparison with previous findings on the interactome of proteins related to the RNA editing machinery suggested that the lack of co-elution in the IEX data could be also because of less stable direct protein-protein interactions. For example, although MRB8170 (Tb927.8.8170) and TbRGG2

(Tb927.10.10830) did not co-eluted in the IEX experiment, they were reported to directly interact in Y2H assays (Ammerman et al. 2012; Foda et al. 2012), or AP-based studies (Madina et al. 2011). Importantly, a previous study has demonstrated the TEV co-elution of the two proteins in a RNA-enhanced manner (Kafkova et al. 2012). Hence, the lack of co-elution in the IEX fractionation experiment for MRB8170 and TbRGG2 proteins is likely due to less stable interaction of the two proteins. Moreover, fractionation data indicated that members of KPAP1 polyadenylation complex reproducibly fractionated differently from those of core editosome, and suggested their association with mitochondrial ribosome, which is consistent with their functional role that couple the mitochondrial editing with the translation process (Aphasizheva et al. 2011). Also, the fractionation patterns successfully captured interactions of KPAP1 protein (Tb927.11.7960) with both editing and ribosomal complexes (Fig 5.12.c), recapitulating the results obtained by mass spectrometric and immunochemical experiments (Aphasizheva et al. 2011). The $TbCF_{HC}$ net also suggested that six new proteins (Tb927.1.3010, Tb927.10.7910, Tb927.1.1730, Tb927.10.5830, Tb927.6.1200, and Tb927.10.1730) play a role in the mitochondrial post-transcriptional gene regulation. Fractionation patterns of GG experiment for these candidates indicated their co-sedimentation with core editosome and accessory elements, but separation from subunits of KPAP1 polyadenylation complex. Importantly, IEX fractionation patterns suggested that two of these candidates (Tb927.6.1200 and Tb927.10.1730) strikingly co-fractionated with members of the core-editosome and some accessory elements, but the co-elution of the other four proteins were, in varying degrees, sensitive to the presence of salt with Tb927.10.7910 being the most salt sensitive one.

Examining the localization pattern for three of our candidate proteins (Tb927.10.7910, Tb927.10.1730, and Tb927.1.1730) clearly confirmed their exclusive presence in the mitochondria of the parasite. We also performed immunoprecipitation experiments on tagged versions of these three proteins. These experiments verified their interactions with TbRGG2 and MRB8170 proteins. Consistent with the observed IEX fractionation patterns, we found that the interaction of all three candidates with MRB8170 was abolished following the RNase treatment (Fig 5.13.a). Likewise, we found that only Tb927.10.1730 remains still bound to TbRGG2 after RNase treatment (Fig 5.13.a). Consistent with this work, another study reported four out of our six candidates (including Tb927.1.3010, Tb927.1.1730, Tb927.6.1200, Tb927.10.1730) form a novel complex involved in the post-transcriptional regulation of mtRNAs, termed

polyadenylation mediator complex, confirming the predictions on these four proteins (Aphasizheva et al. 2014). To further assess the role of the candidate proteins in the RNA editing process, we performed tetracycline (Tet)-inducible RNA interference (RNAi) knockdown of Tb927.10.7910 in the procyclic form of *T. brucei*. RNAi induction for this gene led to growth-defect phenotype, reflecting its essential role in normal growth of the parasite in the procyclic life stage (Fig 5.13.b). Follow up quantitative RT-PCR verified the knockdown of the candidate transcript compared with the control, uninduced cells (Fig 5.13.c). We next quantified the relative changes in mitochondrial-encoded pre-edited, edited, and never-edited transcripts in the RNAi-knock down background. We also considered three precursor RNAs to examine whether this protein play a role in the precursor RNA processing. This experiment indicated that Tb927.10.7910 affect the RNA editing process as judged by the accumulation or reduction of pre-edited or edited transcripts for different target RNAs (Fig 5.13.c). Interestingly, our results suggested that Tb927.10.7910 affect the editing process of the Cyb transcript (i.e. upregulation of pre-edited mRNA and down-regulation of the edited mRNA). The knock down of Tb927.10.7910 also led to the down regulation of RPS12 edited transcript, and also accumulation of MURF2 pre-edited as well as COIII and A6 edited mtRNAs, suggesting the multiple functionality of Tb927.10.7910 in the mitochondrial post transcriptional regulatory network of *T. brucei*. Consistently, a previous study has suggested the essentiality of Tb927.10.7910 protein in the bloodstream life stage of the parasite (Alsford et al. 2011).

Conclusions

We have presented a systematic study of protein complexes in *T. brucei* using two complementary biochemical fractionation approaches. Our results led to the assignment of many previously uncharacterized proteins to complexes. The quality of predictions was verified by independent follow up experiments on newly characterized proteins associated with RNA editing machinery. Interestingly, we found that at least five out of six predictions of $TbCF_{HC}$ net are truly associated with the RNA editing machinery, and also that one of them preferentially affect the editing process of Cyb transcript, a developmentally regulated mitochondrial mRNA. Further experiments are required to clarify the roles of these proteins in the mitochondrial gene regulatory network of *T. brucei*.

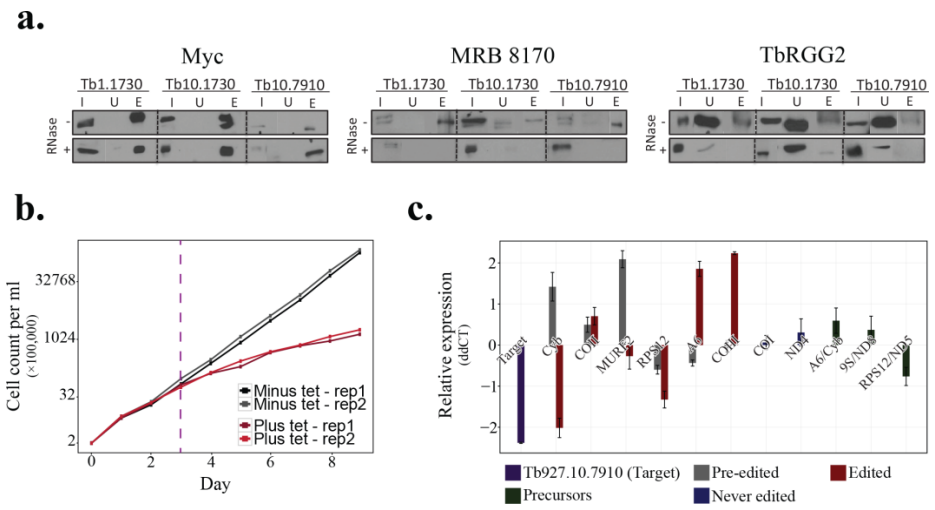


Fig 5.13. Experimental validation of the candidate proteins.
a) Tb927.1.1730 (Tb1.1730), Tb927.10.1730 (Tb10.1730) and Tb927.10.7910 (Tb10.7910) proteins possess interactions with TbRGG2 sub complex.Immunoprecipitation of cmyc-Tb927.1.1730, cmyc-Tb927.10.1730, and cmyc-Tb927.10.7910 from mitochondrial extracts either RNase inhibited (-RNase) or RNase treated (+RNase). Proteins from input (I), unbound (U), and eluate (E) were electrophoresed on 10% SDS-polyacrylamide gel and the blot was probed with specific antibodies against myc (to detect cmyc tag), MRB8170 and TbRGG2. b) Growth curve for Tb927.10.7910 RNAi-knock down experiment. The dashed purple line indicates the selected day (day-3) for collecting RNA sample to examine the knock down effect on mitochondrial transcripts. c) The RNA-editing machinery works on pre-edited mRNA substrates to make edited mRNA; therefore, interfering with this machinery is expected to lead to accumulation of pre-edited and down-regulation of edited mRNAs. We observed this phenotype for the Cyb transcript in knock-down experiment of Tb927.10.7910. The fact that some other edited and pre-edited mtRNAs are also affected can be suggesting that the protein plays multiple functions in the mitochondria gene regulation network of *T. brucei*. The knock down experiment was performed with two independent biological replicates and three technical replicates (six replicates in total).

Despite the employment of stringent filtration criteria, our preliminary network (TbCF net) contains false positive interactions (estimated precision of 34%) that arose from limitations of the employed fractionation approaches. However, integration of results from two different fractionation approaches led to significant boosting of the precision (~2-fold), reflecting the importance of data integration for accurate predictions. Due to the lack of an unbiased, high-confidence, and large-enough protein complex map for *T. brucei*, we were not able to apply sophisticated machine-learning approaches to increase the precision of TbCF net by incorporating other data sources such as transcriptome. The KEGG interactions imply functional rather than physical associations. The interactions deposited in the STRING database were mostly inferred based on the indirect evidences that support functional associations. The literature-derived network was highly biased towards specific protein complexes and also

contaminated with non-specific interactors. According to the primary literature-derived network, for example, there were ~360 co-purified protein groups with known members of RNA editing machinery. Of these, 242 protein groups were detected in our experiments. However, the TbCF net predicted strikingly similar co-fractionation pattern for only 49 of these protein groups, reflecting the high-contamination rate of literature-derived network. By stringent filtration of data, we focused on the high confidence sub-network of TbCF net ($TbCF_{HC}$ net), composed of interactions whose existences were supported by at least one other independent resource. The $TbCF_{HC}$ net refines and extracts new information from previous data and computational predictions on the interactome of *T. brucei*. However, this additional step can lead to the loss of information on proteins for which these types of evidences were not available. Therefore, we can expect that the number of high-confidence interactions will increase with the availability of more experimental data on *T. brucei* protein-protein interactions. Consistent with this, we have found that integration of results from other experiments (e.g. AP or immunoprecipitation) with TbCF net (which is a mixture of high- and low-confidence interactions) leads to the elimination of false positive results from both sources. Importantly, our analysis suggested that integration of TbCF net with other orthogonal resources leads to the overall false discovery rate of less than 40%. This finding has a great impact since most protein interactions in *T. brucei* are inferred by applying stringent conditions at the expense of an increased false negative rate, i.e., losing transient interactions. Our results suggest that this criterion can be relaxed by considering TbCF net as an orthogonal validation resource.

Mass spectrometry-based experiments are known to have limitations in the detection of low abundance proteins (including many regulator proteins), as they become masked in the sample by more abundant proteins. This issue was partially addressed in our approach by fractionating cell-extracts before mass spectrometry. Although comparable with previous SILAC experiments, our experiments cover 42% of proteins associated with expressed mRNAs in the procyclic stage (Kolev et al. 2010). Despite the fact that some of these transcripts can be translationally silent, the fraction of identified proteins is still potentially low. Our fractionation experiments demonstrated that enrichment for specific cellular compartments offers a viable solution to this issue. Hence, a greater depth could be obtained by performing similar experiments for other subcellular compartments of the cell such as the nucleus or mRNA enriched extracts. Moreover, our analysis clearly indicated an under-representation of membrane proteins in the mass

spectrometry data that stems from the employed experimental procedure. Recovery of this class of proteins could likely be increased by use of detergents that can solubilize various membrane associated complexes and protein sub-domains.

Materials and methods related to the analysis section

Construction of primary co-fractionation networks

Although the stoichiometric ratio among proteins that are involved in one complex is expected to be linear, this relationship can be more complicated for those participating in several complexes (As an example, see the fractionation patterns for proteins related to the RNA editing process presented in the validation section). For stringent analysis of data, we used a previously developed mutual information-based approach, termed context likelihood of relatedness (CLR) (Faith et al. 2007), to infer pairwise interactions among the proteins based on the observed fractionation patterns. In contrast to association measures such as Pearson correlation coefficient, mutual information is a measure of association that does not have strong prior assumption about the association type (Steuer et al. 2002). After calculating fractionation pattern similarity scores for each possible interaction using mutual information, CLR determines significant pairwise interactions by comparing the similarity score of each protein pair to a background joint distribution, which was obtained by the calculation of similarities for all possible interactions that each of the interacting proteins can be involved in. To construct the co-fractionation networks, we set the false discovery rate cut-off threshold at 0.05. False discovery rates were calculated using the provided functions in the CLR package (Faith et al. 2007). However, since mutual information does not discriminate the type of association (positive or negative), we considered only those interactions as valid that their fractionation patterns were non-negatively correlated with each other as judged by Pearson correlation coefficient. Comparison of results for the mutual information-based version of CLR with the correlation-based version of this algorithm as well as conventional Pearson correlation coefficient-based method indicated that the former is more consistent with the previous knowledge on *T. brucei* protein complexes.

To assess whether the observed co-fractionation patterns for low abundant proteins is significant or unreliable, we generated 100 noisy datasets from each fractionation experiments. Expecting a

uniform distribution of noises in the dataset, added noise to each cell in a dataset were modeled by Poisson distribution with the lambda equal to the value of the cell plus the noise term of [the lowest identified ion intensity in the dataset]/[No. of fractions in the dataset]. Those protein pairs that were significantly co-fractionated (FDR ≤0.05) in the original dataset, but lost their significant co-fractionation (p-value >0.05) in at least 10% of noisy datasets were discarded from the corresponding co-fractionation network.

Modulation score

Modulation score examines the density of connections among a pre-specified group of nodes (e.g., proteins) in a given-network with the interaction density that is expected to occur by chance (Khosravi et al. 2014). The density of interactions for a group of nodes is defined as the number of within-group interactions divided by the total number of interactions of the members of the group. This ratio is then compared to the density distribution generated over random groups with the same number of nodes in the network. In the original implementation of the modulation score, the density distribution was estimated by normal distribution assumption (Khosravi et al. 2014); however, the distribution may not necessarily be normal. Therefore, we developed a modified version of the algorithm that uses a kernel smoothing function to estimate the distribution.

Curation of high-confidence network

To find high-confidence interactions, we considered following available orthogonal resources that contain protein-protein interaction data:

STRING database

Interactions restricted to the proteins identified in this study with a medium confidence score (default setting in the STRING) were extracted from the database (version 10).

KEGG pathway

All pathways related to *T. brucei*-associated proteins were extracted from the KEGG database.

Interlog mapping

Although trypanosomatid organisms are highly diverged, they share some common processes and, consequently, common protein complexes with other well-studied eukaryotes. To map these conserved complexes, we first transferred protein interaction data from a highly conserved interactome of eukaryotic cells (see below) to *T. brucei* proteins. To remove those interactions that are not conserved in *T. brucei*, only the transferred interactions whose existence was also supported by our fractionation network (TbCF net) were considered as valid. To determine a highly conserved protein interaction map of eukaryotic organisms, we extracted the high confidence interaction networks (interaction scores above 0.9) of human and yeast from the STRING database. Using the orthologous groups defined by the InParanoid database (Ostlund et al. 2010), we extracted a sub-network that is common between these two networks. Considering the large evolutionary distance between yeast and humans, the created sub-network was highly enriched for the basic processes that are vital for eukaryotic cells. Next, the sub-network was mapped to *T. brucei* proteins based on the InParanoid database.

Literature search

Extensive literature searches were performed to find published interactions from small-scale studies of trypanosomatid proteins. In total, 24 studies constituting 81 experiments (including AP, immunoprecipitation, and Y2H experiment types) were considered (Only studies/experiments that provide interaction information for at least two proteins identified in this study was considered.). To construct interaction maps for experiments that contained protein complex information (e.g., AP and immunopercipitation), we considered a matrix model which assumes any two co-purified proteins can be directly connected to each other (von Mering et al. 2002). By restricting the network to proteins identified in this study, the preliminary literature-based network consisted of 13,307 interactions, containing both genuine and false positive interactions. Of these, 1421 interactions whose existence was also supported by the TbCF net, were considered as high-confidence.

Orthogonal reproducibility

Our analysis has suggested a high precision (estimated precision rate of 59%) for the orthogonally reproducible part of TbCF network ($TbCF_{OR}$ net), the sub-network of TbCF net that were reproducible in both glycerol gradient and ion exchange high performance liquid chromatography experiments. Therefore, we considered as high confidence those protein pairs

that were present in the $TbCF_{OR}$ net, but there were not any orthogonal evidences about the interacting partners of at least one of the proteins in any of the four orthogonal evidences mentioned above.

Estimation of precision and recall of the networks

To estimate the precision (percentage of interactions that are true) and recall (the fraction of overall true interactions that are present in the networks) of the constructed networks, benefiting from our extensive literature search, we created a gold standard set of >200 proteins in 20 distinct protein complexes, allowing us to evaluate our network against current literature. It should be noted that some of these complexes are partially identified and there might be some subunits that are still uncharacterized. We also ignored putative subunits that were suggested solely based on a single experiment like pull down without further experimental characterizations and/or verifications. The interactions between proteins in the same complex were defined as the positive set. Negative set were defined as interactions between proteins of different complexes. For more stringent analysis, we added to the negative set the other interactions that the subunits of the 20 complexes had with the rest of proteins in the network, unless they had external experimental evidences (e.g., pull down). Clearly, the defined negative set can contain some of genuine interactions that were not identified previously, leading to an underestimation of the calculated precision for the networks. Precision was defined as the number of interactions in a network that are in the positive set, divided by the total number of interactions in the network that belong to either positive or negative sets. For each network, two types of recalls were defined: 1) Recall for proteins: defined as the number of distinct subunits of the twenty complexes that are present in the network divided by the total number of known subunits identified in our mass spectrometry experiments; 2) Recall for the interactions: defined as the total number of interactions that the identified subunits of the same complex have in the network divided by the maximum number of interactions that they can have theoretically. Since in a network, the subunits of the same complex are often much sparser than the theoretically assumed fully-connected model, the calculated recalls for the interactions are usually small (Havugimana et al. 2012; Wan et al. 2015).

Network visualization and topological analysis

All networks were visualized using Cytoscape, a network visualization tool for Genome Space workflows (Saito et al. 2012). NetworkAnalyzer, a plugin of Cytoscape, was used for the topological analysis of networks, including: node-degree distribution, shortest path-length distribution, and topological coefficient distribution (Assenov et al. 2008).

Measurement of semantic similarity between gene ontology terms

To examine the gene ontology-biological process similarity of interacting genes in the constructed networks, we used Resnik's approach to quantify semantic similarity between gene ontology terms. Semantic similarities between GO categories of biological processes and cellular compartments were calculated for each interacting protein pair present in a network, if both proteins were annotated in the uniprot database (Barrell et al. 2009). When calculating semantic similarities, we ignored those terms with evidence codes of NR (Not recorded), ND (No biological data available), and IEA (Inferred from Electronic Annotation). The GO-BP sematic similarities were calculated by GOssTO tool using default parameters (Caniza et al. 2014).

Statistical analysis

MATLAB R2014b software (The MathWorks Inc., Natick, MA) was used for the CLR score false discovery rate estimation and kernel-based p-value estimation. C# programing language was used for the calculation of the KEGG pathways modulation scores as well as construction of random networks. Other statistical analysis was performed using the R programing language.

Chapter 6

TrypsNetDB: an integrated framework for the functional characterization of trypanosomatid proteins

Genome sequencing of multiple trypanosomatid parasites has revealed that while significant sequence similarity exists at the intra-family level, they are highly diverged from other eukaryotes, making homology-based annotation transfer unreliable. On the positive side, this high divergence can be exploited to develop trypanocidal compounds with less toxicity to the host. Reasoning that physically interacting proteins as well as sets of co-regulated genes are significantly enriched for genes with similar biological roles, we developed a web-based database for functional annotations of trypanosomatid proteins based on the integration of the obtained results from second and fourth chapters of this work with other available resources such as transcriptome, proteome, gene ontology, biological pathways, and protein sequence features (Jensen et al. 2009; Kabani et al. 2009; Minning et al. 2009; Queiroz et al. 2009; Rochette et al. 2009; Aslett et al. 2010; Kramer et al. 2010b; Nilsson et al. 2010; Veitch et al. 2010; Alsford et al. 2011; Haanstra et al. 2011; Gunasekera et al. 2012; Urbaniak et al. 2012; Urbaniak et al. 2013; Fadda et al. 2014; Kanehisa et al. 2014; Gazestani and Salavati 2015; Gene Ontology 2015; Gazestani et al. 2016b). As of Aug. 7th 2016, the database harbors 94,764 interactions among 12,236 proteins from 16 different trypansomatid species and/or strains.

Background

Trypanosomatid parasites cause life-threatening diseases in humans and major production losses in animals. Although they pose global threats, various issues are associated with available drugs against trypanosomatids (including tolerability, cost, and resistance), necessitating the identification of novel essential parasitic-specific pathways/genes as potential drug targets (Gehrig and Efferth 2008; Simarro et al. 2008; Brun et al. 2010; Barrett et al. 2011; Stich et al. 2013). However, supported by whole genome sequencing data, it is long known that species of trypanosomatid family, while showing high similarity in proteome with each other, are highly diverged from other eukaryotes (Berriman et al. 2005; El-Sayed et al. 2005a; El-Sayed et al. 2005b; Ivens et al. 2005). This makes the annotation transfer of nearly half of their proteome by homology-based approaches from model organisms unreliable (El-Sayed et al. 2005b).
Over the past two decades, several genome-wide as well as focused studies have been conducted to functionally characterize trypanosomatid proteins. Construction of global and local protein interaction maps have served as one of the main resources for functional annotation by reflecting the molecular context of proteins in a cell (Li and Wang 2002; Luz Ambrosio et al. 2009; Hernandez et al. 2010; Acestor et al. 2011; Aphasizheva et al. 2011; Ammerman et al. 2012; Cestari et al. 2013; Ridlon et al. 2013; Freire et al. 2014; Gazestani et al. 2016b). Several experimental techniques exist to identify the interacting partners of proteins that differ in selectivity and sensitivity. Therefore, one major challenge in study of protein interactions is to be able to distinguish between the correctly associated proteins from confounding elements present in the results of these experiments. It is also helpful to know the potential interacting proteins that are missing in the results of an experiment based on the previously known knowledge on the species of interest or other related trypanosomatid species. Several databases have been developed to represent the experimentally identified or computationally inferred physical and functional protein interactions (Bader et al. 2003; Ruepp et al. 2010; Kerrien et al. 2012; Franceschini et al. 2013; Zuberi et al. 2013; Chatr-Aryamontri et al. 2015). Such databases have greatly helped researchers to interrogate cellular processes and have a systems level view on protein(s) of choice. Although of critical importance for studies on trypanosomatids, only a limited number of databases cover information on protein interactions of these parasites and such interactions are mostly predicted by transferring the available data from other eukaryotes, missing most part of published data on trypanosomatid species (Franceschini et al. 2013).

Another major approach for functional characterization of proteins stems from recent technological advances that has allowed measuring transcriptome, proteome, and transcript half-life changes in response to environmental changes, different life stages, or cell conditions (Jensen et al. 2009; Kabani et al. 2009; Minning et al. 2009; Queiroz et al. 2009; Rochette et al. 2009; Aslett et al. 2010; Kramer et al. 2010b; Nilsson et al. 2010; Veitch et al. 2010; Alsford et al. 2011; Haanstra et al. 2011; Gunasekera et al. 2012; Urbaniak et al. 2012; Urbaniak et al. 2013; Fadda et al. 2014; Kanehisa et al. 2014; Gazestani and Salavati 2015; Gazestani et al. 2016b). Moreover, it is possible to gain insights on function of a protein by gathering information on: 1) its annotation from resources such as gene annotation and KEGG pathways (Kanehisa et al. 2014; Gene Ontology 2015); 2) protein sequence features such as motifs, isoelectric point, molecular weights, and number of transmembrane domains; 3) the essentiality of the gene knock-down on the cell survival (Alsford et al. 2011); 4) potential cis-regulatory elements present in the 3′-UTR of the gene and the collective response of genes containing that motif to the environmental changes (Gazestani and Salavati 2015). Currently, TriTrypDB database is devoted to the kinetoplastid parasites and provides a wide-range of information, in a gene-centric framework, on a queried protein or terms ranging from genomic sequence and position to captured responses of the interested protein in previously reported studies (Aslett et al. 2010). However, in many cases, researchers are interested to know the collective response of a list of pre-specified proteins along with their interacting partners according to large-scale studies rather than focusing on one protein. Combining interaction data with enrichment analysis of gene ontology, molecular pathways, gene essentiality, and protein sequence features is the key to perceive the function of proteins.

Here, we describe TrypsNetDB, a user friendly, integrated database that tries to fill the aforementioned gaps by not only depositing the current interactome knowledge on trypanosomatid proteins, but also combining such information with other available resources accompanied with related statistical analysis. Moreover, the database automatically performs inter-species mapping of available data and information to help for better characterization of queried proteins in the species of interest. Finally, based on the built in features, the database is able to help researchers on their interactome related experiments to distinguish between the likely binding partners of a protein from confounding elements identified in their experimentsand suggest other potentially interacting proteins that are missing from the list of queried proteins.

Built on powerful asp.Net programming, the database performance is fast and reliable. TrypsNetDB is freely available at trypsNetDB.org.

Program description and methods

Overall view of the databasae

The current release of the database is focused on physical protein interaction data that are already published in trypanosomatid field, supporting 16 trypanosomatid parasites including *T. cruzi* strain CL Brener, *T. cruzi* CL Brener Esmeraldo-like, *T. cruzi* CL Brener Non-Esmeraldo-like, *T. brucei* gambiense DAL972, *T. brucei* Lister strain 427, *T. vivax* Y486, *T. evansi* strain STIB 805, *T. brucei* TREU927, *L. major* strain Friedlin, *L. mexicana* MHOM/GT/2001/U1103, *L. infantum* JPCM5, *L. donovani* BPK282A1, *L. braziliensis MHOM/BR/75/M2903, L. braziliensis MHOM/BR/75/M2904, L. arabica strain LEM1108, L. enriettii* strain LEM3045. The interaction data come from a variety of techniques including affinity purification, immunoprecipitation, Y2H, fractionation patterns, other possible experimental techniques, and ortholog mapping. We have only considered synthenic orthologs reported by TriTrypDB to transfer the inter-species information. Users can query the database based on either tritrypDB IDs (recognizing IDs of recent and older versions of the database) or gene names. Support for the rest of trypanosomatid species are scheduled to be added in the coming months. In cases that gene names match to multiple organisms, the user will be asked to select the species of interest from a dropdown box. As shown in Fig 6.1, querying protein(s) will redirect the user to the interaction page composed of three main elements: information panel, network, and reference section.

Information panel can be used to explore details on three sections annotations, protein descriptions, and the constructed network. The annotation tab contains gene set enrichment analysis of the combined set of queried proteins with the suggested proteins by the database for enrichment in five different categories: 1) GO & KEGG: genes are examined for enrichment in Gene ontology and KEGG pathway annotations using hypergeometric tests. Terms with Benjamini-Hochberg corrected p-value less than 0.05 are reported back to the user. Hovering

over each term will highlight the proteins that are associated with the term. Clicking on each represented term will show description of the term, its category, number of proteins in the network associated with that term, and the corresponding corrected p-value. 2) Sequence & Structure: The sequence and structural features of proteins are examined such as protein motifs, isoelectric points, molecular weights, and predicted number of transmembrane domains. This information can provide complementary view on the function of the proteins. As an illustration, a group of soluble interacting proteins are expected to have significantly low number of transmembrane domains. Likewise, proteins interacting with RNA and DNA are expected to have high isoelectric points. Similar to GO and KEGG enrichment results, hovering over significantly enriched protein motifs will highlight associated proteins in the network. 3) Expression patterns: proteins are examined for their collective response across 48 distinct samples. Each sample is color coded where yellow and blue colors indicate over-expression/enrichment and under-expression/depletion, respectively. Statistically significant terms with p-value less than 0.05 are highlighted by darker colors, while non-significant conditions are semi-transparent. The 48 considered cell states come from genome-wide experiments on *T. brucei*, *T. cruzi*, and *L. infantum*. By ortholog mapping, the database automatically propagates the information to other trypanosomatid species. Clicking on each sample will open information on the title of the sample, description of the results of statistical test, calculated p-value, the title of the study that published the sample and its PMID with a link to the PubMed abstracts. 4) Gene essentiality: The essentiality of proteins in four different cell conditions of *T. brucei* are examined based on the results of a genome-wide phenotyping study. Ortholog mapping is performed for cases that the queried organism is other that *T. brucei*. 5) 3' regulatory elements: Using a novel approach, we recently predicted 88 cis-regulatory elements that are potentially involved in the developmental regulation of *T. brucei*. Although only a limited number of functional elements have been identified thus far, by rigorous analysis of results, we showed that 11 predicted motifs strikingly resemble previously identified regulatory elements in trypanosomatids, suggesting the high accuracy of the predictions. This section examines whether

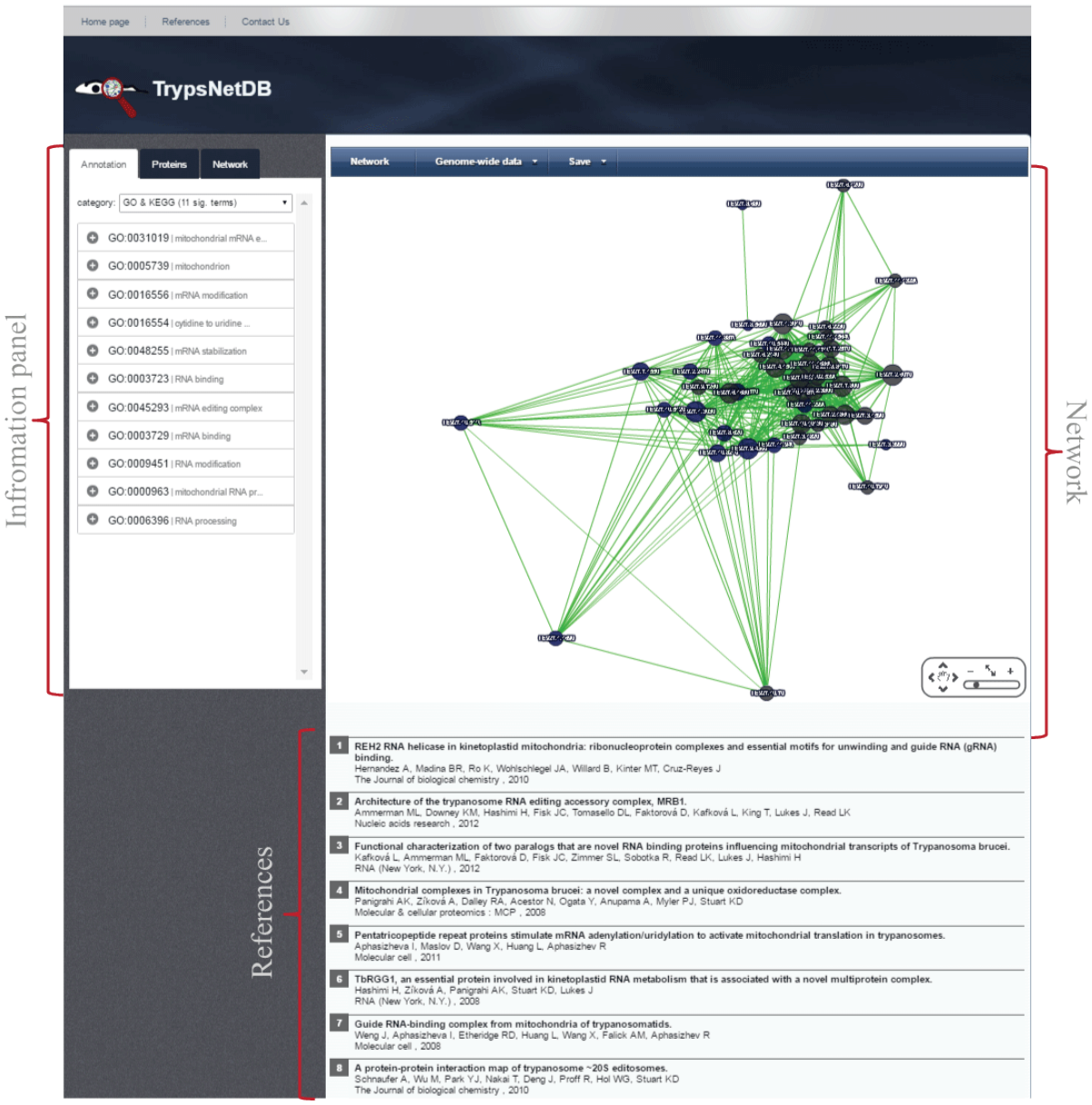


Fig 6.1. The main result page of TrypsNetDB.
The main result page is composed of three elements of infromaion panel, network section, and references section. Information panel contains enrichment analysis results, brief characteristics on the proteins, and interaction types present in the illustrated network. The network section can be used to explore interactions among the proteins. Using the menu on top of the interaction section, users can change the visualization style of the network, access the dynamic heatmaps from various genome-wide data, and save the results. References section include studies from which the interaction data of the illustrated network was extracted.

3′-UTRs of orthologs of the set of proteins in *T. brucei* are significantly enriched for any of the predicted 88 motifs. In cases that enrichment is found, the motif logo along with the transcriptome and proteome response of the motif in different cell conditions would be reported.
The proteins tab provides a brief information (such as transcript and protein length, isoelectric point, molecular weight, etc.) with a link to the TriTrypDB database in a sorted way starting from the queried proteins and ending by the proteins that were included in the network by the program (these suggested proteins have been highly connected to the queried proteins based on the literature derived interactions).
The network tab can be used to explore the contribution of each of experimental techniques in the construction of the illustrated network. In two cases of affinity purification and immunoprecipitation that interactions can show indirect associations, database distinguishes between interactions that are identified based on RNase treatment of samples from those that are not. Hovering over each technique will highlight the interactions that they support. It is also possible to filter out some of the techniques by unchecking the corresponding checkbox and clicking on the “set filters” button.
The network section, using a dynamically interactive interface, represents the interactions among the proteins where each protein is indicated by a circular node. It is possible to zoom in or out the network and reposition the proteins. Queried proteins and other proteins suggested by the database are shown in blue and gray colors, respectively. The Node size of proteins indicates the number of interactions that they have in the global network with larger nodes representing nodes with higher number of interactions. Selecting a protein by clicking on it will highlight the first neighbors of that protein and open the corresponding information in the proteins tab of the information panel. Finally, the network option on top-left part of the network section can be used to automatically rearrange the network for perhaps better presentation or show/hide the protein labels, useful for visualization of relatively big networks.
The reference section states the references that interactions where extracted from. The full source of resources used for the extraction of interaction data can be accessed by going to the “References” section from top menu or going directly to trypsNetDB.org/references.aspx webpage.

Genome-wide data section

"Genome-wide data" section on the menu enables users to visualize the genome-wide data available for the queried proteins as well as their interacting partners suggested by the database. The supported genome-wide data in the current release of the database are categorized in three main groups of fractionation patterns, gene expressions patterns, and phenotypic effects, each containing sub-categories. Users can select one of the main categories (indicated with blue background) to represent all related sub-categories at once, or directly select the sub-categories. This part of the database is particularly useful for the validation of results obtained from interactome-related experiments (such as affinity purifications) by helping users to distinguish between direct binding partners of a protein from potentially spurious elements. For example, the fractionation heatmaps can be exploited to assess whether the potentially interacting proteins show similar fractionation patterns (Fig 6.2). Moreover, the database provides fractionation patterns for whole-cell, mitochondrial-enriched, and cytosolic-enriched cell extracts which can be informative on the localization of the previously unannotated proteins (i.e., mitochondrial proteins are expected to be identified in the mitochondrial-enriched fractions while depleted in the cytosolic-enriched sample). Fractionation patterns are informative on the nature of interactions as well. As elaborated elsewhere (Gazestani et al. 2016b), glycerol gradient-based fractionation patterns can capture more transient interactions, while ion exchange-based fractionation favors more stable interactions due to the presence of salt gradient. Finally, physically interacting proteins are expected to be involved in similar biological processes and, hence, show similar expression patterns and degree of essentiality in each cell state which can be easily assessed using the corresponding heatmaps.

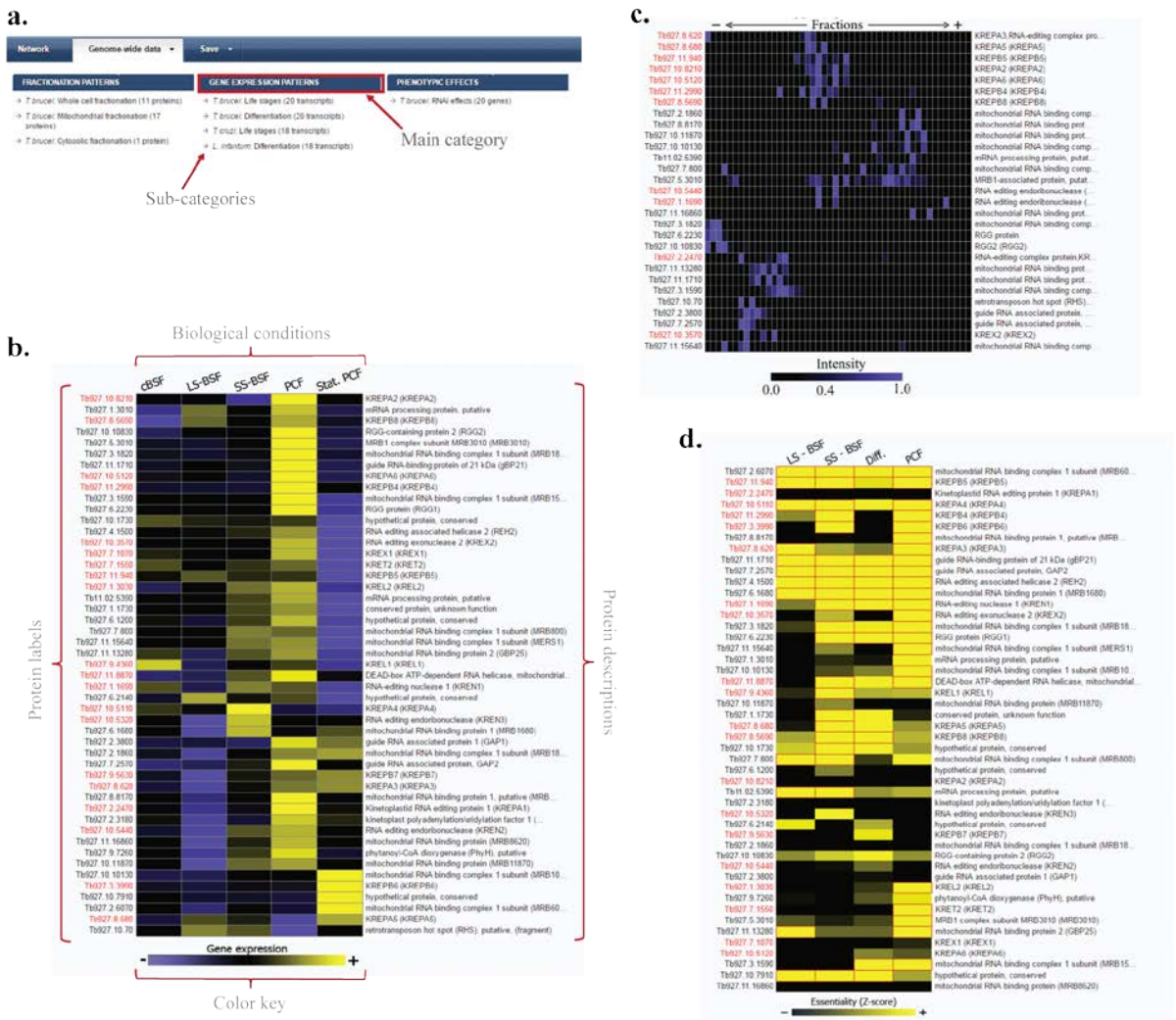


Fig 6.2. Genome-wide section of TrypsNetDB.
(a) Genome-wide section can be accessed from the menu located on top of the interaction section. Users can visualize all relevant data by selecting a category or visualize a more specific dataset by choosing the corresponding sub-category. (b) A sample representation of transcriptome heatmap. Queried proteins are highlighted by red labels on the left of the heatmap, while suggested proteins by the database are with black labels. (c) A sample representation of fractionation heatmap. (d) A sample representation of gene essentiality heatmap. In each life stage (i.e., each column), statistically significant genes are represented by red borders.

Saving the results

By going to the save option on top of the network section, users are able to save the whole represented network or only the sub-networks that are supported by a specific experimental technique. It is also possible to save the enrichment analysis results as well as annotations of

genes such as description, transcript or protein characteristics (length, weight, isoelectric point, identified SNPs), and gene ontology.

Implementation

The web application is developed based on the .Net framework 4.5 technology. To improve the performance, the statistical analysis modules were implemented in C# and added as library to the web application and the performance of modules has been validated by comparing the results with those of MATLAB 2015b on multiple test sets. The network visualization is based on the *cytocapeweb* library that requires flash player for the representation of the network. All analysis is performed in real-time and session for each user will be ended after 1hr of inactivity. For high-performance, the database is implemented in Microsoft SQL Server 2012.

Conclusions and future directions

Protein interaction maps remains as one of the major resources for the functional annotation of proteins. Embedding other lines of information with these maps can help researchers to gain insights on the molecular contexts of the proteins. Here, we introduced TrypsNetDB, a web tool to consolidate the current knowledge on the interactome of the trypanosomatid parasites and dynamically integrate them with wealth of available orthogonal information. We are continuously working to expand the core, literature-derived, protein interaction depository of the database. Future plans also include supports for the rest of trypanosomatid parasites and inclusion of other genome-wide data.

Chapter 6

Conclusion and future work

Trypanosomatid parasites can infect both humans and animals. The spread of these parasites disproportionally affect low income countries, where nursing expenses and animal production losses have greater impacts on the societies. However, the well-developed countries are not immune to the threat of these parasites, and increasing cases are reported in Canada, U.S.A, parts of Europe, as well as Japan and Australia (World Health 2012). Several issues are associated with available treatments including sever side-effects, need of nursing during administration, cost, and most importantly emerging resistance (Baker et al. 2013). History shows us that several major epidemics of trypanosomiasis has occurred in susceptible countries when disease control programs were stopped for some reasons such as fighting wars or treatment costs, underscoring the importance of drug development programs.

Supported by multiple lines of evidence, in my opinion, targeting gene regulatory programs is a highly promising way to combat trypanosomiasis. The main reason for such claim is high divergence of these pathways in trypanosomatids compared with other eukaryotes, including their hosts (Kramer 2012). In the absence of transcriptional control, these parasites rely extensively on rather less complicated post transcriptional regulation mechanisms, where miRNAs play no apparent role in the gene expression (Kramer and Carrington 2011; Kramer 2012; Schwede et al. 2012). Experimental evidence suggests key roles for RBPs in developmental regulation of trypanosomatids. Importantly, comparative studies have experimentally shown that the binding preference of trypanosomatid RBPs are diverged from their counterparts in other eukaryotes during the evolution (Ray et al. 2013). Despite the high importance, the current knowledge on the trypanosomatid gene regulatory programs is only rudimentary. In this work, we sought to gain insights on three different aspects of such processes using novel systematic approaches.

As the first aim, we showed that gene sets harboring same functional RNA regulatory elements form significantly dense modules in co-expression graphs, enabling the systematic integration of different data sources to infer RNA regulatory elements. This property becomes particularly important in the study of non-model organisms with limited whole genome expression data. Application of our approach led to the prediction of 88 motifs that are potentially involved in the developmental and environmental regulation of trypanosomatid genes. However, these predictions need experimental verifications before being exploited for the drug development. One possible approach is to test the functionality of these motifs by using recently developed

RNA footprinting approaches in which the regions of RNAs bound to the proteins are identified using RNase treatments (Baltz et al. 2012; Freeberg et al. 2013; Silverman et al. 2014). Application of such approaches on the two life stages of *T. brucei* can provide the regions that are potentially recognized by RBPs. Matching such results with 88 predicted motifs can not only verify the predictions, but also highlight the functional instances of the motifs. Alternatively, we have designed a dual-colored system that combines flow cytometry with high throughput sequencing techniques to simultaneously examine the regulatory role (in either RNA stability and/or translation) of a library containing a large number of sequences *in vivo*. Our strategy utilises a dual-color reporter system expressing GFP that incorporates the library in its 3′-UTR region and mCherry as the internal control. We expect that insertion of a stabilizer or translational-enhancer element increases the relative light intensity of GFP-to-mCherry, while insertion of a destabiliser or translational suppressor decreases this ratio. To identify RREs and decipher the mechanism of regulation, after the stable transfection of the library, cells can be categorized based on their GFP-to-mCherry ratio and then subjected to high-throughput sequencing. The preliminary results are high promising (data not shown). The beauty of the designed system is that by small modifications, it can be used as a cell-based assay to target the trypanosomatid post-transcriptional gene regulatory circuits. We can expect that by targeting the *cis*-regulatory elements that are critical for the developmental processes of trypanosomatids, we can stop the transmission of the parasite from the host to the fly vector or vice versa.

T. brucei Mitochondria plays a crucial role in the parasites survival by adapting the parasite to varying environmental energy sources. As the second aim, we developed a novel next generation sequencing based approach to identify the 3’ tails of the mitochondrial transcripts. Application of the method on a limited set of *T. brucei* mitochondrial transcripts in PS life stage led to discovery of transcript-specific variation in tails populations of mitochondrial mRNAs. However, the impact of such tail variations on gene regulation is still unknown. Moreover, the relationship among events of polycistron processing, in-tail addition, and RNA editing is largely elusive. Application of the developed approach on a large set of mitochondrial mRNAs in different biological conditions is critical to gain such information.

One of the major challenges for the drug development programs for *T. brucei* is the lack of annotation of nearly half of its genes. In the third aim, we constructed the first, experimentally-derived, protein co-complex map of *T. brucei*. This resource is highly useful for the annotation of

previously uncharacterized proteins. Moreover, by focusing on the mitochondrial post-transcriptional regulatory machinery of *T. brucei*, we were able to identify novel components and provide new insight on the association types of previously identified components with the machinery. We expect that by conducting similar experiments on other trypanosomatid parasites we would be able to discern conserved interaction maps from those of species-specific maps, which would be invaluable for prioritizing novel proteins as potential drug targets. Moreover, we have shown that integration of different data types would highly increase the precision of the protein interaction maps. Consequently, the high confidence section of the fractionation-derived network was extracted by combining the interaction data with other orthogonal resources. However, these orthogonal resources were highly biased towards specific biological processes and cellular compartments, leading to the lack of support for parts of the network. Hence, performing orthogonal experiments (such as affinity purifications) on previously unstudied proteins that are present in the co-fractionation network, would increase the coverage of the high confidence network.

As the last aim, we developed a web-based tool to integrate the results of this work with other available resources for trypanosomatids, and represent the results in a user-friendly and intuitive way to the users. As the first and only resource of its kind in the field of trypanosomatids, TrypsNetDB is scheduled to have HTML mutual cross-references with TriTrypDB, a major database devoted to the kinetoplastid parasites. We are continuously working on the core section of the database, manual extraction of protein interaction data from published studies, and expect to have at least three-fold increase of coverage in this regard. The graphics of the database can be also improved to enhance users' experience. More functional analysis sections, such as dynamic protein complex/module identification, need to be added to the database. We are also interested in the implementation of web-services required to enable querying the database form the Cytoscape software (Cline et al. 2007; Orchard et al. 2010).

Chapter 7

Contribution to knowledge

In this work, we aimed to develop tools and perform analysis that were critical to decipher the post transcriptional gene regulatory circuits in T. brucei with the hope that obtained knowledge can be eventually exploited to identify compounds against trypanosomiasis. The contributions of this thesis to knowledge are:

1- Critical review of previously employed techniques to identify and characterize RBPs and their cognate cis-regulatory elements.

2- Discovering of 88 high-confidence cis-regulatory elements using a novel graph-based approach with an estimated accuracy of more than 60% based on benchmarking on human. These predicted elements are potentially involved in the transmission of the *T. brucei* from the vector to the host. By systematic integration of limited transcriptome data, the developed method bypasses the need for existence of comprehensive transcriptome data to accurately predict funcitonal RREs, allowing the prediction of RREs for non-model organisms, for many of which no other reliable method is available.

3- Introducing a novel next-generation sequencing-based approach to investigate the effect of RBPs on the tails of mitochondrial transcripts. To the best of our knowledge, this work provides highest depth of tail analysis compared to other techniques developed thus far.

4- Establishing the first experimentally derived protein co-complex map of *T. brucei*. Our results led to the assignment of 716 proteins including 635 proteins that lacked experimental annotation to protein complexes.

5- Introducing a protein interaction-centric database to automatically integrate the obtained knowledge from the previous aims with other available resources accompanied with related statistical analysis.

6- Introducing a high-throughput *in vivo* screening approach for experimental validation of RRE predictions as well as a cell-based assay for anti-trypanosomatids lead discovery based on the targeting of trypanosomatid gene regulatory pathways. The preliminary results of both designed

systems are highly encouraging and they are being actively pursued in our lab at McGill University as well as in collaboration with Dr. Zimmer at University of Minnesota.

Bibliography

Acestor N, Zikova A, Dalley RA, Anupama A, Panigrahi AK, Stuart KD. 2011. Trypanosoma brucei mitochondrial respiratome: composition and organization in procyclic form. *Mol Cell Proteomics* **10**: M110 006908.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831-838.

Allers J, Shamoo Y. 2001. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J Mol Biol* **311**: 75-86.

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

Alvar J, Velez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M, Team WHOLC. 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* **7**: e35671.

Ammerman ML, Downey KM, Hashimi H, Fisk JC, Tomasello DL, Faktorova D, Kafkova L, King T, Lukes J, Read LK. 2012. Architecture of the trypanosome RNA editing accessory complex, MRB1. *Nucleic Acids Res* **40**: 5637-5650.

Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. 2003. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**: 570-574.

Anderson BA, Wong IL, Baugh L, Ramasamy G, Myler PJ, Beverley SM. 2013. Kinetoplastid-specific histone variant functions are conserved in Leishmania major. *Mol Biochem Parasitol* **191**: 53-57.

Aphasizhev R, Aphasizheva I. 2011. Mitochondrial RNA processing in trypanosomes. *Res Microbiol* **162**: 655-663.

-. 2014. Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* **100**: 125-131.

Aphasizhev R, Aphasizheva I, Simpson L. 2003. A tale of two TUTases. *Proc Natl Acad Sci U S A* **100**: 10617-10622.

Aphasizhev R, Sbicego S, Peris M, Jang SH, Aphasizheva I, Simpson AM, Rivlin A, Simpson L. 2002. Trypanosome mitochondrial 3' terminal uridylyl transferase (TUTase): the key enzyme in U-insertion/deletion RNA editing. *Cell* **108**: 637-648.

Aphasizheva I, Aphasizhev R. 2010. RET1-catalyzed uridylylation shapes the mitochondrial transcriptome in Trypanosoma brucei. *Mol Cell Biol* **30**: 1555-1567.

Aphasizheva I, Maslov D, Wang X, Huang L, Aphasizhev R. 2011. Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol Cell* **42**: 106-117.

Aphasizheva I, Zhang L, Wang X, Kaake RM, Huang L, Monti S, Aphasizhev R. 2014. RNA binding and core complexes constitute the U-insertion/deletion editosome. *Mol Cell Biol* **34**: 4329-4342.

Archer SK, Inchaustegui D, Queiroz R, Clayton C. 2011. The cell cycle regulated transcriptome of Trypanosoma brucei. *PLoS One* **6**: e18425.

Archer SK, Luu VD, de Queiroz RA, Brems S, Clayton C. 2009. Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle. *PLoS Pathog* **5**: e1000565.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**: D457-462.

Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* **24**: 282-284.

Bader GD, Betel D, Hogue CW. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**: 248-250.

Baker N, de Koning HP, Maser P, Horn D. 2013. Drug resistance in African trypanosomiasis: the melarsoprol and pentamidine story. *Trends Parasitol* **29**: 110-118.

Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M et al. 2012. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* **46**: 674-690.

Barabasi AL. 2009. Scale-free networks: a decade and beyond. *Science* **325**: 412-413.

Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. 2009. The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* **37**: D396-403.

Barrett MP. 2001. Veterinary link to drug resistance in human African trypanosomiasis? *Lancet* **358**: 603-604.

Barrett MP, Boykin DW, Brun R, Tidwell RR. 2007. Human African trypanosomiasis: pharmacological re-engagement with a neglected disease. *Br J Pharmacol* **152**: 1155-1171.

Barrett MP, Vincent IM, Burchmore RJ, Kazibwe AJ, Matovu E. 2011. Drug resistance in human African trypanosomiasis. *Future Microbiol* **6**: 1037-1047.

Beilharz TH, Preiss T. 2011. Polyadenylation state microarray (PASTA) analysis. *Methods Mol Biol* **759**: 133-148.

Berriman M Ghedin E Hertz-Fowler C Blandin G Renauld H Bartholomeu DC Lennard NJ Caler E Hamlin NE Haas B et al. 2005. The genome of the African trypanosome Trypanosoma brucei. *Science* **309**: 416-422.

Bhandari D, Guha K, Bhaduri N, Saha P. 2011. Ubiquitination of mRNA cycling sequence binding protein from Leishmania donovani (LdCSBP) modulates the RNA endonuclease activity of its Smr domain. *FEBS Lett* **585**: 809-813.

Bhat GJ, Souza AE, Feagin JE, Stuart K. 1992. Transcript-specific developmental regulation of polyadenylation in Trypanosoma brucei mitochondria. *Mol Biochem Parasitol* **52**: 231-240.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Broadhead R, Dawe HR, Farr H, Griffiths S, Hart SR, Portman N, Shaw MK, Ginger ML, Gaskell SJ, McKean PG et al. 2006. Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature* **440**: 224-227.

Broder S, Yarchoan R, Collins JM, Lane HC, Markham PD, Klecker RW, Redfield RR, Mitsuya H, Hoth DF, Gelmann E et al. 1985. Effects of suramin on HTLV-III/LAV infection presenting as Kaposi's sarcoma or AIDS-related complex: clinical pharmacology and suppression of virus replication in vivo. *Lancet* **2**: 627-630.

Brun R, Blum J, Chappuis F, Burri C. 2010. Human African trypanosomiasis. *Lancet* **375**: 148-159.

Buhlmann M, Walrad P, Rico E, Ivens A, Capewell P, Naguleswaran A, Roditi I, Matthews KR. 2015. NMD3 regulates both mRNA and rRNA nuclear export in African trypanosomes via an XPOI-linked pathway. *Nucleic Acids Res* **43**: 4491-4504.

Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. *Nat Genet* **27**: 167-171.

Butter F, Bucerius F, Michel M, Cicova Z, Mann M, Janzen CJ. 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.

Caniza H, Romero AE, Heron S, Yang H, Devoto A, Frasca M, Mesiti M, Valentini G, Paccanaro A. 2014. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics* **30**: 2235-2236.

Caro F, Bercovich N, Atorrasagasti C, Levin MJ, Vazquez MP. 2006. Trypanosoma cruzi: analysis of the complete PUF RNA-binding protein family. *Exp Parasitol* **113**: 112-124.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3**: e245.

Castello A, Fischer B, Frese CK, Horos R, Alleaume AM, Foehr S, Curk T, Krijgsveld J, Hentze MW. 2016. Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell*.

Cestari I, Kalidas S, Monnerat S, Anupama A, Phillips MA, Stuart K. 2013. A multiple aminoacyl-tRNA synthetase complex that enhances tRNA-aminoacylation in African trypanosomes. *Mol Cell Biol* **33**: 4872-4888.

Chan CS, Elemento O, Tavazoie S. 2005. Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol* **1**: e69.

Chang H, Lim J, Ha M, Kim VN. 2014. TAIL-seq: genome-wide determination of poly(A) tail length and 3' end modifications. *Mol Cell* **53**: 1044-1052.

Chang JH, Tong L. 2012. Mitochondrial poly(A) polymerase and polyadenylation. *Biochim Biophys Acta* **1819**: 992-997.

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L et al. 2015. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**: D470-478.

Checchi F, Filipe JA, Haydon DT, Chandramohan D, Chappuis F. 2008. Estimates of the duration of the early and late stage of gambiense sleeping sickness. *BMC Infect Dis* **8**: 16.

Cheson BD, Levine AM, Mildvan D, Kaplan LD, Wolfe P, Rios A, Groopman JE, Gill P, Volberding PA, Poiesz BJ et al. 1987. Suramin therapy in AIDS and related disorders. Report of the US Suramin Working Group. *JAMA* **258**: 1347-1351.

Chijioke CP, Umeh RE, Mbah AU, Nwonu P, Fleckenstein LL, Okonkwo PO. 1998. Clinical pharmacokinetics of suramin in patients with onchocerciasis. *Eur J Clin Pharmacol* **54**: 249-251.

Clayton C. 2013. The regulation of trypanosome gene expression by RNA-binding proteins. *PLoS Pathog* **9**: e1003680.

Clayton CE. 2014. Networks of gene expression regulation in Trypanosoma brucei. *Mol Biochem Parasitol* **195**: 96-106.

Clerinx J, Vlieghe E, Asselman V, Van de Casteele S, Maes MB, Lejon V. 2012. Human African trypanosomiasis in a Belgian traveller returning from the Masai Mara area, Kenya, February 2012. *Euro Surveill* **17**.

Clery A, Blatter M, Allain FH. 2008a. RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol* **18**: 290-298.

Clery A, Blatter M, Allain FHT. 2008b. RNA recognition motifs: boring? Not quite. *Curr Opin Struc Biol* **18**: 290-298.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**: 2366-2382.

Collins JM, Klecker RW, Jr., Yarchoan R, Lane HC, Fauci AS, Redfield RR, Broder S, Myers CE. 1986. Clinical pharmacokinetics of suramin in patients with HTLV-III/LAV infection. *J Clin Pharmacol* **26**: 22-26.

Dame DA, Jordan AM. 1981. Control of tsetse flies, Glossina spp. *Adv Vet Sci Comp Med* **25**: 101-119.

Damper D, Patton CL. 1976. Pentamidine transport and sensitivity in brucei-group trypanosomes. *J Protozool* **23**: 349-356.

Das A, Morales R, Banday M, Garcia S, Hao L, Cross GA, Estevez AM, Bellofatto V. 2012. The essential polysome-associated RNA-binding protein RBP42 targets mRNAs involved in Trypanosoma brucei energy metabolism. *RNA* **18**: 1968-1983.

David M, Gabdank I, Ben-David M, Zilka A, Orr I, Barash D, Shapira M. 2010. Preferential translation of Hsp83 in Leishmania requires a thermosensitive polypyrimidine-rich element in the 3' UTR and involves scanning of the 5' UTR. *RNA* **16**: 364-374.

De Gaudenzi J, Frasch AC, Clayton C. 2005. RNA-binding domain proteins in Kinetoplastids: a comparative analysis. *Eukaryot Cell* **4**: 2106-2114.

Decker CJ, Sollner-Webb B. 1990. RNA editing involves indiscriminate U changes throughout precisely defined editing domains. *Cell* **61**: 1001-1011.

Deo RC, Bonanno JB, Sonenberg N, Burley SK. 1999. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98**: 835-845.

Diebel KW, Smith AL, van Dyk LF. 2010. Mature and functional viral miRNAs transcribed from novel RNA polymerase III promoters. *RNA* **16**: 170-185.

Dostalova A, Kaser S, Cristodero M, Schimanski B. 2013. The nuclear mRNA export receptor Mex67-Mtr2 of Trypanosoma brucei contains a unique and essential zinc finger motif. *Mol Microbiol* **88**: 728-739.

Draper DE. 1999. Themes in RNA-protein recognition. *J Mol Biol* **293**: 255-270.

Droll D, Minia I, Fadda A, Singh A, Stewart M, Queiroz R, Clayton C. 2013. Post-transcriptional regulation of the trypanosome heat shock response by a zinc finger protein. *PLoS Pathog* **9**: e1003286.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G et al. 2005a. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. *Science* **309**: 409-415.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C et al. 2005b. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**: 404-409.

Elemento O, Slonim N, Tavazoie S. 2007. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28**: 337-350.

Elkon R, Drost J, van Haaften G, Jenal M, Schrier M, Oude Vrielink JA, Agami R. 2012. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol* **13**: R59.
Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM et al. 2012. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* **46**: 884-892.
Erben E, Chakraborty C, Clayton C. 2014a. The CAF1-NOT complex of trypanosomes. *Front Genet* **4**: 299.
Erben ED, Fadda A, Lueong S, Hoheisel JD, Clayton C. 2014b. A genome-wide tethering screen reveals novel potential post-transcriptional regulators in Trypanosoma brucei. *PLoS Pathog* **10**: e1004178.
Etheridge RD, Aphasizheva I, Gershon PD, Aphasizhev R. 2008. 3' adenylation determines mRNA abundance and monitors completion of RNA editing in T. brucei mitochondria. *EMBO J* **27**: 1596-1608.
Fadda A, Ryten M, Droll D, Rojas F, Farber V, Haanstra JR, Merce C, Bakker BM, Matthews K, 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 AH, Bowman IB. 1977. Trypanosoma brucei: suramin and other trypanocidal compounds' effects on sn-glycerol-3-phosphate oxidase. *Exp Parasitol* **43**: 353-361.
Fairlamb AH, Bowman IB. 1980. Uptake of the trypanocidal drug suramin by bloodstream forms of Trypanosoma brucei and its effect on respiration and growth rate in vivo. *Mol Biochem Parasitol* **1**: 315-333.
Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**: e8.
Feagin JE, Jasmer DP, Stuart K. 1987. Developmentally regulated addition of nucleotides within apocytochrome b transcripts in Trypanosoma brucei. *Cell* **49**: 337-345.
Feagin JE, Stuart K. 1988. Developmental aspects of uridine addition within mitochondrial transcripts of Trypanosoma brucei. *Mol Cell Biol* **8**: 1259-1265.
Fenn K, Matthews KR. 2007. The cell biology of Trypanosoma brucei differentiation. *Curr Opin Microbiol* **10**: 539-546.
Fernandez-Moya SM, Carrington M, Estevez AM. 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.
Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ. 2005. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A* **102**: 17675-17680.
Foat BC, Stormo GD. 2009. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. *Mol Syst Biol* **5**: 268.
Foda BM, Downey KM, Fisk JC, Read LK. 2012. Multifunctional G-rich and RRM-containing domains of TbRGG2 perform separate yet essential functions in trypanosome RNA editing. *Eukaryot Cell* **11**: 1119-1131.
Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C et al. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**: D808-815.

Freeberg MA, Han T, Moresco JJ, Kong A, Yang YC, Lu ZJ, Yates JR, Kim JK. 2013. Pervasive and dynamic protein binding sites of the mRNA transcriptome in Saccharomyces cerevisiae. *Genome Biol* **14**: R13.

Freire ER, Malvezzi AM, Vashisht AA, Zuberek J, Saada EA, Langousis G, Nascimento JD, 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.

Furger A, Schurch N, Kurath U, 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.

Gazestani VH, Hampton M, Abrahante JE, Salavati R, Zimmer SL. 2016a. circTAIL-seq, a targeted method for deep analysis of RNA 3' tails, reveals transcript-specific differences by multiple metrics. *RNA* **22**: 477-486.

Gazestani VH, Lu Z, Salavati R. 2014. Deciphering RNA regulatory elements in trypanosomatids: one piece at a time or genome-wide? *Trends Parasitol* **30**: 234-240.

Gazestani VH, Nikpour N, Mehta V, Najafabadi HS, Moshiri H, Jardim A, Salavati R. 2016b. A Protein Complex Map of Trypanosoma brucei. *PLoS Negl Trop Dis* **10**: e0004533.

Gazestani VH, Salavati R. 2015. Deciphering RNA Regulatory Elements Involved in the Developmental and Environmental Gene Regulation of Trypanosoma brucei. *PLoS One* **10**: e0142342.

Gehrig S, Efferth T. 2008. Development of drug resistance in Trypanosoma brucei rhodesiense and Trypanosoma brucei gambiense. Treatment of human African trypanosomiasis with natural products (Review). *Int J Mol Med* **22**: 411-419.

Gene Ontology C. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049-1056.

Gerstberger S, Hafner M, Ascano M, Tuschl T. 2014a. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv Exp Med Biol* **825**: 1-55.

Gerstberger S, Hafner M, Tuschl T. 2014b. A census of human RNA-binding proteins. *Nat Rev Genet* **15**: 829-845.

Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW. 1999. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat Genet* **21**: 278-283.

Glaumann H, Bronner U, Ericsson O, Gustafsson LL, Rombo L. 1994. Pentamidine accumulates in rat liver lysosomes and inhibits phospholipid degradation. *Pharmacol Toxicol* **74**: 17-22.

Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, Salavati R, Cristea IM, Tavazoie S. 2012. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* **485**: 264-268.

Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, Raj B, Lin ZY, Gingras AC, Blencowe BJ. 2015. An alternative splicing event amplifies evolutionary differences between vertebrates. *Science* **349**: 868-873.

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

Gunzl A. 2010. The pre-mRNA splicing machinery of trypanosomes: complex or simplified? *Eukaryot Cell* **9**: 1159-1170.
Gupta SK, Kosti I, Plaut G, Pivko A, Tkacz ID, Cohen-Chalamish S, Biswas DK, Wachtel C, Waldman Ben-Asher H, Carmi S et al. 2013. The hnRNP F/H homologue of Trypanosoma brucei is differentially expressed in the two life cycle stages of the parasite and regulates splicing and mRNA stability. *Nucleic Acids Res* **41**: 6577-6594.
Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O et al. 2011. A protein complex network of Drosophila melanogaster. *Cell* **147**: 690-703.
Gutteridge WE. 1985. Existing chemotherapy and its limitations. *Br Med Bull* **41**: 162-168.
Haanstra JR, Kerkhoven EJ, van Tuijl A, Blits M, Wurst M, van Nuland R, Albert MA, Michels PA, Bouwman J, Clayton C et al. 2011. A domino effect in drug action: from metabolic assault towards parasite differentiation. *Mol Microbiol* **79**: 94-108.
Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**: 3-12.
Haile S, Estevez AM, Clayton C. 2003. A role for the exosome in the in vivo degradation of unstable mRNAs. *RNA* **9**: 1491-1501.
Halbig K, De Nova-Ocampo M, Cruz-Reyes J. 2004. Complete cycles of bloodstream trypanosome RNA editing in vitro. *RNA* **10**: 914-920.
Hamilton PB, Teixeira MM, Stevens JR. 2012. The evolution of Trypanosoma cruzi: the 'bat seeding' hypothesis. *Trends Parasitol* **28**: 136-141.
Hart GT, Lee I, Marcotte ER. 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8**: 236.
Hashimi H, Zimmer SL, Ammerman ML, Read LK, Lukes J. 2013. Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex. *Trends Parasitol* **29**: 91-99.
Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S et al. 2012. A census of human soluble protein complexes. *Cell* **150**: 1068-1081.
Hawking F. 1978. Suramin: with special reference to onchocerciasis. *Adv Pharmacol Chemother* **15**: 289-322.
Hehl A, Vassella E, Braun R, 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.
Hernandez A, Madina BR, Ro K, Wohlschlegel JA, Willard B, Kinter MT, Cruz-Reyes J. 2010. REH2 RNA helicase in kinetoplastid mitochondria: ribonucleoprotein complexes and essential motifs for unwinding and guide RNA (gRNA) binding. *J Biol Chem* **285**: 1220-1228.
Ho JJ, Marsden PA. 2014. Competition and collaboration between RNA-binding proteins and microRNAs. *Wiley Interdiscip Rev RNA* **5**: 69-86.
Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6**: e255.
Hope R, Ben-Mayor E, Friedman N, Voloshin K, Biswas D, Matas D, Drori Y, Gunzl A, Michaeli S. 2014. Phosphorylation of the TATA-binding protein activates the spliced leader silencing pathway in Trypanosoma brucei. *Sci Signal* **7**: ra85.

Horn D. 2008. Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics* **9**: 2.
Hotz HR, Biebinger S, Flaspohler J, Clayton C. 1998. PARP gene expression: control at many levels. *Mol Biochem Parasitol* **91**: 131-143.
Hu S, Xie Z, Onishi A, Yu X, Jiang L, Lin J, Rho HS, Woodard C, Wang H, Jeong JS et al. 2009. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**: 610-622.
Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE. 2004. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11**: 257-264.
Hudson WH, Ortlund EA. 2014. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol* **15**: 749-760.
Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**: 1205-1214.
Ivens AC Peacock CS Worthey EA Murphy L Aggarwal G Berriman M Sisk E Rajandream MA Adlem E Aert R et al. 2005. The genome of the kinetoplastid parasite, Leishmania major. *Science* **309**: 436-442.
Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC. 2007. mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes. *Proc Natl Acad Sci U S A* **104**: 2035-2042.
Jankowsky E, Harris ME. 2015. Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol* **16**: 533-544.
Jefferies D, Tebabi P, Pays E. 1991. Transient activity assays of the Trypanosoma brucei variant surface glycoprotein gene promoter: control of gene expression at the posttranscriptional level. *Mol Cell Biol* **11**: 338-343.
Jens M, Rajewsky N. 2015. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nat Rev Genet* **16**: 113-126.
Jensen BC, Ramasamy G, Vasconcelos EJ, Ingolia NT, Myler PJ, Parsons M. 2014. Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei. *BMC Genomics* **15**: 911.
Jensen BC, Sivam D, Kifer CT, Myler PJ, Parsons M. 2009. Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei. *BMC Genomics* **10**: 482.
Jha BA, Gazestani VH, Yip CW, Salavati R. 2015. The DRBD13 RNA binding protein is involved in the insect-stage differentiation process of Trypanosoma brucei. *FEBS Lett* **589**: 1966-1974.
Jordan AM. 1973. Observations on the spermatheca of tsetse flies. *Trans R Soc Trop Med Hyg* **67**: 298.
Jordan AM. 1978. Principles of the eradication or control of tsetse flies. *Nature* **273**: 607-609.
Joshi PP, Shegokar VR, Powar RM, Herder S, Katti R, Salkar HR, Dani VS, Bhargava A, Jannin J, Truc P. 2005. Human trypanosomiasis caused by Trypanosoma evansi in India: the first case report. *Am J Trop Med Hyg* **73**: 491-495.
Kabani S, Fenn K, Ross A, Ivens A, Smith TK, Ghazal P, Matthews K. 2009. Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in Trypanosoma brucei. *BMC Genomics* **10**: 427.

Kafkova L, Ammerman ML, Faktorova D, Fisk JC, Zimmer SL, Sobotka R, Read LK, Lukes J, Hashimi H. 2012. Functional characterization of two paralogs that are novel RNA binding proteins influencing mitochondrial transcripts of Trypanosoma brucei. *RNA* **18**: 1846-1861.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**: D199-205.

Kao CY, Read LK. 2005. Opposing effects of polyadenylation on the stability of edited and unedited mitochondrial RNAs in Trypanosoma brucei. *Mol Cell Biol* **25**: 1634-1644.

Kao CY, Read LK. 2007. Targeted depletion of a mitochondrial nucleotidyltransferase suggests the presence of multiple enzymes that polymerize mRNA 3' tails in Trypanosoma brucei mitochondria. *Mol Biochem Parasitol* **154**: 158-169.

Keene JD. 2007a. Biological clocks and the coordination theory of RNA operons and regulons. *Cold Spring Harb Symp Quant Biol* **72**: 157-165.

Keene JD. 2007b. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**: 533-543.

Keiser J, Ericsson O, Burri C. 2000. Investigations of the metabolites of the trypanocidal drug melarsoprol. *Clin Pharmacol Ther* **67**: 478-488.

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U et al. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**: D841-846.

Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103-107.

Khosravi P, Gazestani VH, Asgari Y, Law B, Sadeghi M, Goliaei B. 2014. Network-based approach reveals Y chromosome influences prostate cancer susceptibility. *Comput Biol Med* **54**: 24-31.

King H, Lourie EM, Yorke W. 1937. New trypanocidal substances. *Lancet* **2**: 1360-1363.

Kirkwood KJ, Ahmad Y, Larance M, Lamond AI. 2013. Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. *Mol Cell Proteomics* **12**: 3851-3873.

Kitano H. 2002. Systems biology: a brief overview. *Science* **295**: 1662-1664.

Kitano H. 2004. Biological robustness. *Nat Rev Genet* **5**: 826-837.

Kligun E, Mandel-Gutfreund Y. 2015. The role of RNA conformation in RNA-protein recognition. *RNA Biol* **12**: 720-727.

Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. 2010. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathog* **6**: e1001090.

Kolev NG, Ramey-Butler K, Cross GA, Ullu E, Tschudi C. 2012. Developmental progression to infectivity in Trypanosoma brucei triggered by an RNA-binding protein. *Science* **338**: 1352-1353.

Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909-915.

Koslowsky DJ, Bhat GJ, Perrollaz AL, Feagin JE, Stuart K. 1990. The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. *Cell* **62**: 901-911.

Kramer S. 2012. Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. *Mol Biochem Parasitol* **181**: 61-72.
Kramer S, Carrington M. 2011. Trans-acting proteins regulating mRNA maturation, stability and translation in trypanosomatids. *Trends Parasitol* **27**: 23-30.
Kramer S, Kimblin NC, Carrington M. 2010a. Genome-wide in silico screen for CCCH-type zinc finger proteins of Trypanosoma brucei, Trypanosoma cruzi and Leishmania major. *BMC Genomics* **11**: 283.
Kramer S, Queiroz R, Ellis L, Hoheisel JD, Clayton C, Carrington M. 2010b. The RNA helicase DHH1 is central to the correct expression of many developmentally regulated mRNAs in trypanosomes. *J Cell Sci* **123**: 699-711.
Kramer S, Queiroz R, Ellis L, Webb H, Hoheisel JD, Clayton C, Carrington M. 2008. Heat shock causes a decrease in polysomes and the appearance of stress granules in trypanosomes independently of eIF2(alpha) phosphorylation at Thr169. *J Cell Sci* **121**: 3002-3014.
Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP et al. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**: 637-643.
Kuepfer I, Schmid C, Allan M, Edielu A, Haary EP, Kakembo A, Kibona S, Blum J, Burri C. 2012. Safety and efficacy of the 10-day melarsoprol schedule for the treatment of second stage Rhodesiense sleeping sickness. *PLoS Negl Trop Dis* **6**: e1695.
Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**: 631-634.
Lai DH, Hashimi H, Lun ZR, Ayala FJ, Lukes J. 2008. Adaptations of Trypanosoma brucei to gradual loss of kinetoplast DNA: Trypanosoma equiperdum and Trypanosoma evansi are petite mutants of T. brucei. *Proc Natl Acad Sci U S A* **105**: 1999-2004.
Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**: 2167-2180.
Lee M, Kim B, Kim VN. 2014. Emerging roles of RNA modification: m(6)A and U-tail. *Cell* **158**: 980-987.
Leppek K, Stoecklin G. 2014. An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Res* **42**: e13.
Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, Burley SK. 2000. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100**: 323-332.
Li X, Kazan H, Lipshitz HD, Morris QD. 2014. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA* **5**: 111-130.
Li X, Quon G, Lipshitz HD, Morris Q. 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA* **16**: 1096-1107.
Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. 2016. RNA splicing is a primary link between genetic variation and disease. *Science* **352**: 600-604.
Li Z, Wang CC. 2002. Functional characterization of the 11 non-ATPase subunit proteins in the trypanosome 19 S proteasomal regulatory complex. *J Biol Chem* **277**: 42686-42693.
Lindner AK, Priotto G. 2010. The unknown risk of vertical transmission in sleeping sickness--a literature review. *PLoS Negl Trop Dis* **4**: e783.

Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. 2004. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res* **14**: 451-458.

Lizier JT, Pritam S, Prokopenko M. 2011. Information dynamics in small-world Boolean networks. *Artif Life* **17**: 293-314.

Lueong S, Merce C, Fischer B, Hoheisel JD, Erben ED. 2016. Gene expression regulatory networks in Trypanosoma brucei: insights into the role of the mRNA-binding proteome. *Mol Microbiol* **100**: 457-471.

Lukong KE, Chang KW, Khandjian EW, Richard S. 2008. RNA-binding proteins in human genetic disease. *Trends Genet* **24**: 416-425.

Luz Ambrosio D, Lee JH, Panigrahi AK, Nguyen TN, Cicarelli RM, Gunzl A. 2009. Spliceosomal proteomics in Trypanosoma brucei reveal new RNA splicing factors. *Eukaryot Cell* **8**: 990-1000.

MacGregor P, Matthews KR. 2012. Identification of the regulatory elements controlling the transmission stage-specific gene expression of PAD1 in Trypanosoma brucei. *Nucleic Acids Res* **40**: 7705-7717.

Madina BR, Kuppan G, Vashisht AA, Liang YH, Downey KM, Wohlschlegel JA, Ji X, Sze SH, Sacchettini JC, Read LK et al. 2011. Guide RNA biogenesis involves a novel RNase III family endoribonuclease in Trypanosoma brucei. *RNA* **17**: 1821-1830.

Malovannaya A, Lanz RB, Jung SY, Bulynko Y, Le NT, Chan DW, Ding C, Shi Y, Yucer N, Krenciute G et al. 2011. Analysis of the human endogenous coregulator complexome. *Cell* **145**: 787-799.

Mao Y, Najafabadi HS, Salavati R. 2009. Genome-wide computational identification of functional RNA elements in Trypanosoma brucei. *BMC Genomics* **10**: 355.

Maree JP, Patterton HG. 2014. The epigenome of Trypanosoma brucei: a regulatory interface to an unconventional transcriptional machine. *Biochim Biophys Acta* **1839**: 743-750.

Martinez-Calvillo S, Vizuet-de-Rueda JC, Florencio-Martinez LE, Manning-Cela RG, Figueroa-Angulo EE. 2010. Gene expression in trypanosomatid parasites. *J Biomed Biotechnol* **2010**: 525241.

Matia-Gonzalez AM, Laing EE, Gerber AP. 2015. Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat Struct Mol Biol* **22**: 1027-1033.

Matthews KR. 2005. The developmental cell biology of Trypanosoma brucei. *J Cell Sci* **118**: 283-290.

Matthews KR, Tschudi C, 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.

Mattiacio JL, Read LK. 2008. Roles for TbDSS-1 in RNA surveillance and decay of maturation by-products from the 12S rRNA locus. *Nucleic Acids Res* **36**: 319-329.

Maudlin I, Welburn SC. 1989. A single trypanosome is sufficient to infect a tsetse fly. *Ann Trop Med Parasitol* **83**: 431-433.

Mayho M, Fenn K, Craddy P, Crosthwaite S, 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.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673-684.

Meireles-Filho AC, Stark A. 2009. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Curr Opin Genet Dev* **19**: 565-570.
Milone J, Wilusz J, Bellofatto V. 2004. Characterization of deadenylation in trypanosome extracts and its inhibition by poly(A)-binding protein Pab1p. *RNA* **10**: 448-457.
Minning TA, Weatherly DB, Atwood J, 3rd, Orlando R, Tarleton RL. 2009. The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. *BMC Genomics* **10**: 370.
Monk SL, Simmonds P, Matthews KR. 2013. A short bifunctional element operates to positively or negatively regulate ESAG9 expression in different developmental forms of Trypanosoma brucei. *J Cell Sci* **126**: 2294-2304.
Mony BM, MacGregor P, Ivens A, Rojas F, Cowton A, Young J, Horn D, Matthews K. 2014. Genome-wide dissection of the quorum sensing signalling pathway in Trypanosoma brucei. *Nature* **505**: 681-685.
Morgulis A, Gertz EM, Schaffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**: 1028-1040.
Moura DM, Reis CR, Xavier CC, da Costa Lima TD, Lima RP, Carrington M, de Melo Neto OP. 2015. Two related trypanosomatid eIF4G homologues have functional differences compatible with distinct roles during translation initiation. *RNA Biol* **12**: 305-319.
Najafabadi HS, Lu Z, MacPherson C, Mehta V, Adoue V, Pastinen T, Salavati R. 2013. Global identification of conserved post-transcriptional regulatory programs in trypanosomatids. *Nucleic Acids Res* **41**: 8591-8600.
Nash TA, Jordan AM, Trewern MA. 1972. The Langford colonies of tsetse flies. *Trans R Soc Trop Med Hyg* **66**: 308-309.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.
Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* **9**: 471-472.
Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. *PLoS Pathog* **6**: e1001037.
Norbury CJ. 2013. Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Mol Cell Biol* **14**: 643-653.
Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T et al. 2011. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**: 296-309.
Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133**: 1277-1289.
Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH. 2006. Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat Struct Mol Biol* **13**: 160-167.
Orchard S, Albar JP, Deutsch EW, Eisenacher M, Binz PA, Hermjakob H. 2010. implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. *Proteomics* **10**: 1895-1898.

Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**: D196-203.

Panigrahi AK, Ogata Y, Zikova A, Anupama A, Dalley RA, Acestor N, Myler PJ, Stuart KD. 2009. A comprehensive analysis of Trypanosoma brucei mitochondrial proteome. *Proteomics* **9**: 434-450.

Parrish JR, Gulyas KD, Finley RL, Jr. 2006. Yeast two-hybrid contributions to interactome mapping. *Curr Opin Biotechnol* **17**: 387-393.

Pepin J, Milord F. 1994. The treatment of human African trypanosomiasis. *Adv Parasitol* **33**: 1-47.

Perrin R, Lange H, Grienenberger JM, Gagliardi D. 2004a. AtmtPNPase is required for multiple aspects of the 18S rRNA metabolism in Arabidopsis thaliana mitochondria. *Nucleic Acids Res* **32**: 5174-5182.

Perrin R, Meyer EH, Zaepfel M, Kim YJ, Mache R, Grienenberger JM, Gualberto JM, Gagliardi D. 2004b. Two exoribonucleases act sequentially to process mature 3'-ends of atp9 mRNAs in Arabidopsis mitochondria. *J Biol Chem* **279**: 25440-25446.

Proto WR, Coombs GH, Mottram JC. 2013. Cell death in parasitic protozoa: regulated or incidental? *Nat Rev Microbiol* **11**: 58-66.

Queiroz R, Benz C, Fellenberg K, Hoheisel JD, Clayton C. 2009. Transcriptome analysis of differentiating trypanosomes reveals the existence of multiple post-transcriptional regulons. *BMC Genomics* **10**: 495.

Quijada L, Guerra-Giraldez C, Drozdz M, Hartmann C, Irmer H, Ben-Dov C, Cristodero M, Ding M, Clayton C. 2002. Expression of the human RNA-binding protein HuR in Trypanosoma brucei increases the abundance of mRNAs containing AU-rich regulatory elements. *Nucleic Acids Res* **30**: 4414-4424.

Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A et al. 2014. The binary protein-protein interaction landscape of Escherichia coli. *Nat Biotechnol* **32**: 285-290.

Rassi A, Jr., Rassi A, Marin-Neto JA. 2010. Chagas disease. *Lancet* **375**: 1388-1402.

Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol* **27**: 667-670.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172-177.

Read LK, Lukes J, Hashimi H. 2016. Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley Interdiscip Rev RNA* **7**: 33-51.

Read LK, Myler PJ, Stuart K. 1992. Extensive editing of both processed and preprocessed maxicircle CR6 transcripts in Trypanosoma brucei. *J Biol Chem* **267**: 1123-1128.

Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. 2016. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* **44**: W83-89.

Reynolds D, Cliffe L, Forstner KU, Hon CC, Siegel TN, Sabatini R. 2014. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei. *Nucleic Acids Res* **42**: 9717-9729.

Rico E, Rojas F, Mony BM, Szoor B, Macgregor P, Matthews KR. 2013. Bloodstream form pre-adaptation to the tsetse fly in Trypanosoma brucei. *Front Cell Infect Microbiol* **3**: 78.
Ridlon L, Skodova I, Pan S, Lukes J, Maslov DA. 2013. The importance of the 45 S ribosomal small subunit-related complex for mitochondrial translation in Trypanosoma brucei. *J Biol Chem* **288**: 32963-32978.
Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. 2009. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. *Mol Biochem Parasitol* **165**: 32-47.
Rorbach J, Minczuk M. 2012. The post-transcriptional life of mammalian mitochondrial RNA. *Biochem J* **444**: 357-373.
Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2010. CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res* **38**: D497-501.
Rusche LN, Huang CE, Piller KJ, Hemann M, Wirtz E, Sollner-Webb B. 2001. The two RNA ligases of the Trypanosoma brucei RNA editing complex: cloning the essential band IV gene and identifying the band V gene. *Mol Cell Biol* **21**: 979-989.
Ryan CM, Militello KT, Read LK. 2003. Polyadenylation regulates the stability of Trypanosoma brucei mitochondrial RNAs. *J Biol Chem* **278**: 32753-32762.
Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to Cytoscape plugins. *Nat Methods* **9**: 1069-1076.
Schnaufer A, Domingo GJ, Stuart K. 2002. Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. *Int J Parasitol* **32**: 1071-1084.
Schnaufer A, Panigrahi AK, Panicucci B, Igo RP, Jr., Wirtz E, Salavati R, Stuart K. 2001. An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei. *Science* **291**: 2159-2162.
Schnaufer A, Wu M, Park YJ, Nakai T, Deng J, Proff R, Hol WG, Stuart KD. 2010. A protein-protein interaction map of trypanosome ~20S editosomes. *J Biol Chem* **285**: 5282-5295.
Schuster G, Stern D. 2009. RNA polyadenylation and decay in mitochondria and chloroplasts. *Prog Mol Biol Transl Sci* **85**: 393-422.
Schwede A, Kramer S, Carrington M. 2012. How do trypanosomes change gene expression in response to the environment? *Protoplasma* **249**: 223-238.
Shapiro TA, Englund PT. 1990. Selective cleavage of kinetoplast DNA minicircles promoted by antitrypanosomal drugs. *Proc Natl Acad Sci U S A* **87**: 950-954.
Shapiro TA, Englund PT. 1995. The structure and replication of kinetoplast DNA. *Annu Rev Microbiol* **49**: 117-143.
Sharan R, Ulitsky I, Shamir R. 2007. Network-based prediction of protein function. *Mol Syst Biol* **3**: 88.
Shateri Najafabadi H, Salavati R. 2010. Functional genome annotation by combined analysis across microarray studies of Trypanosoma brucei. *PLoS Negl Trop Dis* **4**.
Shlomai J. 2004. The structure and replication of kinetoplast DNA. *Curr Mol Med* **4**: 623-647.
Siegel TN, Hekstra DR, Kemp LE, Figueiredo LM, Lowell JE, Fenyo D, Wang X, Dewell S, Cross GA. 2009. Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. *Genes Dev* **23**: 1063-1076.
Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GA. 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.

Silva MA. 1957. The value of drugs commonly used in the treatment of T. rhodesiense sleeping sickness. *An Inst Med Trop (Lisb)* **14**: 159-170.
Silverman IM, Li F, Alexander A, Goff L, Trapnell C, Rinn JL, Gregory BD. 2014. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol* **15**: R3.
Simarro PP, Jannin J, Cattand P. 2008. Eliminating human African trypanosomiasis: where do we stand and what comes next? *PLoS Med* **5**: e55.
Simpson L. 1987. The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution. *Annu Rev Microbiol* **41**: 363-382.
Slevin MK, Meaux S, Welch JD, Bigler R, Miliani de Marval PL, Su W, Rhoads RE, Prins JF, Marzluff WF. 2014. Deep sequencing shows multiple oligouridylations are required for 3' to 5' degradation of histone mRNAs on polyribosomes. *Mol Cell* **53**: 1020-1030.
Slobodin B, Gerst JE. 2010. A novel mRNA affinity purification technique for the identification of interacting proteins and transcripts in ribonucleoprotein complexes. *RNA* **16**: 2277-2290.
Slomovic S, Schuster G. 2008. Stable PNPase RNAi silencing: its effect on the processing and adenylation of human mitochondrial RNA. *RNA* **14**: 310-323.
Slomovic S, Schuster G. 2013. Circularized RT-PCR (cRT-PCR): analysis of the 5' ends, 3' ends, and poly(A) tails of RNA. *Methods Enzymol* **530**: 227-251.
Souza AE, Myler PJ, Stuart K. 1992. Maxicircle CR1 transcripts of Trypanosoma brucei are edited and developmentally regulated and encode a putative iron-sulfur protein homologous to an NADH dehydrogenase subunit. *Mol Cell Biol* **12**: 2100-2107.
Steinmann P, Stone CM, Sutherland CS, Tanner M, Tediosi F. 2015. Contemporary and emerging strategies for eliminating human African trypanosomiasis due to Trypanosoma brucei gambiense: review. *Trop Med Int Health* **20**: 707-718.
Stelzl U. 2014. E. coli network upgrade. *Nat Biotechnol* **32**: 241-243.
Steuer R, Kurths J, Daub CO, Weise J, Selbig J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18 Suppl 2**: S231-240.
Stewart ML, Burchmore RJ, Clucas C, Hertz-Fowler C, Brooks K, Tait A, Macleod A, Turner CM, De Koning HP, Wong PE et al. 2010. Multiple genetic mechanisms lead to loss of functional TbAT1 expression in drug-resistant trypanosomes. *Eukaryot Cell* **9**: 336-343.
Stich A, Abel PM, Krishna S. 2002. Human African trypanosomiasis. *BMJ* **325**: 203-206.
Stich A, Ponte-Sucre A, Holzgrabe U. 2013. Do we need new drugs against human African trypanosomiasis? *Lancet Infect Dis* **13**: 733-734.
Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**: 249-255.
Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. 2005. Complex management: RNA editing in trypanosomes. *Trends Biochem Sci* **30**: 97-105.
Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66-71.
Tadros W, Goldman AL, Babak T, Menzies F, Vardy L, Orr-Weaver T, Hughes TR, Westwood JT, Smibert CA, Lipshitz HD. 2007. SMAUG is a major regulator of maternal mRNA destabilization in Drosophila and its translation is activated by the PAN GU kinase. *Dev Cell* **12**: 143-155.
Temperley RJ, Seneca SH, Tonska K, Bartnik E, Bindoff LA, Lightowlers RN, Chrzanowska-Lightowlers ZM. 2003. Investigation of a pathogenic mtDNA microdeletion reveals a

translation-dependent deadenylation decay pathway in human mitochondria. *Hum Mol Genet* **12**: 2341-2348.
Thuita JK, Kagira JM, Mwangangi D, Matovu E, Turner CM, Masiga D. 2008. Trypanosoma brucei rhodesiense transmitted by a single tsetse fly bite in vervet monkeys as a model of human African trypanosomiasis. *PLoS Negl Trop Dis* **2**: e238.
Traub N, Hira PR, Chintu C, Mhango C. 1978. Congenital trypanosomiasis: report of a case due to Trypanosoma brucei rhodesiense. *East Afr Med J* **55**: 477.
Ueno N, Lodoen MB. 2015. From the blood to the brain: avenues of eukaryotic pathogen dissemination to the central nervous system. *Curr Opin Microbiol* **26**: 53-59.
Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, Lowe TM, Salama SR, Haussler D. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7**: 995-1001.
Urbaniak MD, Guther ML, Ferguson MA. 2012. Comparative SILAC proteomic analysis of Trypanosoma brucei bloodstream and procyclic lifecycle stages. *PLoS One* **7**: e36619.
Urbaniak MD, Martin DM, Ferguson MA. 2013. Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of Trypanosoma brucei. *J Proteome Res* **12**: 2233-2244.
van Kouwenhove M, Kedde M, Agami R. 2011. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat Rev Cancer* **11**: 644-656.
Vansterkenburg EL, Coppens I, Wilting J, Bos OJ, Fischer MJ, Janssen LH, Opperdoes FR. 1993. The uptake of the trypanocidal drug suramin in combination with low-density lipoproteins by Trypanosoma brucei and its possible mode of action. *Acta Trop* **54**: 237-250.
Vasquez JJ, Hon CC, Vanselow JT, Schlosser A, Siegel TN. 2014. Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages. *Nucleic Acids Res* **42**: 3623-3637.
Veitch NJ, Johnson PC, Trivedi U, Terry S, Wildridge D, MacLeod A. 2010. Digital gene expression analysis of two life cycle stages of the human-infective parasite, Trypanosoma brucei gambiense reveals differentially expressed clusters of co-regulated genes. *BMC Genomics* **11**: 124.
von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399-403.
Waalkes TP, Denham C, DeVita VT. 1970. Pentamidine: clinical pharmacologic correlations in man and mice. *Clin Pharmacol Ther* **11**: 505-512.
Walker SC, Scott FH, Srisawat C, Engelke DR. 2008. RNA affinity tags for the rapid purification and investigation of RNAs and RNA-protein complexes. *Methods Mol Biol* **488**: 23-40.
Walrad P, Paterou A, Acosta-Serrano A, Matthews KR. 2009. Differential trypanosome surface coat regulation by a CCCH protein that co-associates with procyclin mRNA cis-elements. *PLoS Pathog* **5**: e1000317.
Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A et al. 2015. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**: 339-344.
Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. 2011. Understanding the transcriptome through RNA structure. *Nat Rev Genet* **12**: 641-655.

Wang CC. 1995. Molecular mechanisms and therapeutic approaches to the treatment of African trypanosomiasis. *Annu Rev Pharmacol Toxicol* **35**: 93-127.
Wang X, McLachlan J, Zamore PD, Hall TM. 2002. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**: 501-512.
Wardrop NA, Atkinson PM, Gething PW, Fevre EM, Picozzi K, Kakembo AS, Welburn SC. 2010. Bayesian geostatistical analysis and prediction of Rhodesian human African trypanosomiasis. *PLoS Negl Trop Dis* **4**: e914.
Wardrop NA, Fevre EM, Atkinson PM, Kakembo A, Welburn SC. 2012. An exploratory GIS-based method to identify and characterise landscapes with an elevated epidemiological risk of Rhodesian human African trypanosomiasis. *BMC Infect Dis* **12**: 316.
Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet* **26**: 66-74.
Welburn SC, Maudlin I. 1992. The nature of the teneral state in Glossina and its role in the acquisition of trypanosome infection in tsetse. *Ann Trop Med Parasitol* **86**: 529-536.
Welch JD, Slevin MK, Tatomer DC, Duronio RJ, Prins JF, Marzluff WF. 2015. EnD-Seq and AppEnD: sequencing 3' ends to identify nontemplated tails and degradation intermediates. *RNA* **21**: 1375-1389.
Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**: 582-587.
Williamson J. 1962. Chemotherapy and chemoprophylaxis of African trypanosomiasis. *Exp Parasitol* **12**: 323-367.
Windbichler N, Schroeder R. 2006. Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. *Nat Protoc* **1**: 637-640.
Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**: 59-69.
World Health O. 2012. Research priorities for Chagas disease, human African trypanosomiasis and leishmaniasis. *World Health Organ Tech Rep Ser*: v-xii, 1-100.
World Health O. 2013. Control and surveillance of human African trypanosomiasis. *World Health Organ Tech Rep Ser*: 1-237.
Wurst M, Seliger B, Jha BA, Klein C, Queiroz R, Clayton C. 2012. Expression of the RNA recognition motif protein RBP10 promotes a bloodstream-form transcript pattern in Trypanosoma brucei. *Mol Microbiol* **83**: 1048-1063.
Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M. 2011. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* **21**: 1450-1461.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**: 338-345.
Yan X, Mehan MR, Huang Y, Waterman MS, Yu PS, Zhou XJ. 2007. A graph-based approach to systematically reconstruct human transcriptional regulatory modules. *Bioinformatics* **23**: i577-586.
Yousefi M, Hajihoseini V, Jung W, Hosseinpour B, Rassouli H, Lee B, Baharvand H, Lee K, Salekdeh GH. 2012. Embryonic stem cell interactomics: the beginning of a long road to biological function. *Stem Cell Rev* **8**: 1138-1154.

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N et al. 2008. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**: 104-110.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**: 614-620.

Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP. 2005. Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *J Biol* **4**: 6.

Zhang X, Virtanen A, Kleiman FE. 2010. To polyadenylate or to deadenylate: that is the question. *Cell Cycle* **9**: 4437-4449.

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT. 2010. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**: 939-953.

Zheng D, Tian B. 2014. Sizing up the poly(A) tail: insights from deep sequencing. *Trends Biochem Sci* **39**: 255-257.

Zimmer SL, McEvoy SM, Menon S, Read LK. 2012. Additive and transcript-specific effects of KPAP1 and TbRND activities on 3' non-encoded tail characteristics and mRNA stability in Trypanosoma brucei. *PLoS One* **7**: e37639.

Zimmer SL, Schein A, Zipor G, Stern DB, Schuster G. 2009. Polyadenylation in Arabidopsis and Chlamydomonas organelles: the input of nucleotidyltransferases, poly(A) polymerases and polynucleotide phosphorylase. *Plant J* **59**: 88-99.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**: 65-70.

Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q. 2013. GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41**: W115-122.