GENE CO-EXPRESSION NETWORK OF *Trypanosoma brucei* DEVELOPMENTAL STAGES IN THE TSETSE FLY VECTOR G*lossina morsitans*

KENNEDY WANJAU MWANGI

MASTER OF SCIENCE
(Molecular Biology & Bioinformatics)

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY

2022

Gene Co-expression Network of *Trypanosoma brucei* Developmental Stages in the Tsetse Fly Vector *Glossina morsitans*

Kennedy Wanjau Mwangi

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Molecular Biology and Bioinformatics of the Jomo Kenyatta University of Agriculture and Technology

2022

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature……..Date…….

Kennedy Wanjau Mwangi

This thesis has been presented for examination with our approval as University supervisors.

Signature……..Date…….

Dr. Joel Bargul, PhD

JKUAT, Kenya

Signature……Date……..

Dr. Rosaline Macharia, PhD

UoN, Kenya

DEDICATION

I dedicate this thesis to my parents Patrick Mwangi and Lucy Mwangi and my siblings Michael Mwangi and Graceann Mwangi. Thank you for your sacrifice, patience, and encouragement throughout my studies. May the Almighty God bless you abundantly.

ACKNOWLEDGEMENTS

My sincere gratitude to my supervisors Dr. Joel Bargul and Dr. Rosaline Macharia for their immense support right from conception of the project idea, analysis, manuscript preparation, and writing of the thesis. I reckon that this work was truly a great learning experience and I am a better scientist. Many thanks to Dr. Joel Bargul for funding this work without which it would not have been possible to undertake. I gratefully acknowledge *icipe* where this work was carried out and JKUAT where I am registered as a student. I am also thankful to Dr. Caleb Kibet for facilitating my use of the *icipe* HPC cluster to carry out my work and insightful comments on my work. I thank Dr. Dan Masiga and Dr. Jandouwe Villinger for their insights offered on my work during the Bioinformatics team meetings. I acknowledge my colleagues at *icipe*'s Molecular Biology and Bioinformatics Unit; Pauline, Gilbert, and Careen for cheering me up when the analysis pipeline constantly broke down. Above all, I am eternally grateful to Our Lord Jesus Christ without whom we can do nothing: "*quia sine me nihil potestis facere*" (*for without Me you can do nothing*) – John 15:5.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

AAT	Animal African Trypanosomiasis
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
BARP	Brucei Alanine Rich Protein
BIND	Biomolecular Interaction Network Database
BIPS	Biana Interolog Prediction Server
BRENDA	Braunschweig Enzyme Database
CDC	Center for Disease Control and Prevention
ChIP	Chromatin Immunoprecipitation
DALYs	Disability-adjusted life years
DEG	Differentially Expressed Genes
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic Acid
ES	Ectoperitrophic Matrix
FIRE	Finding Informative Regulatory Elements
GBA	Guilt by Association
GCN	Gene Co-expression Network
GO	Gene Ontology
HAT	Human African Trypanosomiasis
NCBI	National Center for Biotechnology Information
PC	Procyclic Trypomastigote

PCA	Principal Component Analysis
PIPE	Protein-Protein Interaction Prediction Engine
PM	Peritrophic Matrix
PPI	Protein-Protein Interaction
PPIN	Protein-Protein Interaction Network
PRISM	Protein Interactions by Structural Matching
PUF9	Pumilio/fem-3 mRNA binding factor 9
RBP	RNA-Binding Protein
RNA	Ribonucleic Acid
RNAi	RNA interference
RNA-Seq	RNA Sequencing
RRE	RNA Regulatory Element
RRM	RNA Regulatory Motif
SAT	Sequential Aerosol Technique
SIT	Sterile Insect Technique
SOM	Self-Organizing Map
SRA	Sequence Read Archive
TIGRESS	Trustful Inference of Gene Regulation with Stability Selection
TOM	Topological Overlap Matrix
TRN	Transcription Regulatory Network
UTR	Untranslated Region
Y2H	Yeast two-hybrid

YLD	Years Lost to Disability
YLL	Years of Life Lost
VSG	Variant Surface Glycoprotein
WGCNA	Weighted Gene Co-expression Network
WHO	World Health Organization

ABSTRACT

Trypanosoma brucei causes both Human African Trypanosomiasis and Animal African Trypanosomiasis. These diseases are transmitted by tsetse flies through saliva infected with *T. brucei* as the vector feeds on a blood meal. One strategy of controlling disease transmission is disrupting the life cycle of *T. brucei* in the tsetse fly. Ongoing studies on *Sodalis glossinidius* have provided the proof-of-concept that tsetse endosymbionts may be used to interfere with pathogen transmission through the vector. This strategy, however, relies on the knowledge of trypanosome biology, especially the developmental events that take place in the tsetse fly vector to allow for its survival. This study aimed at constructing a gene co-expression network to predict key genes in *T. brucei* development in the tsetse fly, the functional roles which these genes are associated with, and also predict 3' untranslated region motifs for gene clustering together in the network. RNA-seq counts data generated from the developing *T. brucei* parasite in the tsetse fly was used in this study. The expression levels of *T. brucei* genes were obtained by running the RNA-seq data through the RNA-seq pipeline. Using the *T. brucei* gene expression data, the weighted gene co-expression network analysis approach was used to construct and analyze the network. Twelve (12) out of 27 functionally enriched gene modules (clusters) of the co-expression network were obtained from the network analysis. The enriched functional roles for the clusters were associated with cell cycle, cell signaling, mitochondrion, protein biosynthesis, and cell surface and highlight important functional processes during the parasite's development on tsetse fly. The hub (key) genes for the 12 modules encoded proteins such as RBP6, Inner arm dynein 5-1 protein, and BARP protein, that have previously been proven crucial in *T. brucei* development in the tsetse fly. The hub genes may be involved in key processes that enable the parasite develop and complete its life cycle in tsetse fly. Other hub genes encoded proteins whose functional roles are still unknown and could serve as candidate genes for further studies. The 3' untranslated region motif prediction for genes clustered together identified 10 significantly enriched motifs that could provide insights into gene regulation during parasite's development in tsetse. The results of this study provide a resource for network-based data mining to identify candidate genes for functional studies. The knowledge obtained from co-expression analysis will provide novel insights on the role of genes in development and *T. brucei* molecular processes that may be targeted by trypanocidal products.

CHAPTER ONE

INTRODUCTION

1.1 Background information

African trypanosomiasis refers to diseases caused by trypanosomes in the sub-Saharan Africa. These diseases affect humans, as in the case of Human African Trypanosomiasis (HAT), or livestock, as in the case of Animal Afrian Trypanosomiasis (AAT). The causative protozoan parasites were discovered early in the 19th century (Bruce, 1897; Dutton, 1902; Forde, 1902). The trypanosome parasites that cause HAT are *Trypanosoma brucei rhodesiense* and *Trypanosoma brucei gambiense* (Franco *et al.*, 2014) while AAT is caused by seven species of trypanosomes and the most important among them that cause disease are *Trypanosoma brucei brucei*, *Trypanosoma vivax*, and *Trypanosoma congolense* (Uilenberg & Boyt, 1998). These parasites are transmitted by several species of tsetse fly vector that belong to the genus *Glossina* which include *Glossina fuscipes fuscipes*, *Glossina morsitans morsitans*, and *Glossina pallidipes* (Krafsur, 2009).

African trypanosomiasis causes devastating losses to the population that lives in the sub-Saharan Africa. Human African trypanosomiasis causes loss of human lives and families incur huge financial cost in the course of the treatment of the disease (Shaw *et al.*, 2010). Additionally, an economic loss of about US$5 billion a year is incurred due to AAT (Angara *et al.*, 2014). To curb the effects of African trypanosomiasis in sub-Saharan Africa, the WHO and other non-governmental institutions including charitable organizations such as Bill and Melinda Gates Foundation and the Wellcome Trust have joined efforts to control the disease (Holmes, 2014).

1.2 Control of African trypanosomiasis

In the 1930s, HAT incidence in many parts of sub-Saharan Africa increased up to about 60,000 cases annually prompting the colonial governments to set up control strategies which led to a drastic decline of new cases by the 1960s (Simarro *et al.*, 2008). However, after independence, most countries neglected disease surveillance which led to

the return of the disease. By the end of the twentieth century, the World Health Organization (WHO) estimated 300,000 new cases every year (WHO, 1998). At the turn of the century, WHO led a global alliance of international and national institutions that also included the private sector to eliminate HAT (WHO, 2002). In 2012, the WHO set a goal to eliminate sleeping sickness as a public health problem by 2020 (WHO, 2012) and elimination of transmission (zero cases) of the *gambiense* form to humans by 2030 (WHO, 2013). Since all these efforts towards eliminating sleeping sickness started early in the 21st century, disease incidence has been decreasing steadily (Figure 1.1) (Franco *et al.*, 2018). In 2020, WHO reported that 977 new cases were reported in 2018 which was below the targeted 2,000 new cases by 2020, indicating that eliminating sleeping sickness as a public health problem was achievable (Franco *et al.*, 2020). Despite the decline in new cases of sleeping sickness over the years, it is estimated that about 70 million people in an area of 1.55 million km^2 in sub-Saharan Africa are at risk of contracting the disease (Simarro *et al.*, 2012).

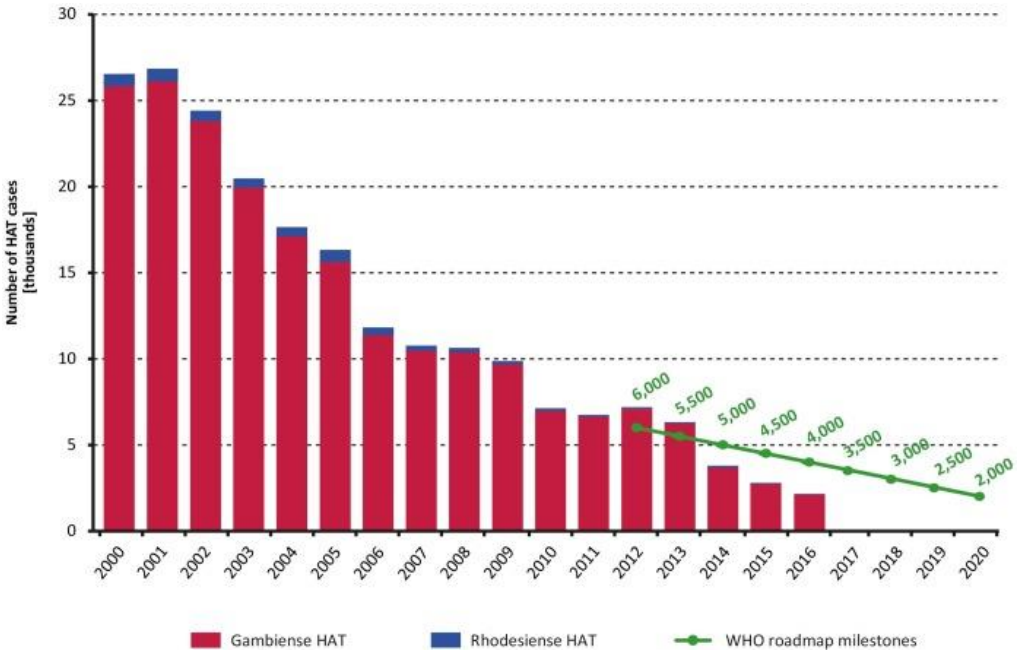


Figure 1.1: **Reported cases of *Gambiense* and *Rhodesiense* HAT per year from 2000 to 2016** (Franco *et al.*, 2018).

1.2.1 Vector control

Vector control plays a crucial role in controlling African trypanosomiasis by reducing their population and is the only preventive measure against the infectious bites from the tsetse flies. Various strategies for controlling the vectors include sequential aerosol technique (SAT), use of bait (live and artificial) technique, and sterile insect technique (SIT) (Holmes, 2013; Vreysen *et al.*, 2013). The SAT technique involves the spraying of low concentrations of insecticides at given time intervals during the reproduction cycle of the tsetse flies (Cooper & Dobson, 1993). The technique has been successful in eliminating the savannah-dwelling *G. morsitans* in southern Africa at Okavango Delta (Kgori *et al.*, 2006) and is being considered for the control of riverine tsetse (Adam *et al.*, 2013; Deken & Bouyer, 2018). However, the SAT technique needs to be adapted to fit the habitat of the riverine tsetse flies. For instance, the forested habitat, unlike the open grasslands, would interfere with the effective dispersal of the aerosolized insecticides (Deken & Bouyer, 2018).

The bait technique involves the use of stationary devices such as traps and targets or live baits such as livestock (Vreysen *et al.*, 2013). Targets attract the tsetse flies which come into contact with insecticides applied on them (Vale, 1993) while traps ensure that the tsetse flies do not escape once they enter into them (Langley *et al.*, 1990). The bait technique of using traps and target ensure that there is a reduction of the vector population. The effectiveness of the traps and targets can be enhanced by using odours and colours to attract the tsetse flies. The blue colour is used to attract the tsetse to the traps and targets, the black colour to trigger their landing, and odours to lure them into the traps (Gibson & Torr, 1999). Some of the traps developed to control tsetse flies population include the NGU trap (Brightwell *et al.*, 1987), biconical traps (Challier & Laveissiere, 1973), H traps (Kappmeier, 2000), sticky panels (Vreysen *et al.*, 1996), and F3 and Epsilon traps (Flint, 1985; Green & Flint, 1986).

Besides traps and targets, live bait can be used to control the population of the vectors. The use of live bait entails spraying livestock which are hosts to tsetse flies with insecticides to kill the vectors as they feed (Leak *et al.*, 1995). This method has

disadvantages in that the host wild animals are not sprayed with insecticides and, therefore, the tsetse flies population is sustained. Early methods of controlling tsetse flies involved destruction of the host wild animals and the clearing of bushes which were habitats for tsetse flies and were thus very effective (Dransfield *et al.*, 1991; Ford & Blaser, 1971). However, they are now unacceptable because they lead to loss of biodiversity and environmental destruction.

The SIT technique of controlling tsetse population involves releasing sterilized male tsetse flies in a target area to compete with the wild type with an aim to affect the reproductive capacity of the vector and thus lower their population (Vreysen, 2001). The SIT technique was successful in the eradication of *Glossina austeni* in Unguja Island in 1997 (Vreysen *et al.*, 2014). The success of this technique in Unguja Island proves that it can be applied successfully for other species to control their population. However, it has challenges such as difficulty in establishing sterile colonies (Enserink, 2007). The SIT technique can be used in combination with other techniques geared towards controlling the transmission of sleeping sickness. Among these techniques is the paratransgenesis technique which has been proposed as an avenue to reduce vector competence (Kariithi *et al.*, 2018).

1.2.2 Paratransgenesis

One of the control strategies of HAT would be to eliminate the parasite in the tsetse populations. Control of the parasite in the tsetse fly is an attractive strategy for disrupting the cycle of transmission of sleeping sickness because *T. brucei* becomes infective in the vector by re-acquiring the variant surface glycoprotein (VSG) coat. One technique that can be used to disrupt the parasite's development in tsetse is paratransgenesis which relies on symbiotic bacteria that live in the gut of the tsetse fly vectors (Aksoy *et al.*, 2008). The paratransgenesis technique involves the expression of anti-parasitic molecules by genetically modified symbiotic bacteria in the insect vectors (Coutinho-Abreu *et al.*, 2010). Evidence shows that the symbiotic bacteria in gut of

tsetse flies play a crucial role in modulating its immunity against trypanosome infection (Weiss *et al.*, 2013). Therefore, the gut microbiota of the tsetse flies provides an avenue to control the transmission of the parasite through the vector.

Sodalis glossinidius, a symbiotic bacterium in the tsetse fly, is under investigation for the paratransgenic approach of controlling *T. brucei* transmission through the tsetse fly. Specifically, successful transinfection of tsetse flies with *Sodalis* has been demonstrated by Weiss *et al* (2006). Additionally, genetically modified *Sodalis* has been demonstrated to pass onto adult tsetse flies after intra-larval microinjection with this *Sodalis* and, more importantly, to the subsequent progeny (De Vooght *et al.*, 2018). The success demonstrated so far in developing the paratransgenic approach in tsetse flies indicates that it is a viable technique for controlling parasite transmission through the vector. However, knowledge of trypanosome biology in the tsetse fly is central to complete the development of paratransgenic approach because it will aid in the identification of drug targets in *T. brucei* for trypanocidal products delivered by *Sodalis*. Comprehensive studies that are needed to elucidate the biology of trypanosomes during their development in the tsetse fly are yet to be conducted.

This study aimed at constructing a gene co-expression network and perform network analysis to predict the key genes in *T. brucei* development in the tsetse fly, and 3' UTR motifs for gene clustering together in the network. Information obtained in this study provides insight into molecular mechanisms underlying the parasite's development in the tsetse fly. The critical molecular processes involved in the development of the parasite in the fly may be targeted by trypanocidal products to control its transmission.

1.3 Problem statement

Human African trypanosomiasis (HAT) is a neglected tropical disease that threatens about 70 million people in sub-Saharan Africa. In 2012, the World Health Organisation's (WHO) set a goal to eliminate HAT as a public health problem by 2020. According to WHO's report as of 2020, this goal was within reach, with 977 cases

reported in 2018 against a target of 2,000 cases by 2020 (Franco *et al.*, 2020). However, the elimination of transmission for the *gambiense* form of the disease is targeted for 2030 and requires innovative disease control strategies because it is chronic and remains undetected thus contributing to flare ups of trypanosomiasis. One of the strategies being studied to control transmission of *gambiense* sleeping sickness is interfering with pathogen transmission through the tsetse fly using the paratransgenesis technique (Aksoy *et al.*, 2008; De Vooght *et al.*, 2018). The success of this strategy lies in the understanding of the trypanosome biology underlying transmission dynamics in the tsetse fly. However, many of the mechanisms of trypanosome developmental biology in tsetse fly remain poorly understood. Studying the development of the trypanosomes in the tsetse fly will aid in opening up opportunities for its control through techniques such as paratransgenesis.

1.4 Justification

Paratransgenesis is a novel approach aimed at reducing vector competence and complements existing trypanosomiasis control strategies. The technique entails genetic modification of symbionts of disease vectors to act as *in vivo* drug delivery vehicles thus control parasite development in the tsetse fly. *Sodalis glossinidius* is being studied for the paratransgenic approach to control *T. brucei* transmission through the tsetse fly (De Vooght *et al.*, 2018). In parallel to these studies, drug targets in the trypanosomes are needed and therefore necessitate a comprehensive understanding of trypanosome biology in the insect vector. Gene expression and its regulation are key to elucidating molecular processes underlying each life cycle stage of the trypanosomes. Using network biology techniques, specifically gene co-expression networks, gene expression data may provide novel information of essential genes and molecular processes that may be targeted by trypanocidal products to control the infection transmission through the tsetse fly. Therefore, this study aimed to construct a gene co-expression network to predict key genes involved in *T. brucei* development in the tsetse fly. The information obtained from the study contributes to the understanding of molecular mechanisms and

processes underlying its developmental events, which may lead to the identification of molecular targets for trypanocidal products.

1.5 Hypothesis

Gene expression contributes to developmental changes in *T. brucei* life cycle stages in the tsetse fly.

1.6 Objectives

1.6.1 Main objective

To characterize *Trypanosoma brucei* genes involved in the developmental stages of the parasite in the tsetse fly vector *Glossina morsitans* using a gene co-expression network.

1.6.2 Specific objectives

1. To develop a gene co-expression network of *T. brucei* genes based on RNA-seq gene expression profile data.
2. To identify and characterize gene modules and the hub genes involved in the life cycle of *T. brucei* in the tsetse fly vector.
3. To identify potential regulatory motifs in the 3' untranslated regions (UTR) for the *T. brucei* genes in the same module.

CHAPTER TWO

LITERATURE REVIEW

2.1 The burden of African trypanosomiasis

The burden of a disease is its impact on the individual, families, or the society and can be used as a basis for its control and other intervention measures (Murray & Lopez, 1996). Quantifying the burden of a disease is important because it enables the government and other stakeholders formulate the intervention measures. The impact of a disease on an individual is used as the basis of quantifying its burden. Specifically, it is measured in terms of disability-adjusted life years (DALYs) lost by an individual suffering from the disease (Murray, 1994). The DALYs is a combination of years lost due to disability (YLD) and years of life lost (YLL) due to premature mortality (Murray, 1994).

In the early 2000s, the estimated DALYs due to untreated HAT was 30 years per death in Angola (Schmid, 2004) and 27 years in the Democratic Republic of Congo (Lutumba *et al.*, 2005). Current DALYs in different disease foci are yet to be estimated. Globally, as of 2016, the total DALYs due to HAT was 128,400 years which is a decrease from 539,000 years estimated in 2006 (Hay *et al.*, 2017). However, the decline in global DALYs does not necessarily reflect corresponding decrease in the impact at the community level because the disease is clustered in different foci which may experience flare-ups. For example, in Zambia, an increase in the total number of DALYs due to rHAT was registered between 2004 and 2014 (Mwiinde *et al.*, 2017).

The nature of the DALYs measure implies that the impact of the disease on the individual will also have consequences to their families and society at large. An important impact of the disease on the family and society is the economic burden associated with controlling the disease and treating the patients (Fèvre *et al.*, 2008). The financial cost for treatment incurred by individuals and their families can be as high as 10 months' worth of income (Shaw *et al.*, 2010). Additionally, the HAT control and elimination programmes in the sub-Saharan Africa are costly. It is projected that either of these programmes will cost about US$1 billion between 2020 and 2030 (Sutherland &

Tediosi, 2019). Besides the impact of African trypanosomiasis on human health, the AAT disease also affects livestock, which are a source of livelihood for the population living in the sub-Saharan Africa. The economic loss associated with AAT is estimated to be around US$5 billion a year and at least $30 million is spent on treating the disease in livestock (Angara *et al.*, 2014). The impact of African trypanosomiasis on the population in the sub-Saharan Africa indicates that concerted effort is needed to control these diseases and eventually eliminate them.

2.2 Vectors of African trypanosomiasis

African trypanosomiasis is transmitted by tsetse flies (Diptera: *Glossinidae*) (Figure 2.1). Depending on their habitat in the sub-Saharan Africa, tsetse flies are classified into three sub-groups; palpalis (riverine), morsitans (savannah), and fusca (forest-dwelling) (Cecchi *et al.*, 2008) (Figure 2.2). *Glossina palpalis spp* and *Glossina fuscipes spp* are the main vectors of *T. b. gambiense.* The vectors for *T. b. rhodesiense* are the forest-dwelling *Glossina fuscipes fuscipes* in Uganda and *Glossina morsitans morsitans* and *Glossina pallidipes* which both live in the savannah (Shereni *et al.*, 2016).

Figure 2.1: **A pregnant female tsetse fly.** (Photo: with permission from Geoffrey M. Attardo, Yale School of Public Health, Yale University, USA)

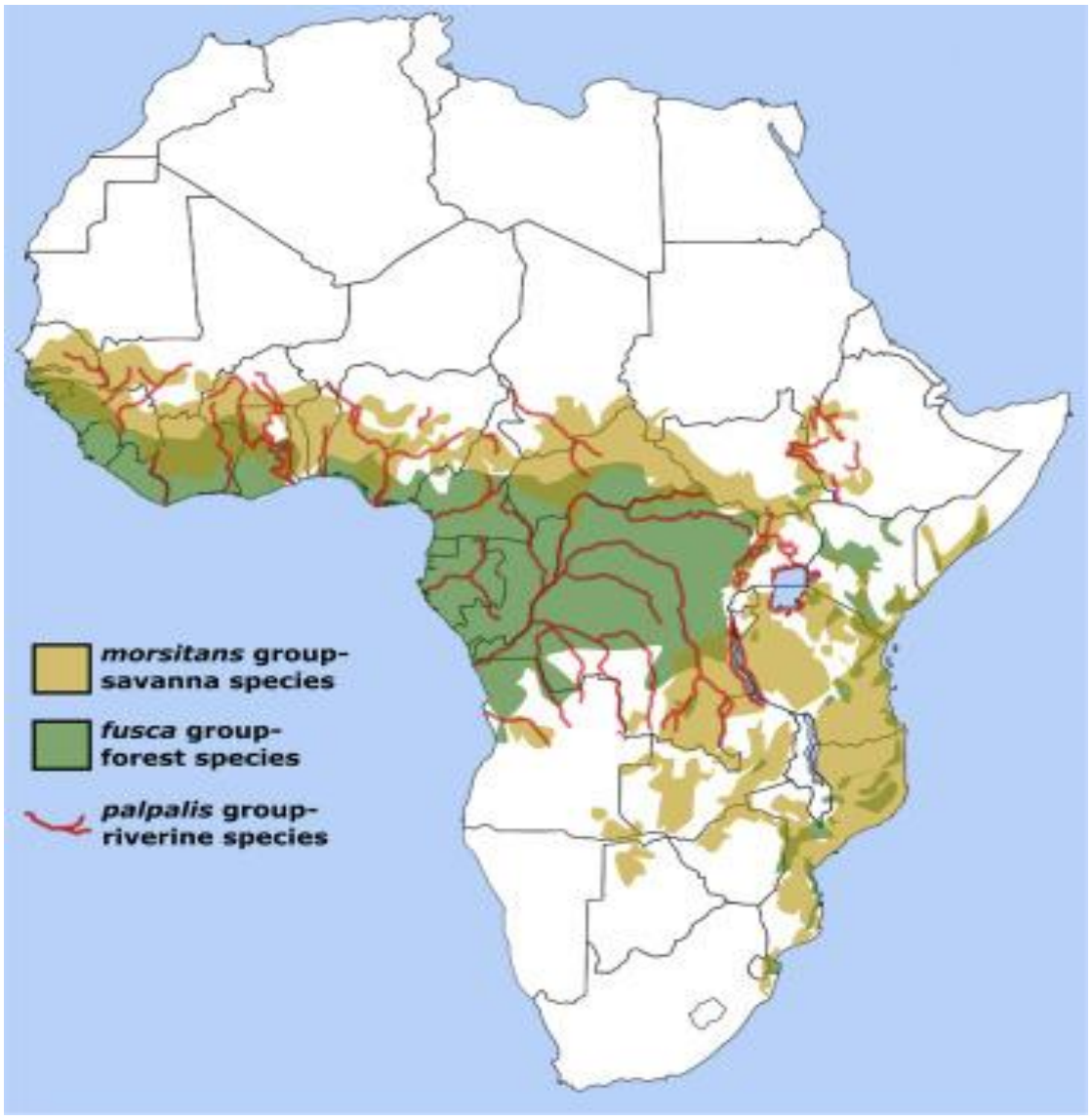


Figure 2.2: **The distribution of tsetse fly species across the sub-Saharan Africa. The tsetse flies are distributed in the savannah grasslands of East Africa, and the forests and rivers of Central and West Africa** (Krinsky, 2019).

2.3 Parasites causing African trypanosomiasis

African trypanosomes (Figure 2.3) are protozoan parasites that cause human African trypanosomiasis (HAT) and animal African trypanosomiasis (AAT) (Bruce, 1897; Dutton, 1902; Forde, 1902). There are three subspecies of *Trypanosoma brucei* which

include: *Trypanosoma brucei brucei*, *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense.* Human African trypanosomiasis, also known as sleeping sickness, is caused by *T. b. gambiense* in Central and West Africa and *T. b. rhodesiense* in East and Southern Africa (Figure 1.1) (Fèvre *et al.*, 2006).

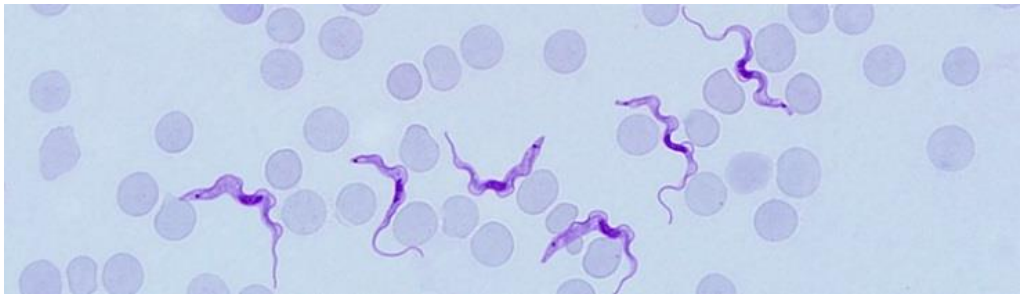


Figure 2.3: **Image of *T. b. rhodesiense* species (trypanosome parasites) that cause Human African trypanosomiasis as observed in a blood smear.** (CDC, 2020).

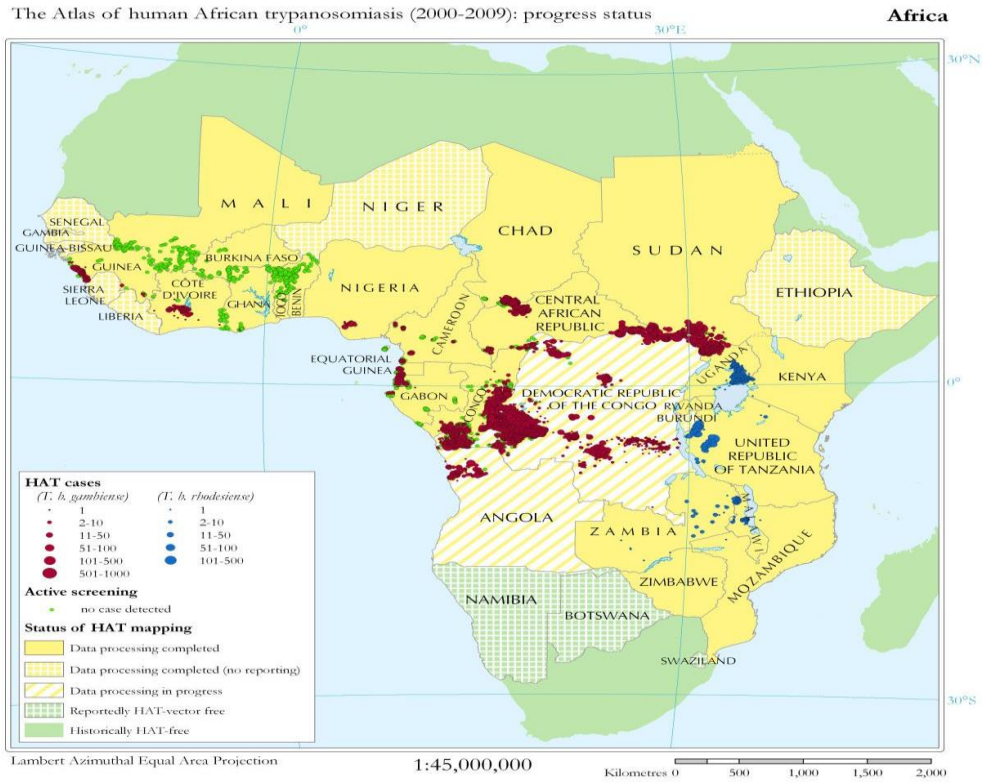


Figure 2.4: **Distribution of *T. b. gambiense* and *T. b. rhodesiense* species that transmit HAT** (Simarro *et al.*, 2010).

T. b. rhodesiense causes a more virulent form of HAT while *T. b. gambiense* HAT is chronic and progresses over the years (Franco *et al.*, 2014). Animal African trypanosomiasis is caused by either of the seven other species of trypanosomes including *T. b. brucei*, *T. vivax* and *T. congolense*, which are the major pathogens that cause AAT (nagana) in cattle (Ooi & Bastin, 2013). Both HAT and AAT are considered fatal if left untreated.

2.4 The life cycle of *Trypanosoma brucei*

Trypanosoma brucei undergoes developmental changes that mark important life cycle stages in the tsetse fly and mammalian host (Figure 2.1). These changes enable the parasite to adapt to the changing environment in the hosts for its survival (Vickerman, 1985).

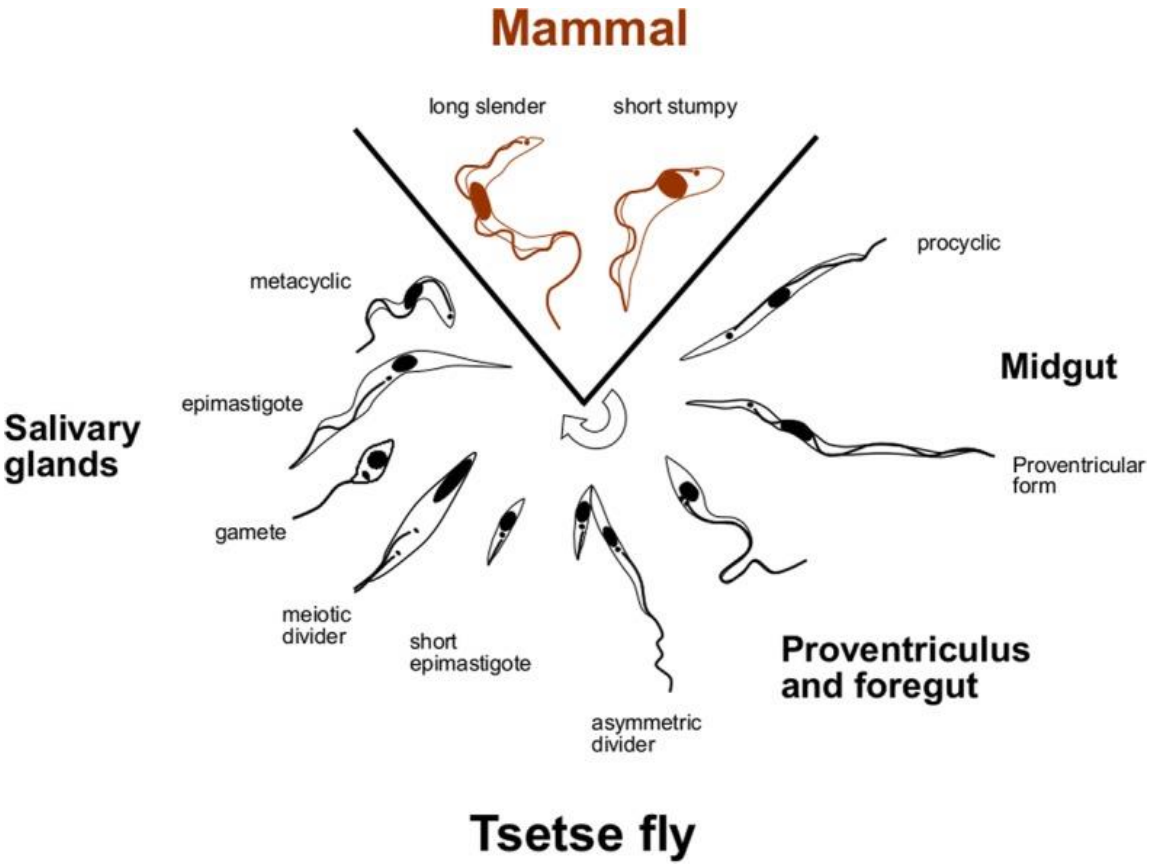


Figure 2.5: **Life cycle stages of *Trypanosoma brucei*** (Gibson & Peacock, 2019).

2.4.1 Life cycle of *Trypanosoma brucei* in the mammalian host

Infection in the mammal begins with the introduction of metacyclic trypomastigotes from the tsetse's saliva as it feeds on a blood meal (Awuoche, 2012). The trypanosomes proliferate as slender forms in the bloodstream. A yet-to-be established proportion of slender forms proceed to develop into non-proliferative stumpy forms (Rico *et al.*, 2013). The stumpy forms are arrested at G_0-G_1 phase of the cell cycle which ensures successful transmission to the tsetse fly during feeding and also reduction in the number of proliferative slender forms to avoid killing the mammalian host (Matthews, 2005). In general, the life cycle stages in the mammalian host are characterized by antigenic variation and slender-to-stumpy transitions which ensures chronicity of infection and optimal potential for parasite transmission (MacGregor *et al.*, 2012; Matthews *et al.*, 2015).

2.4.2 Life cycle of *Trypanosoma brucei* in the tsetse fly

Infection in the tsetse fly begins with the fly feeding on an infected blood meal from a mammalian host. The blood meal is composed of long slender and short stumpy bloodstream-form trypomastigotes. The short stumpy bloodstream trypomastigotes differentiate into procyclic trypomastigotes a few hours later in the posterior midgut while the long slender trypomastigotes die off. The procyclic trypomastigotes (PC) establish the infection in the midgut of the fly (Ooi & Bastin, 2013). Some PCs cross the peritrophic matrix (PM) (figure 2.2) and enter into the ectoperitrophic space (ES) where they elongate to become non-proliferative mesocyclic trypomastigotes. They migrate to the anterior midgut, thus crossing the PM the second time to enter the proventriculus. In the proventriculus, they change their morphology and become thinner and the nucleus migrates to the posterior side of the kinetoplast (Rotureau & Van Den Abbeele, 2013). This morphology is known as the epimastigote form. The epimastigotes divide asymmetrically to give long and short epimastigotes (Ooi & Bastin, 2013). A few of the

short epimastigotes migrate to the salivary glands where they attach to the epithelium via the flagellum, elongate, and acquire brucei alanine-rich protein (BARP) coat whose role is not known (Urwyler *et al.*, 2007). The short epimastigotes mature and form metacyclic trypomastigotes that reacquire the variant surface glycoprotein (VSG) coat (Tetley *et al.*, 1987). The metacyclic trypomastigotes are transmitted to the mammalian host through saliva while the fly is feeding thus completing the transmission cycle in the insect vector.

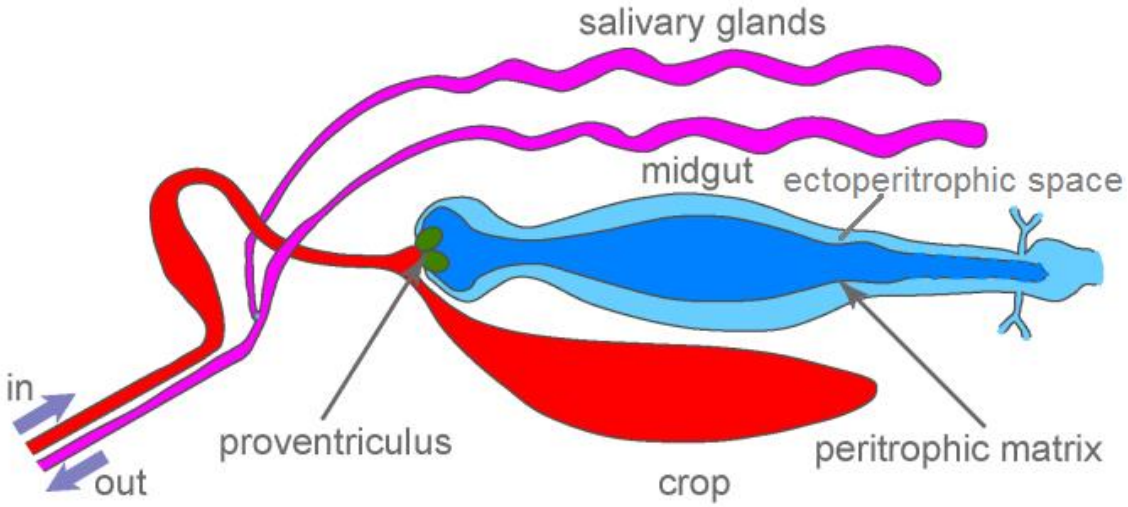


Figure 2.6: **The visceral organs of the tsetse fly** (Sharma *et al.*, 2009)**.**

2.5 *Trypanosoma brucei* survival challenge in the tsetse fly

Trypanosoma brucei encounters changing micro-environments during its passage in the insect vector. One of the first changes it encounters in the tsetse fly's midgut is a change in the source of energy. Proline is the main source of energy for tsetse during flight (Bursell, 1966) and is therefore the available source of nutrients for the procyclic forms of the parasite in the midgut. Therefore, the parasite undergoes metabolic reprogramming; it switches from glucose to proline as the primary source of energy (Mantilla *et al.*, 2017). Reprogramming of the parasite metabolism means that the fly and the parasite may compete for proline as an energy source. However, starving the fly to limit the available proline to both the fly and the parasite does not change the parasite

number, leading to the proposition that the parasite uses programmed cell death to avoid compromising tsetse fitness (Welburn & Maudlin, 1997). In addition to nutritional change, the parasite also encounters proteases, lectins, and changing pH in the gut of tsetse fly (Dyer *et al.*, 2013).

The peritrophic matrix (PM) of insects line their midgut and acts as a physical and biochemical barrier that protects the midgut epithelium from damage by abrasive food particles, ingested toxins, pathogens, and digestive enzymes (Hegedus *et al.*, 2009). In the midgut, the parasite encounters tsetse's digestive enzymes, immune peptides, reactive oxygen species, and serum complement (Bullard *et al.*, 2012; Hu & Aksoy, 2005; MacLeod *et al.*, 2007) and avoids their attack by crossing from the PM into the ectoperitrophic space (ES). However, crossing the PM acts as a hurdle for the parasite because procyclic trypanosomes are larger (several microns long) as compared to the smaller pore size of tsetse PM at ~ 9nm (Miller & Lehane, 1990). The mechanism by which the parasite crosses the PM is yet to be determined. The challenges encountered by *T. brucei* during its development in the tsetse fly (Figure 2.3) may provide novel control strategies upon the elucidation of the molecular mechanism under which the parasite overcomes them.

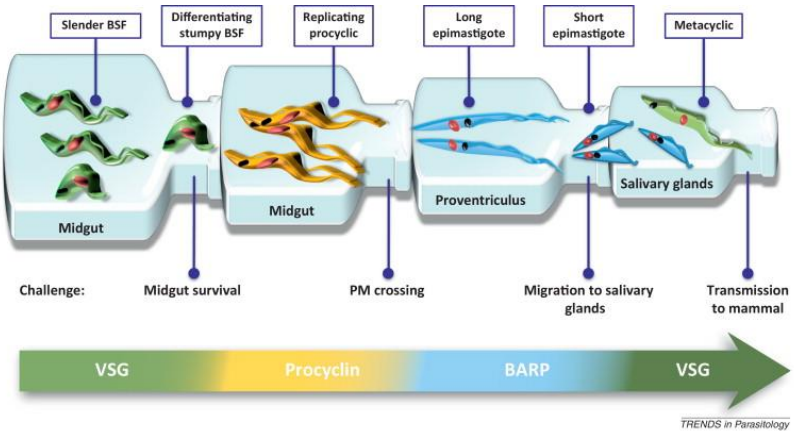


Figure 2.7: **Bottlenecks encountered by *T. brucei* during its developmental stages in the tsetse fly. BSF: Blood stream form; VSG: Variant surface glycoproteins; BARP: *brucei* alanine-rich protein** (Dyer *et al.*, 2013).

2.6 Gene expression during the development of *T. brucei*

The arrangement of genes on the genomes of trypanosomatids is sequential (polycistronic arrangement) as observed by Berriman *et al.* (2005), El-Sayed *et al.* (2005), and Ivens *et al.* (2005). The polycistronic arrangement of these genes means that all genes that belong to the same polycistronic unit are transcribed at the same level even if they do not encode related proteins. Consequently, gene regulation occurs post-transcriptionally (Clayton, 2002, 2019). Following transcription, the polycistronic mRNAs are processed into monocistronic mRNAs through the processes of *trans*-splicing and polyadenylation (Clayton & Shapira, 2007). Besides gene regulation during *trans*-splicing and polyadenylation, regulation also occurs during RNA export, through regulating RNA stability, during the process of protein translation, and after protein synthesis through regulation of protein stability (Clayton, 2002). The stability of mRNA and its degradation are major transcription regulation mechanisms and are effected through the 3' untranslated regions (3′ UTR) sequence motifs on the mRNA (Zubiaga *et al.*, 1995). These sequences contribute to regulation by being recognized and bound by mRNA degradation components or proteins that affect the stability of mRNA.

2.6.1 RNA-binding proteins

RNA-binding proteins (RBPs) are proteins that bind to the 3' UTR of mRNAs by recognizing the mRNA's structural motifs. These structural motifs, also called RNA regulatory elements (RREs), govern the fate of functionally related mRNA in trypanosomatids where gene regulation is coordinated at the post-transcription level (Haile & Papadopoulou, 2007; Ouellette & Papadopoulou, 2009). RNA-binding proteins possess functional domains such as Zinc Finger domains, RNA Recognition Motif (RRM), and Alba and Pumillo domains which interact with mRNA's structural motif (Clayton & Shapira, 2007). The RBPs are crucial regulators of gene expression during the development of *T. brucei* in the insect vector as it faces changing developmental and environmental conditions.

2.6.1.2 Role of RBPs in trypanosome development in the tsetse fly

RNA-binding proteins play different roles in the development of *T. brucei* in its life cycle. These regulatory proteins have been implicated in several functions such as cell differentiation and mRNA degradation. Pumilio/fem-3 mRNA binding factor 9 (PUF9), one of the pumilio-domain protein, was found to play a role in mRNA degradation through binding and stabilizing a small number of mRNA that increases in the late G1 phase of the cell cycle (Archer et al., 2009). PUF9 was also found to target the 'UUGUACC' motif that was over-represented on its target's 3' UTR region. Another RBP, ZFP1, is required for the differentiation of bloodstream-form to procyclic form (Hendriks et al., 2001; Hendriks & Matthews, 2005). ZFP1 is a zinc finger domain protein with the CCCH zinc finger domain. ZC3H20, also a zinc finger RBP protein, was shown to be enriched in *T. brucei* procyclic forms where they stabilized developmentally regulated transcripts (Ling et al., 2011). The function of the majority of RBPs is, however, yet to be known.

Trypanosoma brucei RBP6 is an RNA recognition motif (RRM) whose overexpression *in vitro* in the procyclic form leads to progression through developmental stages; epimastigotes formed and were followed by metacyclic trypomastigotes – the infective form that expresses VSG (Kolev *et al.,* 2012). This was a remarkable breakthrough that yielded an *in vitro* system for studying developmental stages in the insect vector. However, many important questions on trypanosome development in the tsetse remain, starting with how the molecular mechanism of RBP6 in non-experimental conditions is mediated (Ooi & Bastin, 2013). Additionally, RBP10 is another RRM protein that was found to be up-regulated in the bloodstream-form trypanosomes (Wurst *et al.*, 2012). Its expression in procyclic forms leads to their development into bloodstream-form (Mugo & Clayton, 2017). These examples highlight the crucial role that RBPs play in post-transcription gene regulation during the development of *T. brucei* in the tsetse vector, and the need to decipher their role in trypanosome development. RNA-binding proteins

may provide novel strategies to control parasite transmission through the insect vector by interfering with its modulation of gene expression which occurs post-transcriptionally.

2.7 Controlling the parasite in the insect vector

T. brucei encounters important bottlenecks during its developmental progress in the tsetse fly vector. In the tsetse's midgut, the parasite encounters anti-parasitic immune responses which it must evade (Aksoy *et al.*, 2003). The insect's immune response is an important component of the vector competence, that is, its genetic ability to transmit pathogens (Weiss & Aksoy, 2011). Besides the immune response, the parasite also encounters the gut microbiota. This microbiota may influence vector competence by directly interacting with the parasite and therefore provides novel avenues to interfere with the transmission of the parasite. Engineering tsetse refractoriness is one of the active research areas geared towards eliminating the parasite from tsetse fly populations (Kariithi *et al.*, 2018).

2.7.1 Paratransgenesis

Paratransgenesis entails isolation and genetic modification of symbiotic bacteria from a host, followed by their re-introduction to express anti-pathogenic products (Coutinho-Abreu *et al.*, 2010). In the tsetse fly, *Sodalis glossinidius* is one of the endosymbiotic bacteria that reside in the midgut, salivary glands, milk gland, muscle, and fat body tissues (Cheng & Aksoy, 1999). Other major symbionts found in tsetse fly are *Wigglesworthia glossinidia* and *Wolbachia spp*. These endosymbionts can be used to reduce vector competence through paratransgenesis. Towards this end, *Sodalis glossinidius* has been engineered to express trypanocidal products and maybe a potential *in vivo* drug delivery vehicle especially because it resides in tissues close to trypanosomes in the tsetse fly (De Vooght *et al.*, 2012, 2014; Haines *et al.*, 2003). This innovative approach provides proof of the ability to control the parasite transmission in the vector. However, it is dependent on the knowledge of trypanosome biological

processes and molecules that can be targeted by trypanocidal products. This knowledge may be obtained from constructing biological networks such as gene co-expression networks because genes and their protein products carry out biological processes as functional clusters and interact with each other through complex networks.

2.8 Biological networks

Biological functions arise from molecular interactions between key components of the cell such as proteins, DNA, RNA, and small molecules. These molecular interactions form biological networks (Barabási & Oltvai, 2004). Graphically, networks are represented as nodes (or vertices) and edges which connect the nodes. Nodes represent biological components such as genes or proteins while edges represent interactions between the connecting pair of nodes. Various types of biological networks can be constructed using the available data from transcriptomic, genomic, and proteomic experiments. The biological networks which can be constructed from the data include protein-protein interaction networks involving proteins, metabolic networks composed of metabolites, reactions, and enzymes, transcription regulatory networks made up of gene and transcription factors, signal transduction networks, and functional association networks of genes such as gene interaction networks and gene co-expression networks (Albert, 2005). Some of the biological networks that have been constructed include protein interactions networks for *D. melanogaster* (Guruharsha *et al.*, 2011), *T. brucei* (Gazestani *et al.*, 2016), *C. elegans* (Li *et al.*, 2004), gene interaction networks for *S. cerevisiae* (Costanzo *et al.*, 2010), and gene co-expression network for *Arabidopsis thaliana* (Mao *et al.*, 2009).

2.8.1 Types of biological networks

2.8.1.1 Protein-protein interaction networks

Interaction between proteins in cells, protein-protein interactions (PPIs), plays critical and fundamental roles to their proper functioning. Therefore, the study of PPIs through protein-protein interaction networks (PPINs) is crucial to our understanding of cellular

functions (De Las Rivas & Fontanillo, 2012). Interactions between proteins can be through direct physical contact, regulatory factors such as transcription factors, genetic interactions through genes that encode them, and functional association through their involvement in the same metabolic or signaling pathway.

Protein-protein interaction networks are constructed from data generated through experimental methods such as yeast two-hybrid (Y2H) and affinity purification mass spectrometry (AP-MS) (Koh *et al.*, 2012). PPIs can also be predicted computationally through tools such as Biana Interolog Prediction Server (BIPS) (Garcia-Garcia *et al.*, 2012), Protein-Protein Interaction Prediction Engine (PIPE) (Pitre *et al.*, 2006), and protein interactions by structural matching (PRISM) (Tuncbag *et al.*, 2011). Upon the creation of PPINs, the networks are stored in online databases where the information about interactions between proteins can be assessed. These databases include the biomolecular interaction network database (BIND) (Bader *et al.*, 2003), the database of interacting proteins (DIP) (Xenarios *et al.*, 2002), and IntAct database (Hermjakob *et al.*, 2004).

2.8.1.2 Metabolic networks

Metabolism is a fundamental aspect of life and presents an avenue for advancing our knowledge of living systems (Sévin *et al.*, 2015). One strategy of studying metabolism in organisms is via metabolic networks. In network biology, metabolic networks were among the first to be studied in different organisms (Jeong *et al.*, 2000). Metabolic networks are made up of metabolites and the metabolic reactions in which they participate (Barabási & Oltvai, 2004). These networks can be reconstructed using genomic data or metabolomics data (Cottret & Jourdan, 2010). Using genomic data, functional annotation is performed to identify regions of the genome coding for enzymes, followed by the identification of metabolic reactions encoded by referencing databases such as Braunschweig Enzyme Database (BRENDA) (Schomburg *et al.*, 2004) and KEGG (Kanehisa & Goto, 2000). Alternatively, metabolomics data which consists of quantitative measures of metabolites during chemical reactions can be used

to construct a metabolic network (Cottret & Jourdan, 2010). Metabolic networks are used in biomedical research to identify metabolites involved in diseases such as cancer (Armitage & Barbas, 2014) and infectious diseases involving host and pathogens to identify therapeutic targets (Sévin *et al.*, 2015).

2.8.1.3 Transcription regulatory networks

Transcription regulatory networks (TRNs) represents physical and regulatory interactions between transcription factors and their target genes. Transcription factors are made up of two domains; a DNA-binding domain which binds to the DNA and a transcription regulation domain which binds to other regulatory proteins (Mitchell & Tjian, 1989). Elucidating the role of transcription factors in gene regulation through regulatory networks can provide insights into mechanism of pathogenesis during disease states. Transcription regulatory networks can be deciphered from data generated through experimental techniques such as RNA interference (RNAi) and chromatin immunoprecipitation (ChIP) (Blais & Dynlacht, 2005). Alternatively, TRNs can be constructed through computational modelling using gene expression and ChIP data (He & Tan, 2016). Some of the tools used in the modelling of TRNs include Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Margolin *et al.*, 2006), ChIPXpress (Wu & Ji, 2013), and Trustful Inference of Gene Regulation with Stability Selection (TIGRESS) (Haury *et al.*, 2012).

2.8.1.4 Signal transduction networks

Signal transduction involves the sensing of the environment by an organism through molecular components such as proteins (Gomperts *et al.*, 2015). Understanding the mechanism in which organisms transmit and respond to signals can provide opportunities to study diseases affecting them. For example, signal transduction is thought to play a significant role in tumorigenesis and therefore presents an avenue for developing drugs against cancer (Kolch *et al.*, 2015). In signal transduction, the molecular components are mainly proteins (Gomperts *et al.*, 2015). Hence, PPINs

generated from gene expression data using techniques such as DNA microarrays, Co-Immunoprecipitation, and Y2H are used in the elucidation of signaling pathways (Steffen *et al.*, 2002). Some of the tools used in the construction of signal transduction networks include NetSearch (Steffen *et al.*, 2002), CyTRANSFINDER (Politano *et al.*, 2016), PathFinder (Bebek & Yang, 2007), and BowTieBuilder (Supper *et al.*, 2009).

2.8.1.5 Gene co-expression networks

A gene co-expression network (GCN) is a network where nodes represent genes and edges represent the correlation of gene expression. A GCN identifies genes showing coordinated behaviour of gene expression across a group of samples (van Dam *et al.*, 2018). Co-expression networks are constructed from gene expression data such as microarray or RNA-seq gene expression data and indicate which genes are active simultaneously based on their correlation scores. It has been shown that similar co-expression patterns usually indicate functional linkages between genes (Kharchenko *et al.*, 2005). Co-expression network analysis is done to find novel genes implicated in a given biological process, perform functional gene annotation, identify regulatory genes, and for gene prioritization (van Dam *et al.*, 2018). Generally, a gene co-expression network provides novel insights into the molecular processes of an organism.

2.8.2 The topology of biological networks

Biological insight from networks is obtained through network topology analysis. Network topology refers to the arrangement of nodes and edges within a network (Albert, 2005). Topological properties are used to obtain relevant biological information about nodes which represent biological component and edges that represent their interaction. Node degree is the number of edges connecting to a given node (Albert, 2005). Nodes with a high degree of connections and a high clustering coefficient become hubs and are likely to be associated with essential genes in the network (Albert, 2005; Carlson *et al.*, 2006). Based on the node degree of biological networks, most of the networks are considered to be scale-free networks (Albert, 2005; Barabási & Oltvai,

2004). In scale-free networks, most nodes are connected to a few neighbours while a small number of nodes, the hubs, are connected to a high number of neighbours. Biological networks also form modules (Hartwell *et al.*, 1999). Modules are formed by highly connected genes in the network which form sub-networks that correspond to specific biological function. Modules in a biological network act as starting points for further study and thus reduce global network complexity. One way in which modules can be studied is through enrichment analysis. Enrichment analysis allows for associating a module with biological functions through the 'guilt-by-association' (GBA) rule which states that interacting genes are more likely to share a function (Wolfe *et al.*, 2005).

2.8.3 Construction and analysis of gene co-expression networks

Construction of gene co-expression network begins with defining the relationship between each pair of genes. Correlation measures such as Pearson correlations have been used to describe the similarity of gene expression between gene pairs (D'haeseleer, 2005; Steuer *et al.*, 2002). The co-expression network is represented as a gene-gene similarity matrix created from the correlation scores. From the matrix, a weighted or un-weighted and signed or unsigned co-expression network can be constructed where nodes represent genes and edges indicate the presence and strength of a co-expression relationship (van Dam *et al.*, 2018). Absolute correlation values are used in unsigned networks thus negatively correlated genes also indicate co-expression. A signed network scales correlation scores between 0 and 1 such that values less than 0.5 indicate a negative correlation while values greater than 0.5 indicate a positive correlation. In a weighted network, the edges indicate the strength of co-regulation through continuous weights between 0 and 1 while in un-weighted networks, the connection between gene pairs is either 0 indicating no connection, or 1 indicating the presence of a connection (van Dam *et al.*, 2018).

After the construction of a gene co-expression network, network analysis is performed to identify modules. Genes with similar expression patterns in multiple samples are

grouped through clustering techniques such as k-means clustering, self-organizing map (SOM), or hierarchical clustering (D'haeseleer, 2005). A widely used gene co-expression network analysis pipeline is the Weighted Gene Co-expression Network Analysis (WGCNA), which uses hierarchical clustering (Zhang & Horvath, 2005). In WGCNA, module hub genes are identified as genes with high module membership and are representative of the module's overall function (Horvath, 2011). Enrichment analysis for the genes in a module is used to associate modules with functions through guilt-by-association (GBA).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Data acquisition and quality assessment

RNA-Seq datasets of *Glossina morsitans morsitans* (tsetse fly) trypanosome-infected midgut, proventriculus, and salivary glands tissues were obtained from European Nucleotide Archive (ENA) (Leinonen *et al.*, 2011) under accession numbers SRP002243 and SRR965341. The dataset consisted of 18 samples; seven (7) midgut, four (4) proventriculus, and seven (7) salivary glands (Savage *et al.*, 2016; Telleria *et al.*, 2014). The quality of the data was assessed using FastQC version 0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Before reads mapping, *T. brucei* genome and *G. morsitans* scaffolds genome were obtained from TriTrypDB (Release 43) (Aslett *et al.*, 2010) and VectorBase (Lawson *et al.*, 2007), respectively, and concatenated to create a chimeric genome. The RNA-Seq reads were mapped to the chimeric genome of *T. brucei* and *G. morsitans* using HISAT2 version 2.1.0 (Kim *et al.*, 2015) to remove ambiguously mapped reads. Duplication rates were computed after read mapping using the MarkDuplicates tool from Picard toolkit version 2.20.3 (http://broadinstitute.github.io/picard/) to mark duplicate reads. Further, dupRadar Bioconductor R package version 1.18.0 was used to assess the RNA-Seq data for the presence of PCR duplicates (Sayols *et al.*, 2016). Samples that had PCR duplicates (reads made from the same original cDNA fragment during PCR) were excluded from downstream analysis.

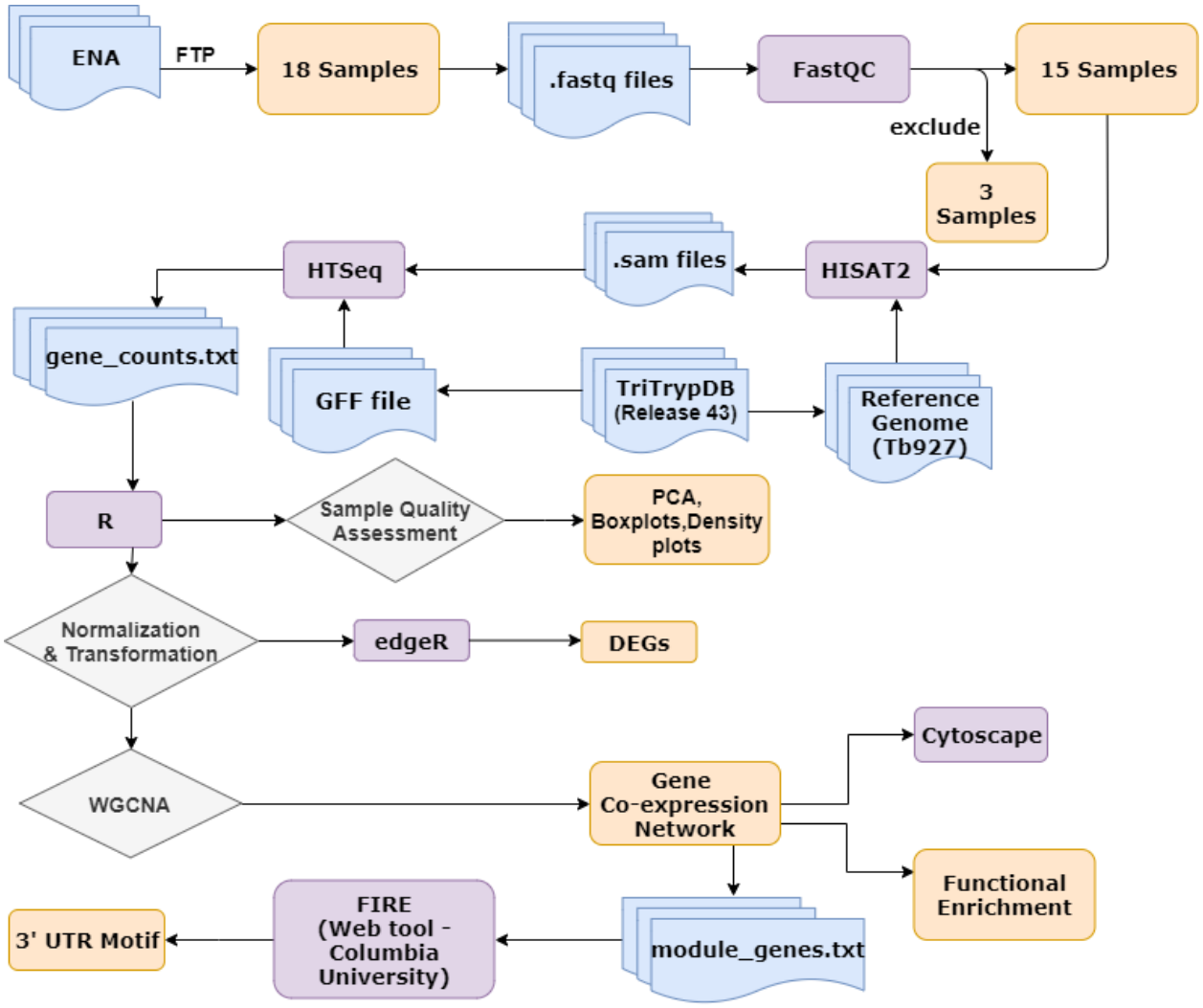


Figure 3.1: **A schematic showing workflow of the methodology**.

3.2 Read quantification

The reads that mapped to the *T. brucei* genome were counted using HTSeq version 0.11.2 (Anders *et al.*, 2015) using the annotation file of *T. brucei* downloaded from TriTrypDB (Release 43). Non-protein coding genes (ncRNA, snRNA, snoRNA, pseudogenic transcripts, rRNA, and tRNA) were excluded from the read counts as this study focused on protein-coding genes and their functional analysis.

3.3 Sample quality assessment and data normalization

Genes with low expression levels were removed from the read counts data using *filterByExpr* function from R package edgeR version 3.8 (Robinson *et al.*, 2010). Sample quality was assessed using Pearson correlation heatmaps, Principal Component Analysis (PCA), and box plots in R version 3.6.0 (R Core Team, 2020). Trimmed mean of M-values (TMM) was used as a normalization method using *calcNormFactors* function in edgeR (Robinson *et al.*, 2010). The normalized read counts were then converted to counts per million and $\log_2$ transformed for downstream analysis. Batch effects were adjusted for using the ComBat method from sva R package version 3.32.1 (Leek *et al.*, 2012).

3.4 Construction of gene co-expression network

The weighted gene co-expression network was constructed using the WGCNA R package version 1.66 (Langfelder & Horvath, 2008). First, soft-thresholding power, β, was determined using the *pickSoftThreshold* function from the WGCNA package. This was followed by the construction of a weighted adjacency matrix using the *adjacency* function, after which the matrix was computed into Topological Overlap Matrix (TOM) using the *TOMsimilarity* function (Zhang & Horvath, 2005). The TOM measure between pairs of genes was used as input for average linkage hierarchical clustering by first creating a dissimilarity matrix (dissTOM = 1 - TOM) and then using the *flashClust* function to create the gene tree dendrogram. The Dynamic Tree Cut algorithm was used to identify modules using the gene tree dendrogram as input for *cutreeDynamicTree* function from dynamicTreeCut R package version 1.63-1 (Langfelder *et al.*, 2008). The *chooseTopHubInEachModule* function from the WGCNA package was used to identify the hub genes.

3.5 Network functional enrichment analysis and visualization

The goseq R package version 1.36.0 (Young *et al.*, 2010) was used to test for enrichment of gene ontology (GO) (Ashburner *et al.*, 2000) and Kyoto Encyclopedia of Gene and

Genomes (KEGG) (Kanehisa & Goto, 2000) annotations for each of the identified modules. The GO and KEGG annotations were obtained from TriTrypDB. The generated lists of GO terms for the modules were summarized using REVIGO (http://revigo.irb.hr/) (Supek *et al.*, 2011). Cytoscape version 3.7.1 (Shannon *et al.*, 2003) was used to visualize the network using the *exportNetworkToCytoscape* function from the WGCNA package.

3.6 Prediction of 3' UTR motifs

All the genes in the gene co-expression network and their corresponding cluster/module index were used to generate an expression file that was used as input for the tool FIRE, version 1.1a (Elemento *et al.*, 2007). This expression file was submitted online to FIRE (https://tavazoielab.c2b2.columbia.edu/FIRE/). Default parameters were applied in prediction of 3' UTR motifs.

The analysis code used in data pre-processing, network construction, and functional analysis is archived at https://github.com/wanjauk/tbrucei_gcn.

CHAPTER FOUR

RESULTS

4.1 Data quality assessment and pre-processing

Eighteen (18) samples were analyzed in this study. Quality assessment of the data revealed that three (3) samples had poor quality due to PCR duplicates and were excluded from the analysis (Table 4.1). The remaining 15 samples generated a total of 7,390 genes after removing non-protein-coding and lowly-expressed genes from the read counts.

No.	Library Name	Run	Sample Name	Tissue	Source Study
1	mg1	SRR039378	MG1	MG	Savage et al. (2016)
2	mg1	SRR039381	MG1	MG	Savage et al. (2016)
3	mg1	SRR039453	MG1	MG	Savage et al. (2016)
4	MG2_SL	SRR039454	MG2	MG	Savage et al. (2016)
5	MG2_SL	SRR039455	MG2	MG	Savage et al. (2016)
6	PV2_SL	SRR039456	PV2	PV	Savage et al. (2016)
7	PV2_SL	SRR039457	PV2	PV	Savage et al. (2016)
8	SA2_SL	SRR039939	SA2	SG	Savage et al. (2016)
9	SA2_SL	SRR039940	SA2	SG	Savage et al. (2016)
10	mg1	SRR039948	MG1	MG	Savage et al. (2016)
11	MG2_SL	SRR039949	MG2	MG	Savage et al. (2016)
12	PV2_SL	SRR039950	PV2	PV	Savage et al. (2016)
13	SA2_SL	SRR039952	SA2	SG	Savage et al. (2016)
14	PV2_SL	SRR042429	PV2	PV	Savage et al. (2016)
15	SA1	SRR965341	SA1	SG	Telleria et al. (2014)
16	SA1	SRR039950	SA1	SG	Savage et al. (2016)
17	SA1	SRR039937	SA1	SG	Savage et al. (2016)
18	SA1	SRR039938	SA1	SG	Savage et al. (2016)

Table 4.1: **Metadata for samples downloaded for analysis in this study.**

Samples highlighted in red were excluded from analysis because of a failed quality assessment. MG: Midgut, PV: Proventriculus, SG: Salivary glands.

4.2 Sample quality assessment

The relationship between the samples and the reproducibility of biological replicates was determined using Principal Component Analysis (PCA) and Pearson correlation heatmap analysis before (Appendix 1) and after (Figure 4.1**,** 4.2) adjusting for batch effects which could have resulted from biological replicates. The PCA and Pearson correlation heatmap plots showed that the samples grouped based on the developmental stages of *T. brucei* in the insect vector rather than their biological replicates (Figure 4.1**,** 4.2).

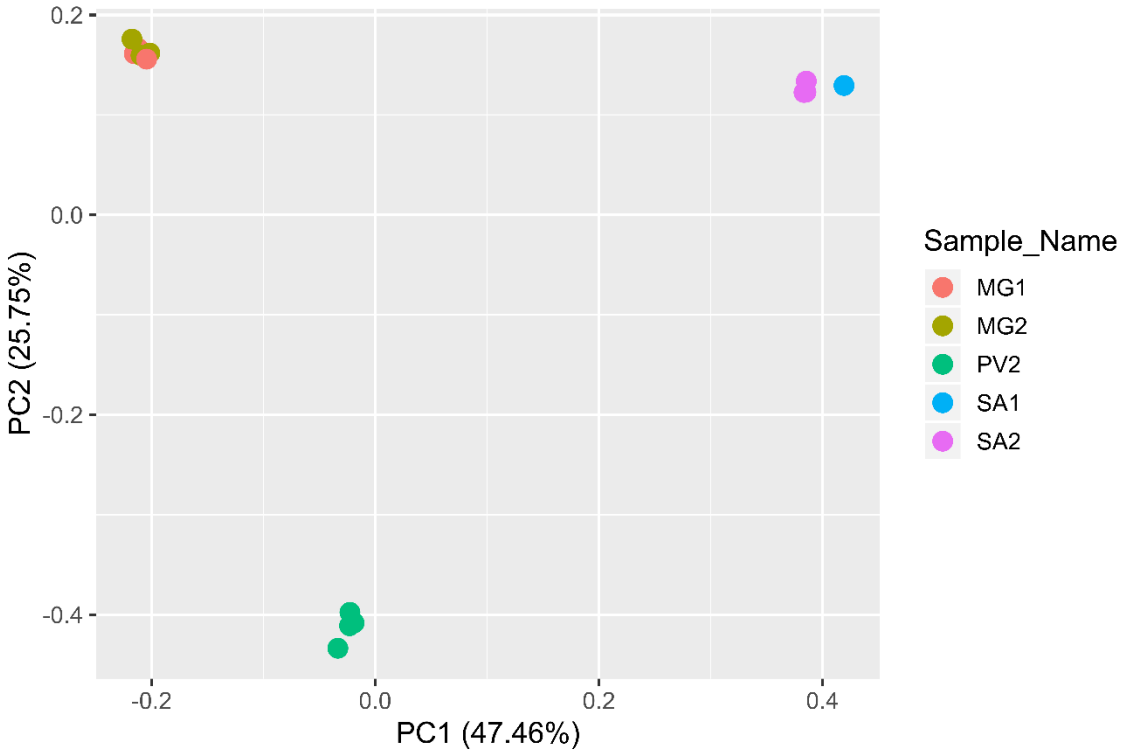


Figure 4.1: **Principal component analysis (PCA) plot. Each point in the PCA plot represents a sample and point color indicates a batch that consists of the biological replicates. MG1 and MG2 are midgut samples, PV2 are proventriculus samples, and SA1 and SA2 are salivary gland samples.**

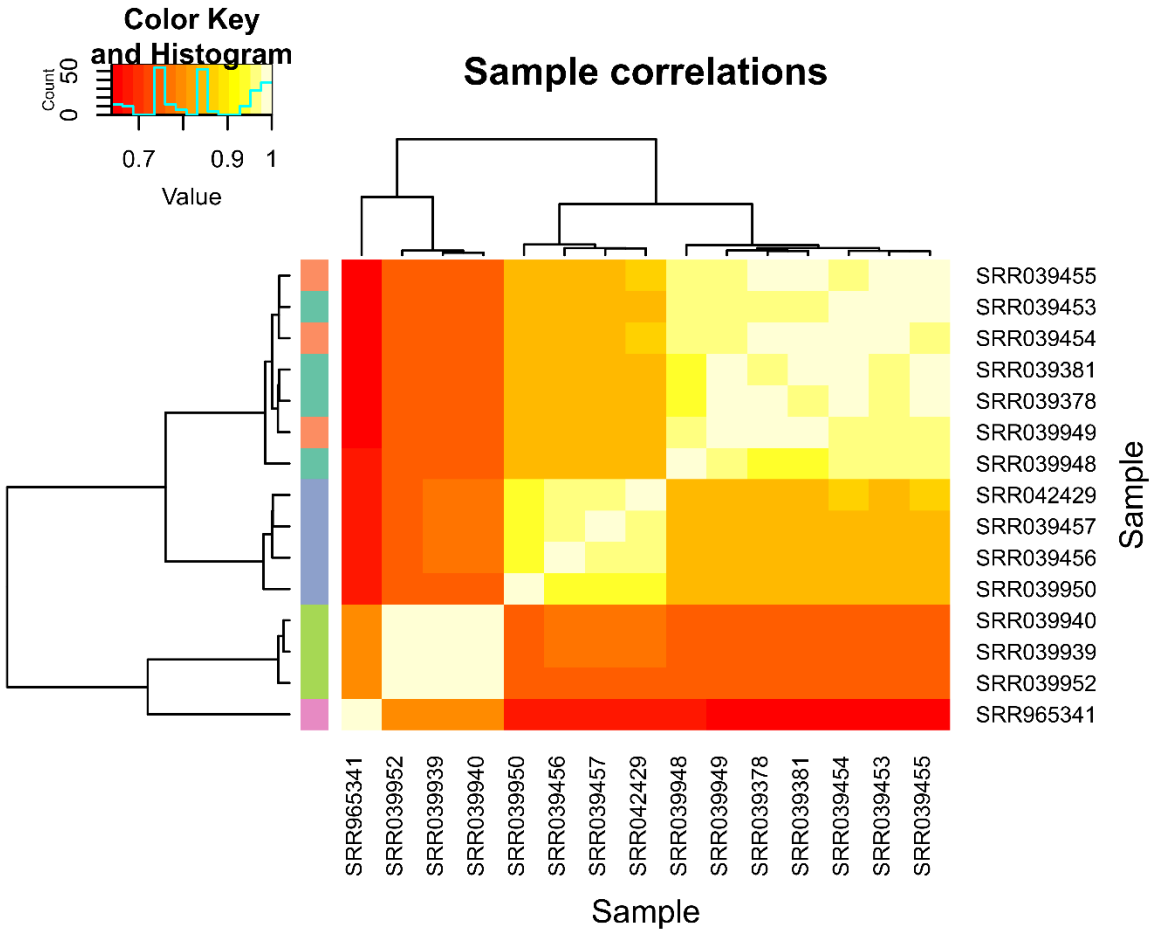


Figure 4.2: Sample correlation heatmap using hierarchical clustering. Color codes along the left side of the sample correlation heatmap indicate samples based on the batch they belong to.

An assessment of the distribution of per-gene read counts per sample showed a median steady-state expression level of ~ 6.5 $\log_2$ counts per million in all the 15 samples before (Appendix 2) and after (Figure 4.3) normalization and batch effect adjustment.

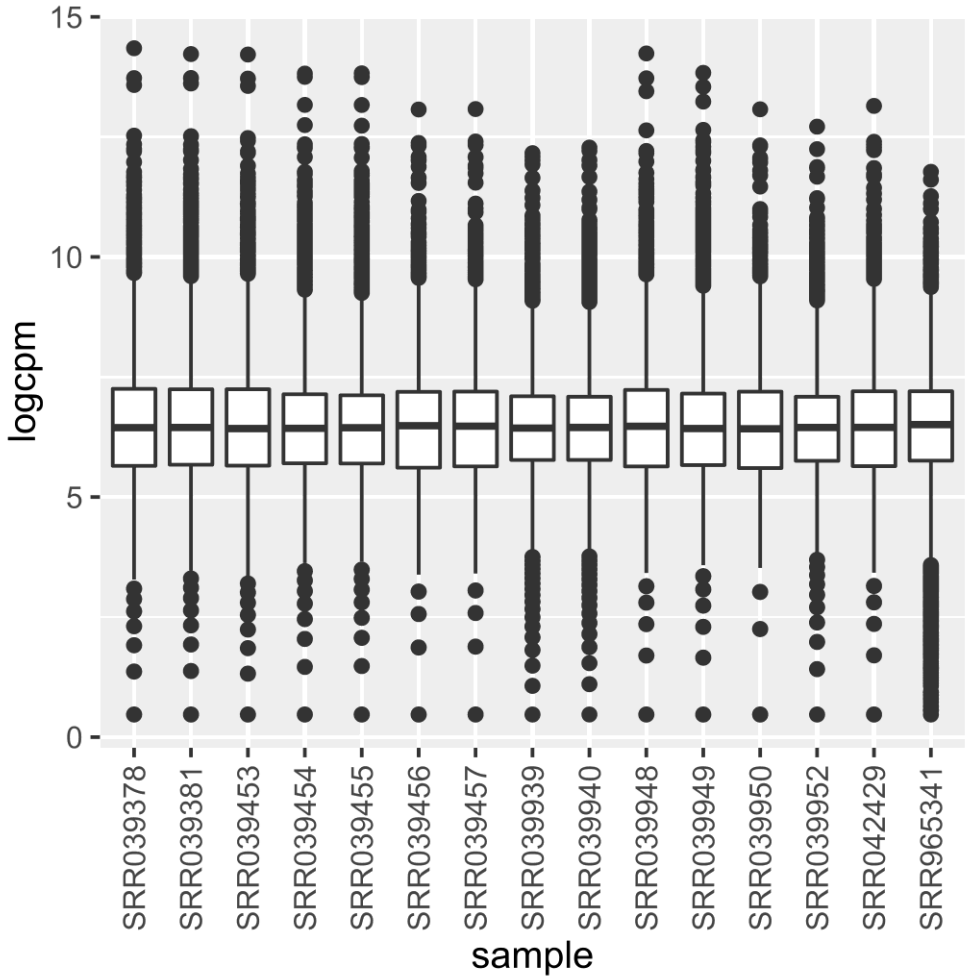


Figure 4.3: **Assessment of per-gene read counts distribution per sample using boxplot after normalization and batch effect adjustment. All the samples have similar median expression level values after normalization.**

4.3 Gene co-expression network

Soft-thresholding power is the power to which co-expression similarity (correlation) is raised to reduce the noise of the correlations in the adjacency matrix. Soft-thresholding power 14, the power for which the scale-free topology fitting index (R^2) was ≥ 0.8, was chosen after analysis of thresholding powers from 1 to 20 (Figure 4.4).

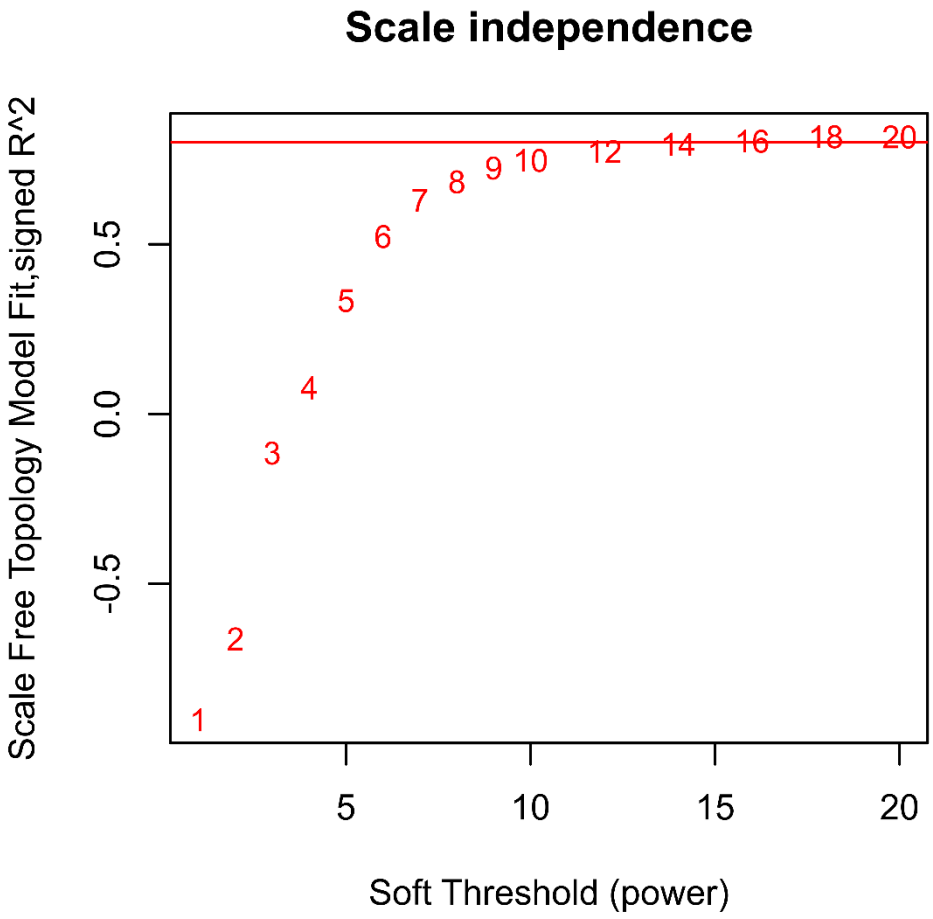


Figure 4.4: **Scale-free topology plot for selecting the power β for the signed correlation network. Scale-free topology index (*y*-axis) as a function of powers, β, 1 to 20 (*x*-axis).**

A total of 28 distinct modules were generated for 7,390 protein-coding genes from the hierarchical clustering tree (dendrogram) using the dynamic tree cut algorithm (Figure 4.5). The resulting network of the 28 modules visualized using Cytoscape is shown in Figure 4.6.

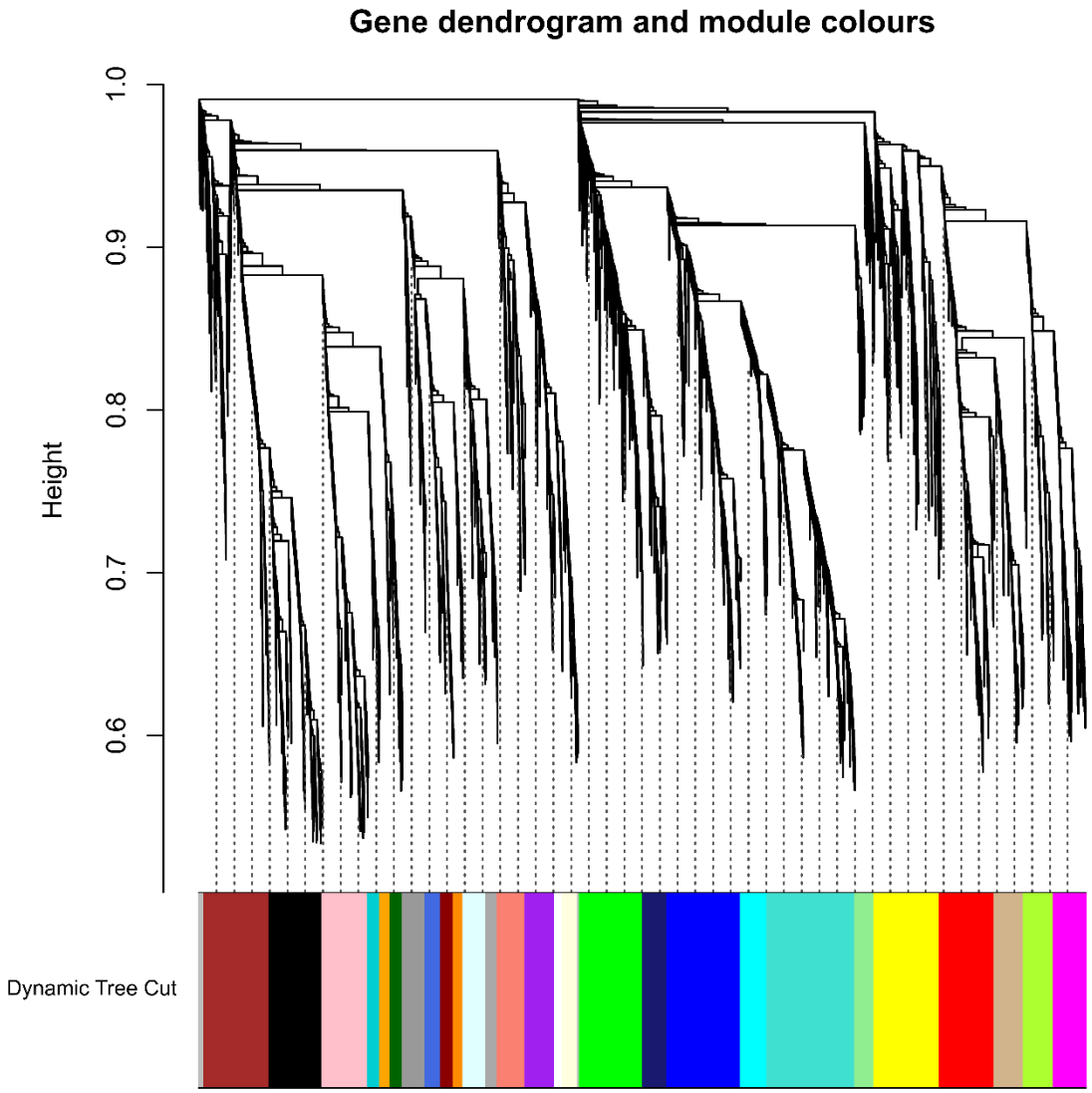


Figure 4.5: **Hierarchical cluster dendrogram. The *x*-axis represents the co-expression distance of the genes, while the *y*-axis represents the genes. A dynamic tree-cutting algorithm identified the modules by splitting the tree at significant branching points. Modules are represented by different colors as shown by the dendrogram.**

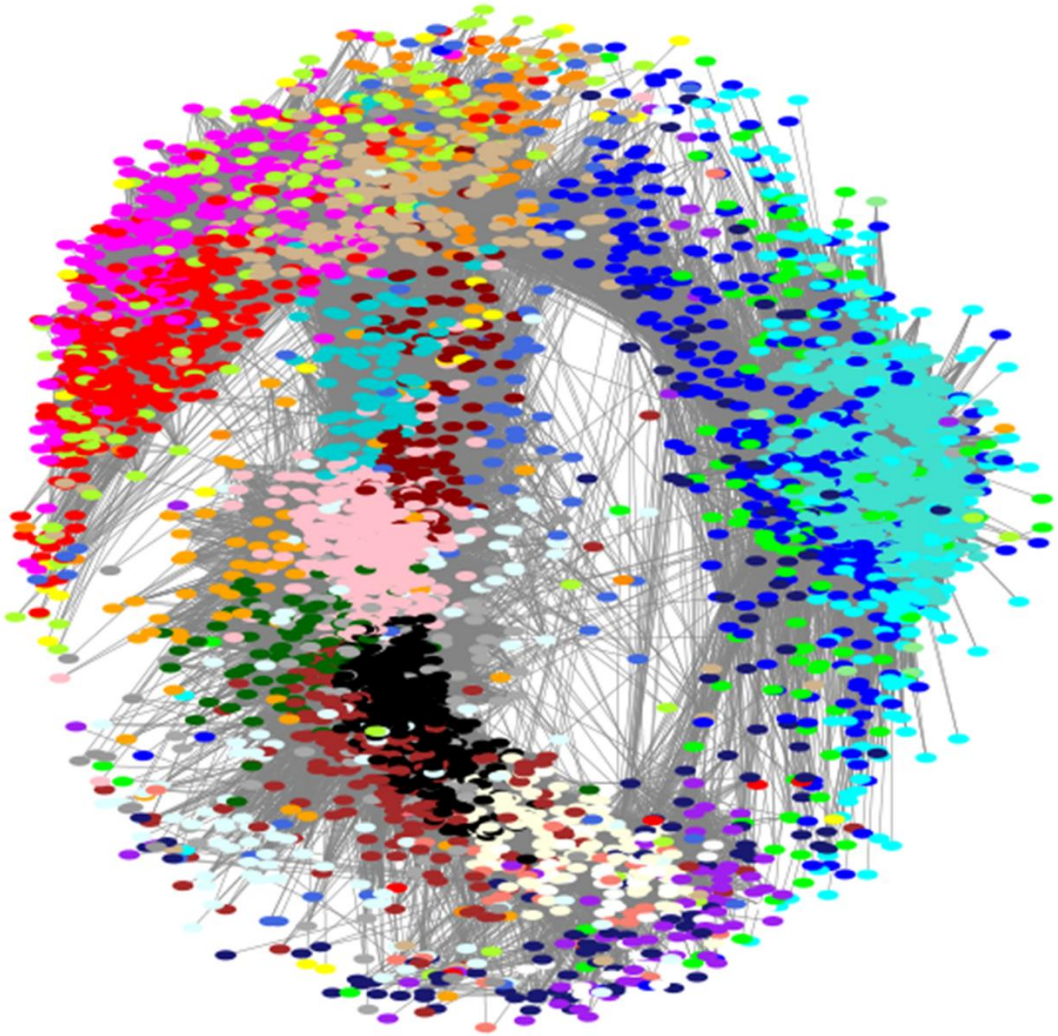


Figure 4.6: **Co-expression network resulting from weighted gene co-expression network analysis (WGCNA) based on topological overlap measure (TOM) greater than 0.3 for visualization. Each point (or node) on the network represents a gene and points of the same color form a gene module. Lines (edges) on the network connecting the nodes represents a relationship between the genes.**

The grey module, which contained 59 genes that could not be assigned to any module, was excluded from the analysis (Figure 4.7). Thus, a total of 27 modules were used in subsequent analysis. The module with the least genes was the white module with 61 genes while the turquoise module had the largest number of genes with 732 genes

(Figure 4.7). The full list of genes with the modules they were assigned can be accessed here:
https://github.com/wanjauk/tbrucei_gcn/blob/master/results/tables/coexpression_network_genes.txt

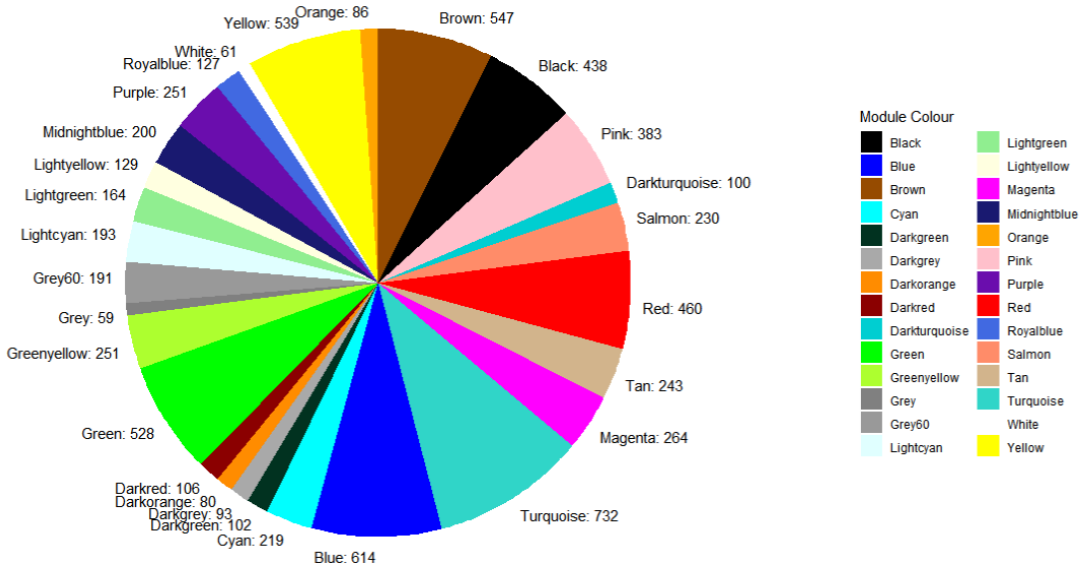


Figure 4.7: **Number of genes identified in each module. A total of 28 modules are shown including the Grey module which was excluded from the analysis.**

4.3.1 Modules functional and pathway enrichment

Fourteen (14) out of 27 modules were found to be enriched for GO terms; 12 were over-represented and two (2) (blue and green modules) were under-represented for GO terms.

Cell cycle-related GO terms that were over-represented included spindle pole in the tan module (Figure 4.8) and DNA replication, DNA metabolic process, and chromosome organization in the red module (Figure 4.9).

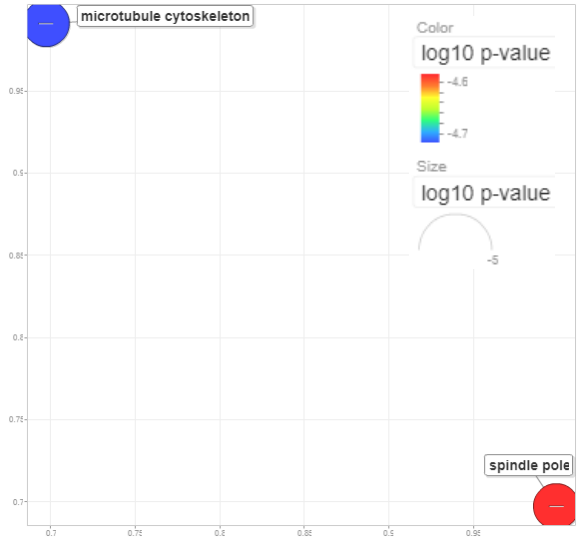


Figure 4.8: **Tan module over-represented Cellular Component GO terms. There were no biological process and molecular functions GO terms that were enriched.**

a **b**

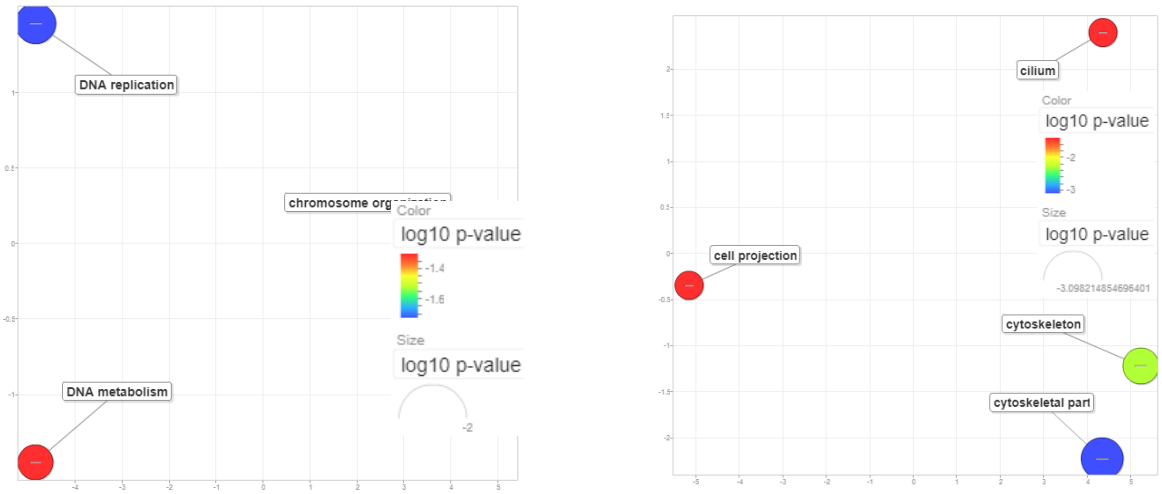


Figure 4.9: **Red module over-represented GO terms. a) Biological Process GO terms b) Cellular Component GO terms. There were no molecular functions GO terms that were enriched.**

The tan and red modules also had GO terms related to microtubule cytoskeleton possibly associated with segregation of organelles during the cell cycle. Other modules with the cytoskeleton as an over-represented GO term included magenta (Figure 4.10) and greenyellow modules (Figure 4.11).

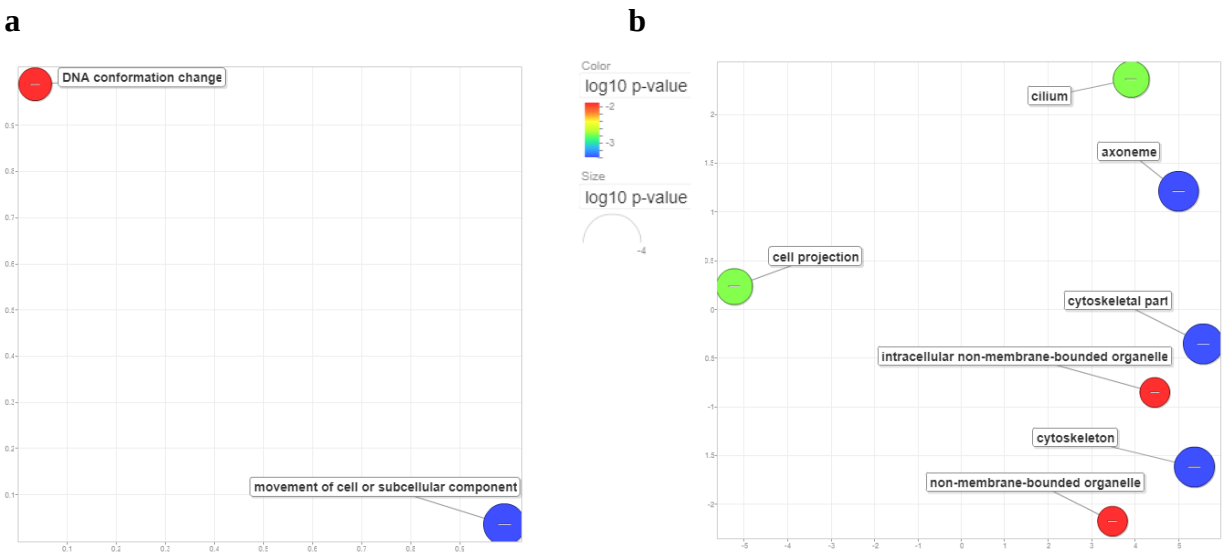
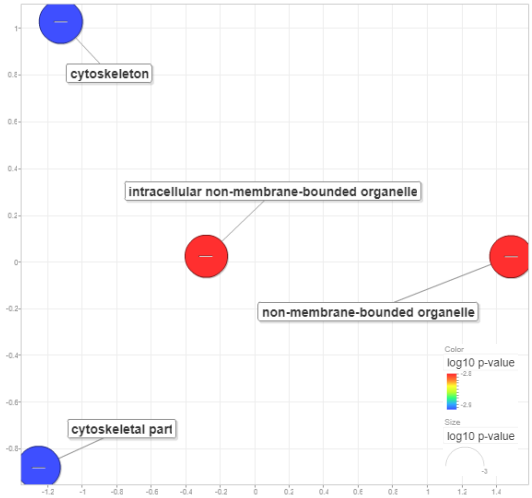


Figure 4.10: **Magenta module over-represented GO terms. a) Cellular Component GO terms b) Molecular Function GO terms. There were no biological process GO terms that were enriched.**

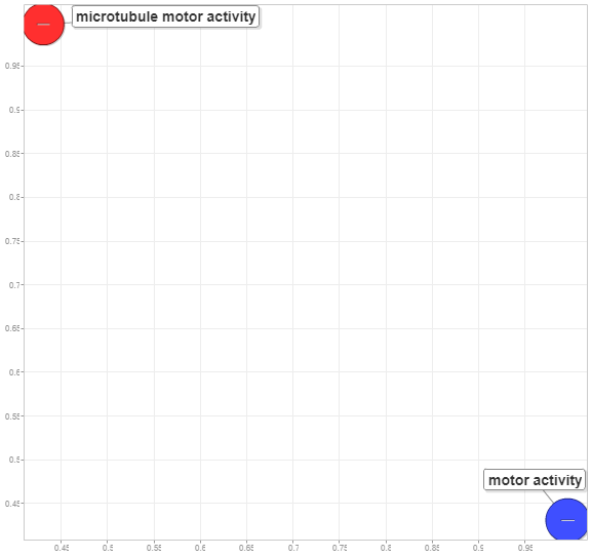
a


arginine kinase activity
phosphotransferase activity, nitrogenous group as acceptor
Color
log10 p-value
Size
log10 p-value

b


cytoskeleton
intracellular non-membrane-bounded organelle
non-membrane-bounded organelle
cytoskeletal part
Color
log10 p-value
Size
log10 p-value

c

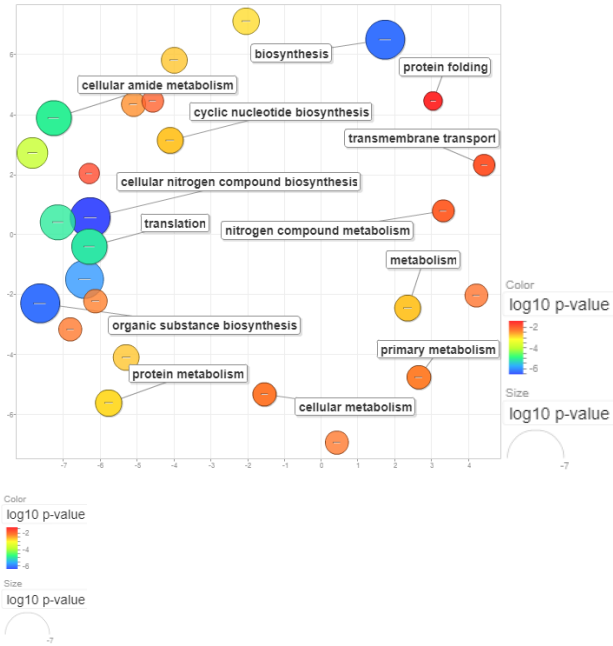

microtubule motor activity
motor activity


Color
log10 p-value
Size
log10 p-value

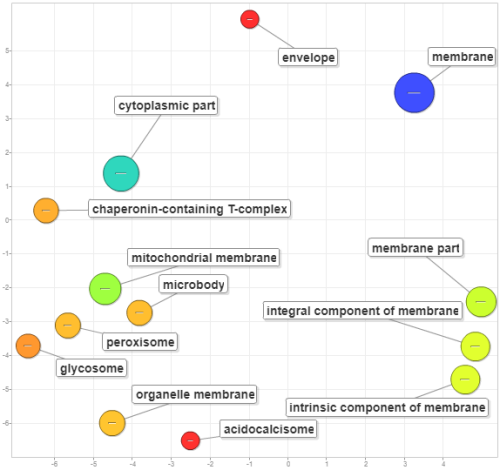
Figure 4.11: Greenyellow module over-represented GO terms. a) Biological Process GO terms b) Cellular Component GO terms c) Molecular Function GO terms.

Cell signaling related GO terms including adenylate cyclase activity, cyclase activity, cyclic nucleotide biosynthetic process, and cyclic nucleotide metabolic process were over-represented in the black (Figure 4.12) and brown (Figure 4.13) modules.

a



b



c

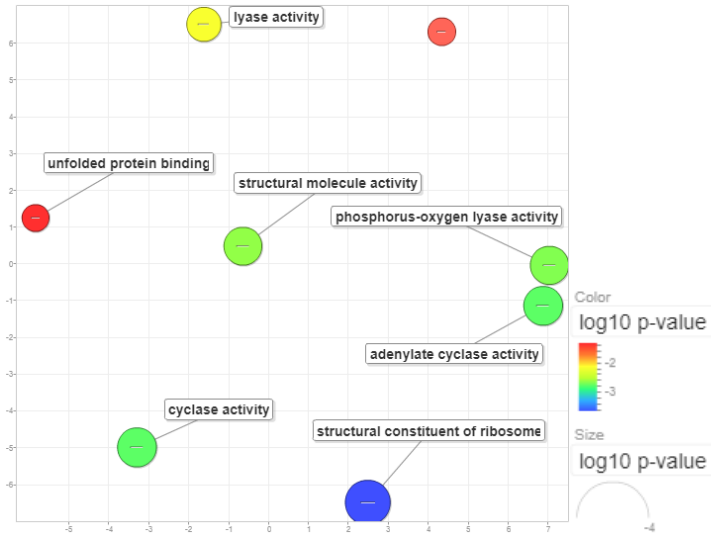


Figure 4.12: **Black module over-represented GO terms. a) Biological Process GO terms b) Cellular Component GO terms c) Molecular Function GO terms.**

a **b**

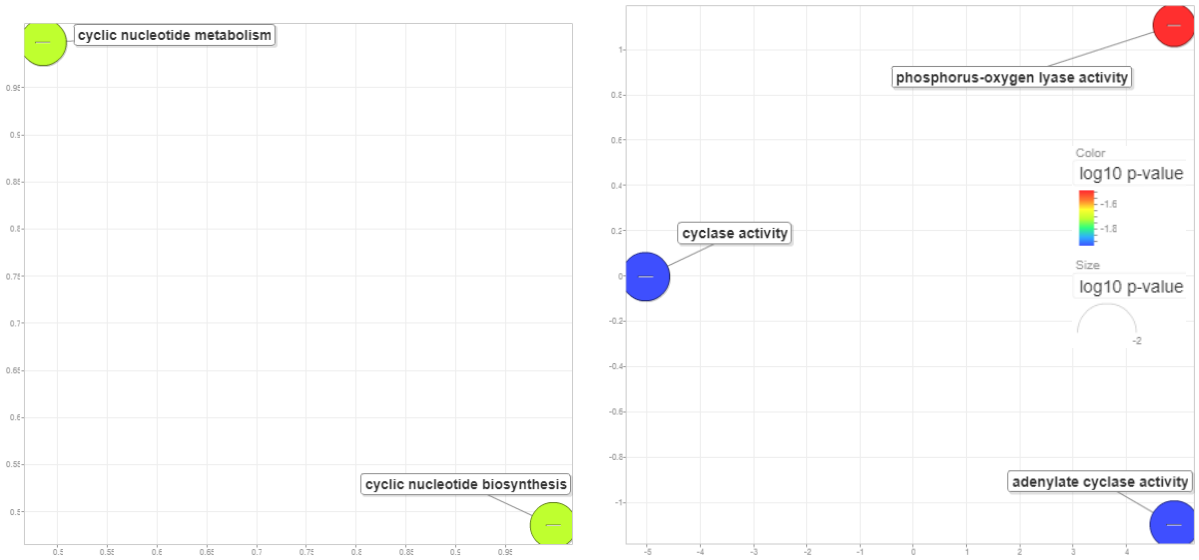


Figure 4.13: **Brown module over-represented GO terms. a) Biological Process GO terms b) Molecular Function GO terms. There were no cellular components GO terms that were enriched.**

Cell membrane-associated GO terms were over-represented in the turquoise module. These GO terms included membrane, membrane part, integral component of membrane, membrane raft, and cell surface (Figure 4.14). Despite having the highest number of genes (732) among the modules, no GO terms associated with other functions were enriched.

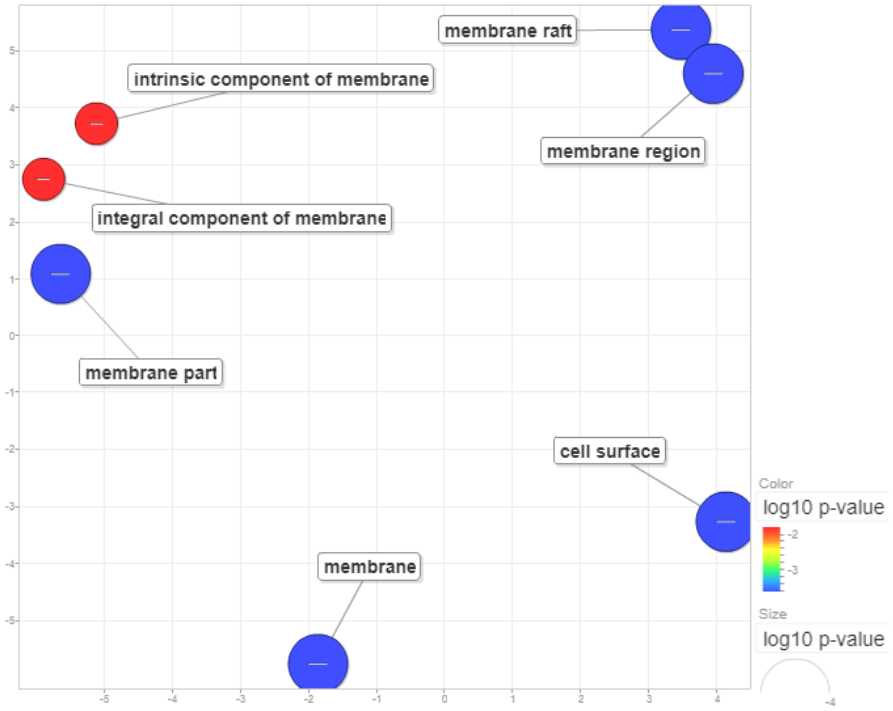


Figure 4.14: **Turquoise module over-represented Cellular Component GO terms. There were no biological and molecular process GO terms that were enriched.**

Protein biosynthesis was associated with over-represented terms in the black (Figure 4.12) and lightyellow module (Figure 4.15). The GO terms in the black module related to protein biosynthesis included translation, peptide biosynthetic process, protein folding, and unfolded protein binding (Figure 4.12). In the lightyellow module, the GO terms included ribosome, translation, peptide biosynthetic process, protein metabolic process, rRNA binding, and structural constituent of ribosome (Figure 4.15).

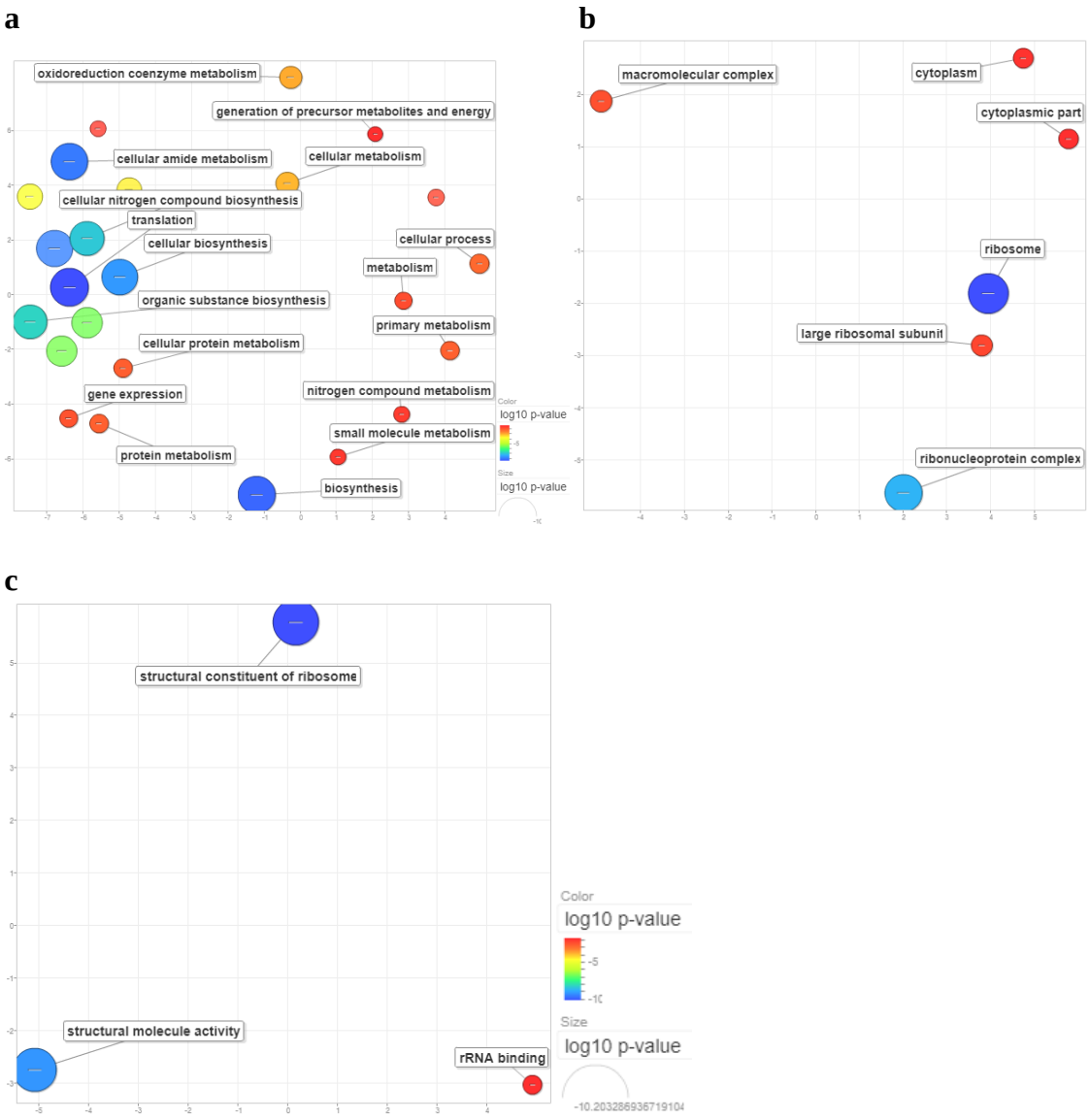


Figure 4.15: Lightyellow module over-represented GO terms. a) Biological Process GO terms b) Cellular Component GO terms c) Molecular Function GO terms.

The lightyellow module was over-represented for one (1) GO terms associated with gene expression (GO:0010467: gene expression) (Figure 4.15). Gene regulation is related to gene expression and may occur post-transcriptionally. GO terms that may be related to gene regulation were over-represented in the salmon module and included RNA binding, nucleic acid binding, and RNA metabolic process (Figure 4.16).

a **b**

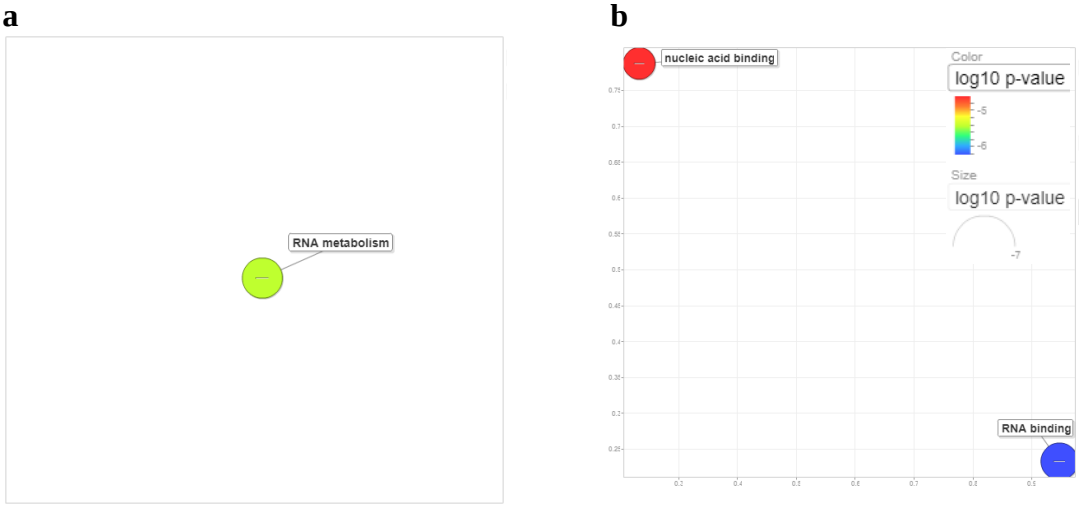


Figure 4.16: Salmon module over-represented GO terms. a) Biological Process GO terms b) Molecular Function GO terms. There were no cellular components GO terms that were enriched.

The pink and purple modules were over-represented for terms associated with the mitochondrion. These terms included mitochondrial protein complex, mitochondrial matrix, mitochondrial large ribosomal subunit in the purple module (Figure 4.17), and cytochrome complex in the pink module (Figure 4.18). The darkturquoise module was over-represented for GO terms associated with transfer of phosphorous containing groups (Figure 4.19).

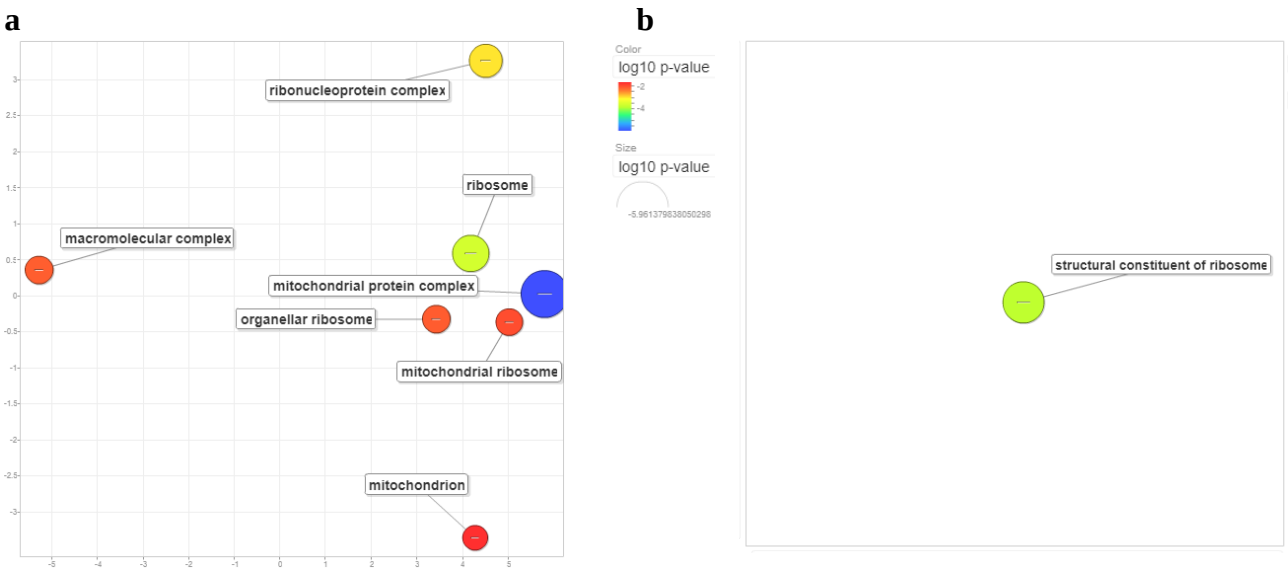


Figure 4.17: Purple module over-represented GO terms. a) Cellular Component GO terms b) Molecular Function GO terms. There were no biological process GO terms that were enriched.

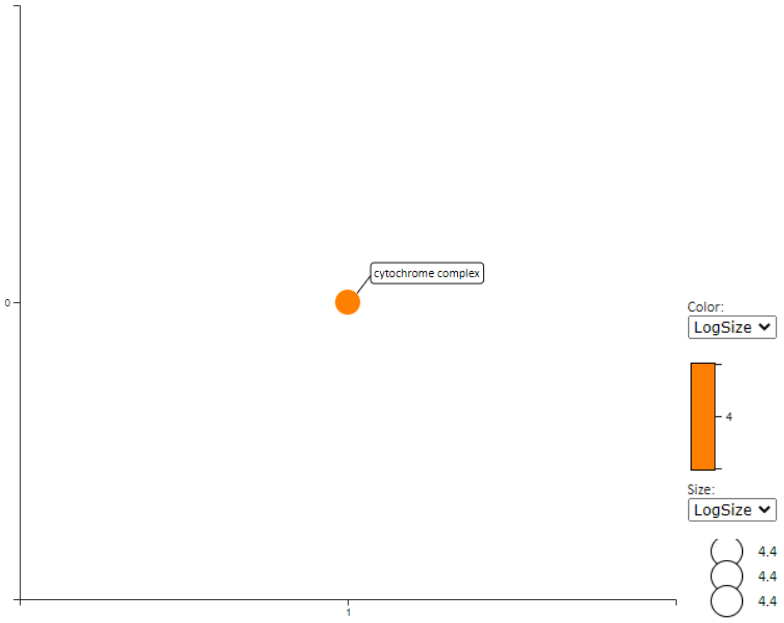


Figure 4.18: Pink module over-represented cellular component GO terms. There were no biological process and molecular functions GO terms that were enriched.

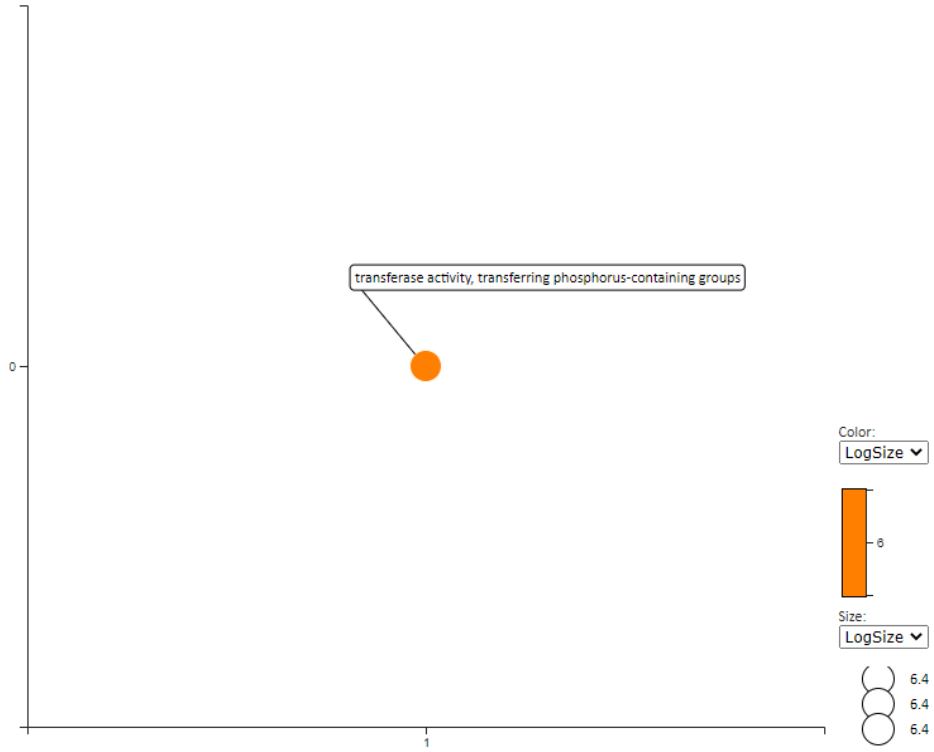


Figure 4.19: Darkturquoise module over-represented molecular function GO terms. There were no biological process and cellular components GO terms that were enriched.

Seven (7) out of the 27 modules were enriched following KEGG pathway enrichment analysis, from which five (5) were over-represented and two (2) (lightcyan and blue modules) were under-represented for KEGG pathway terms (Table 4.2). Out of the 12 modules with over-represented GO terms, four (4) modules were over-represented for KEGG pathway terms, and one module (yellow module) was over-represented for a KEGG pathway term (endocytosis), but not GO terms.

Table 4.2: **KEGG enrichment for seven modules.**

Module	**GO Category**	**Genes in GO Category**	**Total Genes**	**Adjusted P-Value**	**KEGG Term**	**Enrichment Status**
Lightyellow	path:tbr03010	26	134	1e-07	Ribosome	Over-represented
Lightcyan	path:tbr01100	2	378	0.0078292	Metabolic pathways	Under-represented
Magenta	path:tbr00230	12	82	0.0088518	Purine metabolism	Over-represented

Blue	path:tbr03010	1	134	0.0143821	Ribosome	Under-represented
Red	path:tbr03440	6	14	0.0148137	Homologous recombination	Over-represented
Yellow	path:tbr04144	8	35	0.0249234	Endocytosis	Over-represented
Pink	path:tbr03013	8	42	0.0318986	RNA transport	Over-represented

4.3.2 Module hub genes

Highly connected genes in a module are referred to as intra-modular hub genes. Higher connectivity for a gene implies more importance in the module's functional role. These hub genes are considered functionally significant in the enriched functions of the modules. Genes with the highest connectivity in the 27 modules were determined and considered to be the hub genes. Hub genes for the 12 modules with over-represented GO terms are shown in Table 4.3.

Table 4.3: **Identified hub genes and their encoding proteins for the 12 modules with over-represented GO terms.**

Module	Hub gene	Encoding protein
Brown	Tb927.11.1570	Hypothetical protein, conserved
Black	Tb927.7.1790	Adenine phosphoribosyltransferase, putative
Pink	Tb927.10.6200	Hypothetical protein, conserved
Darkturquoise	Tb927.8.6650	RNA-binding protein, putative
Salmon	Tb927.11.1450	2-oxoglutarate dehydrogenase E1 component, putative
Purple	Tb927.1.600	Phosphate-repressible phosphate permease, putative
Lightyellow	Tb927.10.2560	Mitochondrial malate dehydrogenase

Red	Tb927.7.6920	Hypothetical protein, conserved
Tan	Tb927.3.2930	RNA-binding protein RBP6, putative
Greenyellow	Tb927.7.920	Inner arm dynein 5-1
Magenta	Tb927.9.6290	Arginine kinase
Turquoise	Tb927.9.15630	BARP protein

4.4 Prediction of 3' untranslated regions (UTRs)

Genes in a given module are hypothesized to be co-regulated as they are assumed to have similar functions. Consequently, their *cis*-regulatory element should be similar. Following this hypothesis, a total of 10 statistically significant RNA motifs each over-represented in different gene modules were identified using FIRE (Figure 4.20).

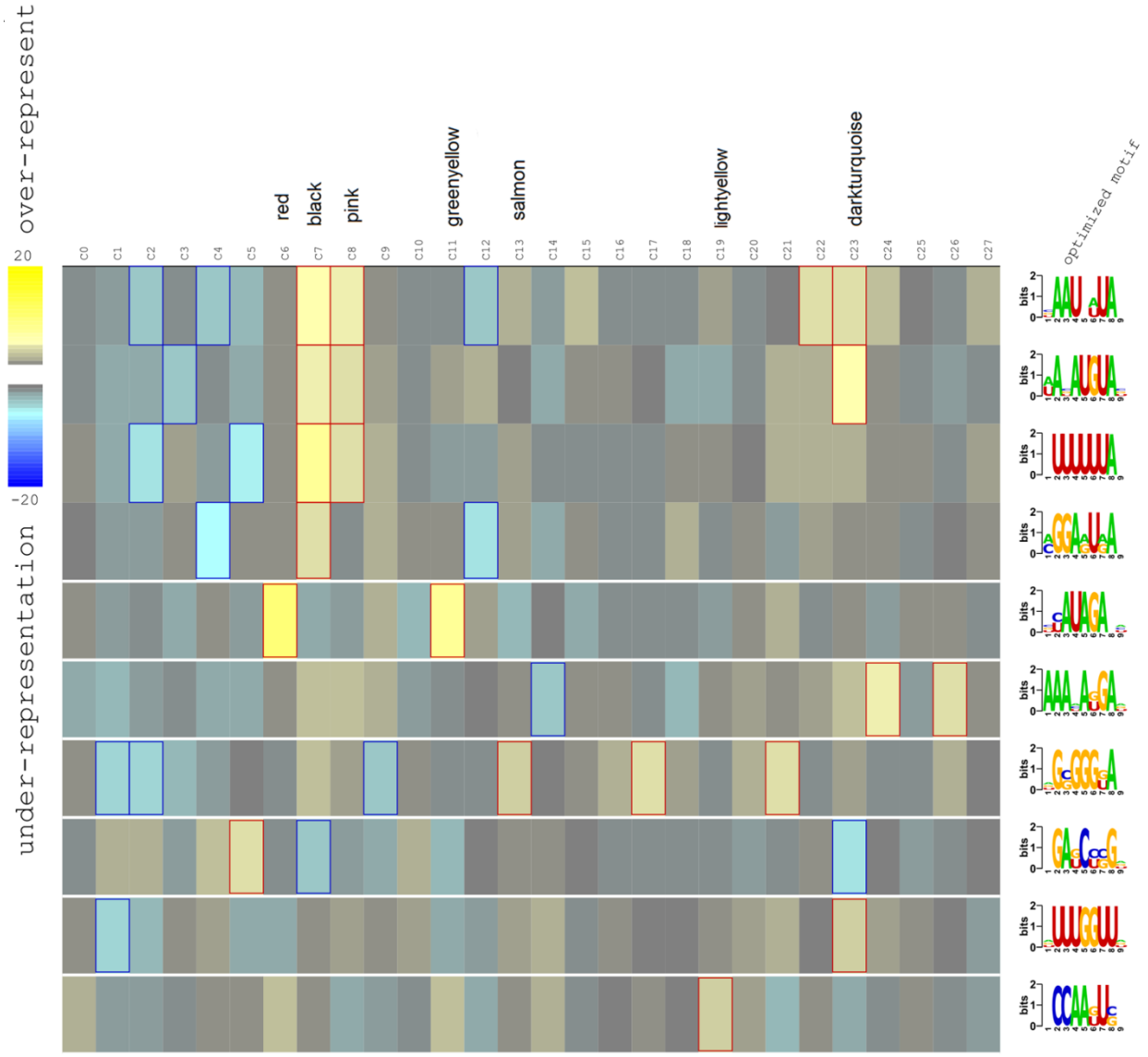


Figure 4.20: **Prediction of regulatory elements in the 3' untranslated regions (UTR) based on gene co-expression modules. A heatmap of the predicted motifs for the gene modules is shown. Columns represent gene modules, while rows represent the predicted motifs with consensus sequence on the right side. Over-representation of a motif for a given gene module is indicated by yellow color with significant over-representation highlighted by red frames. Blue color map and frames indicate under-representation.**

The motifs [CGU]AAU.[AU]UA. and [AU]A[CGU]AUGUA[CGU] were over-represented in the black, pink, and darkturquoise modules. The motif .UUUUUUA. was over-represented in the black and pink modules. [AC]GGA[AG]U[AG]A. was over-represented in the red and greenyellow modules (Figure 4.8). [AU]A[CGU]AUGUA[CGU] co-occurs with other motifs which are [CGU]AAU.[AU]UA., .UUUUUUA., [AC]GGA[AG]U[AG]A. and [AGU]UUUGGUU[AGU] (lighter colors in Figure 4.21).

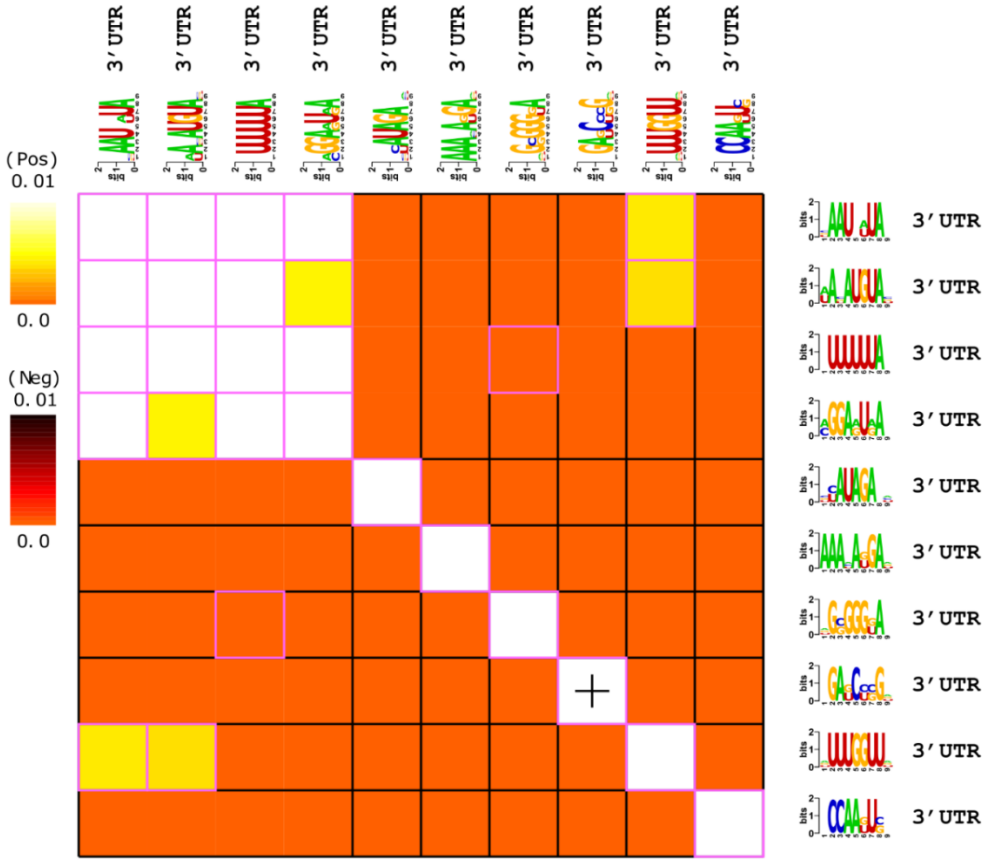


Figure 4.21: **Motif pairs co-occurring in the 3' UTR are shown in the heatmap where each row and each column correspond to a predicted motif. Light colors indicate the presence of another motif within the same 3' UTR while dark colors indicate that the motifs are absent in the same 3' UTR. "+" indicates significant spatial co-localization between pairs of motifs.**

CHAPTER FIVE

DISCUSSION

This study constructed the *T. brucei* weighted gene co-expression network using the WGCNA method and identified highly connected genes in a module, known as intra-modular hub genes (Langfelder & Horvath, 2008). These hub genes may be key drivers of a module's molecular process or act as a representative of the predominant biological function of the module. The modules were enriched in functional roles associated with cell cycle, cell signaling, mitochondrion, protein biosynthesis, and cell surface. These functional roles and their associated genes are discussed below and indicate that the network recapitulated known biology about the parasite.

T. brucei has a complex life cycle involving differentiation and proliferative cell divisions both in the human host and insect vector. Significant progress has been made over the years in understanding the cell cycle in *T. brucei*. (Farr & Gull, 2012; Hu *et al.*, 2019; Zhou *et al.*, 2014, 2016a, 2018a). Several genes involved in the cell cycle were assigned to the red and tan modules which were enriched for GO terms associated with the cell cycle. The red module was functionally enriched for GO terms such as DNA replication and chromosome organization, and KEGG pathway term homologous recombination. It contained genes such as BOH1 (Tb927.10.12720), which cooperates with *Tb*PLK to initiate cytokinesis and flagellum inheritance (Pham *et al.*, 2019), and Cytokinesis Initiation Factor 2 (CIF2) (Tb927.9.14290) which is involved in the initiation of cytokinesis (Zhou *et al.*, 2016b). Additionally, it contained nucleus- and spindle-associated protein 1 (NuSAP1) (Tb927.11.8370) that is required in chromosome segregation and NuSAP2 (Tb927.9.6110) that promotes G_2/M transition (Zhou *et al.*, 2018b) indicating that its genes are involved in the progression of the cell cycle. The hub gene for the red module is a hypothetical gene (Tb927.7.6920) that may play a key role in the progression of the cell cycle and requires further study to ascertain its role in the cell cycle.

The tan module's enriched GO terms were spindle pole and microtubule cytoskeleton. Among the assigned genes include CIF4 (Tb927.10.8240), TLK1 (Tb927.4.5180), and FPRC (Tb927.10.6360) that are involved in cytokinesis (Hu *et al.*, 2019; Li *et al.*, 2008). The hub gene for the tan module is RNA-binding protein RBP6. Its over-expression *in vitro* has been demonstrated to recapitulate the parasite's tsetse fly stage developmental form that was previously elusive in culture (Kolev *et al.*, 2012). Although the exact role of RBP6 during the parasite's development in the tsetse fly is unknown, based on its assignment to the tan module, it is likely to be involved in regulating a key step during the progression of the cell cycle.

Signal transduction is essential during the life cycle of *T. brucei* as it enables the parasite to sense and adapt to the changing environment in the mammalian host and the insect vector. The parasite has been documented to utilize signaling for evading the host (Salmon *et al.*, 2012b), social motility (Imhof *et al.*, 2014; Oberholzer *et al.*, 2015), and cytokinesis (Salmon *et al.*, 2012a). In the tsetse fly, the parasite interacts with the vector's tissue surfaces while moving across and through tissues from the midgut to the salivary glands via the proventriculus. This requires not only the parasite sensing its environment, but also its motility (Rotureau *et al.*, 2014). Adenylate cyclase (AC) genes are among the genes involved in signaling and have been identified in procyclic trypanosomes (Saada *et al.*, 2014). Among them, ACP3 (Tb927.7.7470), ACP4 (Tb927.10.13040), ACP5 (Tb927.11.13740), and ACP6 (Tb927.9.15660) were assigned to the black module which was enriched in adenylate cyclase GO terms. Adenylate cyclases catalyze the formation of cyclic AMP (cAMP) from ATP. Additionally, ACs have been proposed to interact with extracellular ligands directly owing to the lack of G-protein coupled receptors (GPCRs) in *T. brucei* (Seebeck *et al.*, 2001). The hub gene for the black module is adenine phosphoribosyltransferase (APRT) which plays a crucial role in the purine salvage pathway in *T. brucei*, as this parasite lacks a *de novo* purine biosynthetic pathway (Hammond & Gutteridge, 1984). Purine nucleotides are precursors of DNA and RNA as well as constituents of second messengers in signaling such as

cyclic AMP. In this regard, APRT is important in enriched module functions such as cyclic nucleotide biosynthesis and consequently, signaling functional roles of the black module. Further studies are required to reveal the role of APRT in signaling.

Besides the black module, the brown module was also enriched in functions associated with signaling as depicted by GO terms such as adenylate cyclase signaling activity. The brown module contained Proteins Associated with Differentiation (PADs) – PAD1 and PAD3. PADs are a carboxylate-transporter family that is required for the perception of the differentiation signal (Dean *et al.*, 2009). Other PAD genes were included in the black module: PAD2 (Tb927.7.5940), PAD5 (Tb927.7.5970), and PAD7 (Tb927.7.5990). The hub gene for the brown module is a hypothetical gene (Tb927.11.1570) which may be crucial in the signaling in *T. brucei* during its life cycle in the tsetse fly.

T. brucei expresses different surface coats in its life cycle stages both in the mammalian host and insect vector (Pays & Nolan, 1998; Urwyler *et al.*, 2007). Insect-stage trypanosomes express two classes of procyclins; GPEET and EP (EP1, EP2, and EP3 isoforms) (Vassella *et al.*, 2001), and GPI-anchored proteins known as *brucei* alanine-rich proteins (BARPs) (Urwyler *et al.*, 2007). Procyclin genes were assigned to the black module (EP1 (Tb927.10.10260) and GPEET (Tb927.6.510)) and pink module (EP2 (Tb927.10.10250) and EP3-2 (Tb927.6.520)). However, it is not clear what roles these genes play concerning enriched functions of the modules in which they are assigned. The GPEET procyclin is the surface coat of early procyclic trypanosomes and is replaced by the EP procyclins in late procyclic trypanosomes (Acosta-Serrano *et al.*, 2001). The BARP protein genes were assigned to the turquoise module which was enriched in GO terms associated with the cell membrane and cell surface. BARP is found in epimastigote trypanosomes (Urwyler *et al.*, 2007) and its role is presently unknown. The hub gene for the turquoise module is BARP (Tb927.9.15630) and could be important for the survival of the parasite in the insect vector. Out of 16 BARP protein genes in the network, 14 were assigned to the turquoise module while two (2) were

assigned to the blue module which was under-represented for GO enrichment terms. Although the complete set of functions of these surface coats are yet to be elucidated, the procyclins have been suggested to protect the parasite from the proteolytic environment in the tsetse fly midgut (Acosta-Serrano *et al.*, 2001).

Protein biosynthesis was associated with over-represented terms in the black and lightyellow module. The hub gene for the black module is adenine phosphoribosyltransferase (APRT). APRT plays a crucial role in the purine salvage pathway in *T. brucei*, as this parasite lacks a *de novo* purine biosynthetic pathway (Hammond & Gutteridge, 1984). Purine nucleotides are precursors of DNA and RNA as well as constituents of second messengers in signaling such as cyclic AMP. In this regard, APRT is important in enriched module functions such as cyclic nucleotide biosynthesis and synthesis of the structural constituent of the ribosome particularly ribosomal RNA, and consequently, signaling and protein biosynthesis functional roles of the black module. The lightyellow module was enriched in functions associated with gene expression, translation, protein metabolic process, and the ribosome. Its hub gene is mitochondrial malate dehydrogenase (Tb927.10.2560). However, mitochondrial malate dehydrogenase has no direct involvement in protein biosynthesis and has been suggested to be involved in fatty acids biosynthesis (van Weelden *et al.*, 2005). Since *T. brucei* undergoes metabolic reprogramming during its life cycle to adjust to the energy source (Smith *et al.*, 2017), mitochondrial malate dehydrogenase may be involved in this process. However, further studies are required to establish its role in metabolic reprogramming of the parasite during its development in the tsetse fly.

The salmon module was enriched in RNA binding and RNA metabolism and its hub gene is 2-oxoglutarate dehydrogenase E1 component which is a tricarboxylic acid cycle (TCA) enzyme. 2-oxoglutarate dehydrogenase in the mitochondrion has been implicated in the degradation of proline and glutamate to succinate which may subsequently be used for gluconeogenesis in procyclic trypanosomes (van Weelden *et al.*, 2005). This hub gene may be important in the role of the mitochondrion in responding to stress as a

result of a change in energy source in insect-stage trypanosomes (Kramer *et al.*, 2013). An RNA-binding protein related to stress response, ZC3H30 (Tb927.10.1540), together with an associated stress response granule (Tb927.8.3820) (Chakraborty & Clayton, 2018), were assigned to the salmon module. The enriched functions of RNA binding may either involve binding of the mRNA by RNA-binding proteins (RBPs) as a post-transcriptional gene regulation mechanism in *T. brucei* (Clayton, 2013; Kolev *et al.*, 2014) or binding by translation initiation factors for protein synthesis (Clayton & Shapira, 2007).

T. brucei genes have a polycistronic arrangement and therefore regulation of gene expression occurs almost exclusively post-transcriptionally (Clayton, 2002; Clayton & Shapira, 2007; Queiroz *et al.*, 2009). Post-transcriptional regulation of mRNA abundance mainly involves the interaction of the *cis*-regulatory element and a *trans*-acting element such as an RNA-binding protein (Haile & Papadopoulou, 2007). Genes with similar functions are co-regulated together thus their mRNAs are hypothesized to have similar *cis*-regulatory elements (Elemento *et al.*, 2007). Since the gene modules of a co-expression network are composed of genes with similar functions, they can be used as a basis for identifying potential regulatory elements in the untranslated regions of mRNA.

The predicted motif [AU]A[CGU]AUGUA[CGU] was over-represented in the black, pink, and darkturquoise modules. The motif contains the core sequence, UGUA, that is recognized by the PUF family of RNA-binding proteins (Gerber *et al.*, 2006) and has previously been identified in *T. brucei* as targeting transcripts involved in the cell cycle (Archer *et al.*, 2009, 2011; Najafabadi *et al.*, 2013). The motif co-occurs with other motifs which are [CGU]AAU.[AU]UA., .UUUUUUA., [AC]GGA[AG]U[AG]A. and [AGU]UUUGGUU[AGU] (lighter colors in Figure 4.21). This means that they co-localize within the same untranslated region (UTR) which indicates that the presence of one motif implies the presence of its putative counterpart (Elemento *et al.*, 2007). Their

co-localization may provide more information on post-transcriptional regulation because it could imply the physical interaction of their binding elements, hence their functional interaction (Elemento *et al.*, 2007).

The motif [CGU][CU]AUAGA.[ACU] was over-represented in the red and greenyellow modules. It contains the core AUAGA sequence similar to CAUAGAA that has been implicated in cell cycle regulation (Avliyakulov *et al.*, 2003; Mahmood *et al.*, 1999), and was previously predicted in *T. brucei* (Najafabadi *et al.*, 2013). Notably, genes in the red module were enriched for cell cycle functions while those in the greenyellow module were enriched for microtubule-associated functions, including motility. Motility in *T. brucei* is mediated through the flagellum (Langousis & Hill, 2014). Importantly, flagellum motility is essential for completion of the cell division (Broadhead *et al.*, 2006; Ralston *et al.*, 2006) suggesting co-regulation of genes in the greenyellow module together with those in the red module. These findings indicate the importance of post-transcriptional gene regulation and could provide insights into key regulatory functions that could be targeted to control the transmission of trypanosomiasis.

CHAPTER SIX

CONCLUSION & RECOMMENDATIONS

6.1 Conclusion

The life cycle stages of *T. brucei* in the insect vector are less explored and therefore the parasite's biology during this stage is largely unknown. It is crucial to understand the parasite's biology in tsetse fly as it provides insights on mechanisms for interrupting disease transmission. This study aimed at identifying key genes involved in *T. brucei* development in the tsetse fly, the functional roles which these genes are associated with, and prediction of their 3' untranslated regions. From this study:

i. A gene co-expression network was constructed. The enriched functional roles of the network modules are associated with cell cycle, cell signaling, mitochondrion, protein biosynthesis, metabolic regulation, and cell surface. These enriched functional roles highlight important processes and cell features that may be involved in *T. brucei* development in tsetse fly and which may be targeted to cut short the parasite's division in tsetse fly.

ii. Genes that have previously been identified as crucial in *T. brucei* development were among the identified hub (key) genes indicating that they may play crucial role in *T. brucei* life cycle in tsetse fly. Aditionally, genes with unknown functions were identified as hub genes and should be the focus of futher studies to determine their function. Together, these hub genes have the potential for being trypanocidal drug targets to stop the parasite's life cycle in the tsetse fly thus controlling trypanosomiasis transmission.

iii. The 3' untranslated region motifs involved in post-transcriptional gene regulation were predicted. Some of these motifs had consensus sequences that have been previously implicated in gene regulation in *Trypanosoma brucei* and possibly play key regulatory roles during its development in tsetse fly.

These findings indicated that the co-expression network recapitulated known trypanosome biology hence its usefulness in providing insights into parasite

development in the tsetse fly. One of the limitations, however, is that these findings pertain to *T. brucei* and other species of trypanosomes may have different genes involved during their development if they have differerent insect vectors or life cycle stages. Nevertheless, this study provides a resource for network-based data mining to identify candidate genes for functional studies in *T. brucei* and may inform studies on key candidate genes in other species of trypanosomes. Understanding trypanosome biology in tsetse will increase the chances of identifying potential molecular targets for disease control.

6.2 Recommendations

1. More transcriptomics data should be generated for the life cycle stages of *T. brucei* as it traverses the tsetse fly vector. The data will aid in creation of more co-expression networks leading to robust characterization of genes involved in the development of the parasite in tsetse fly.

2. The three (3) hypothetical protein genes that were identified as hub genes should be studied through techniques such as RNAi knockdown to experimentally validate their functional roles in trypanosome development in the tsetse fly.

3. The functional roles of 3' untranslated regions (UTR) motifs should be studied further to determine their activities on gene regulation through mechanisms such as regulating mRNA stability and translation.

REFERENCES

Acosta-Serrano, A., Vassella, E., Liniger, M., Renggli, C. K., Brun, R., Roditi, I., & Englund, P. T. (2001). The surface coat of procyclic *Trypanosoma brucei*: Programmed expression and proteolytic cleavage of procyclin in the tsetse fly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 1513–1518.

Adam, Y., Cecchi, G., Kgori, P. M., Marcotty, T., Mahama, C. I., Abavana, M., … Bouyer, J. (2013). The sequential aerosol technique: A major component in an integrated strategy of intervention against Riverine Tsetse in Ghana. *PLoS Neglected Tropical Diseases*, *7*, e2135.

Aksoy, S., Gibson, W. C., & Lehane, M. J. (2003). Interactions between tsetse and trypanosomes with implications for the control of trypanosomiasis. *Advances in Parasitology*, *53*, 1–83.

Aksoy, S., Weiss, B., & Attardo, G. (2008). Paratransgenesis applied for control of tsetse transmitted sleeping sickness. *Advances in Experimental Medicine and Biology*, *627*, 35–48.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, *118*, 4947–4957.

Allsopp, R., & Hursey, B. (2004). *Insecticidal control of tsetse*. Oxfordshire, UK: Cabi Publishing.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*, 166–169.

Angara, T., Ismail, A., & Ibrahim, A. (2014). An overview on the economic impacts of animal trypanosomiasis. *Global J Res Anal*, *3*, 275–276.

Archer, S. K., Inchaustegui, D., Queiroz, R., & Clayton, C. (2011). The cell cycle regulated transcriptome of *Trypanosoma brucei*. *PLOS ONE*, *6*, e18425.

Archer, S. K., Luu, V.-D., de Queiroz, R. A., Brems, S., & Clayton, C. (2009). *Trypanosoma brucei* PUF9 regulates mRNAs for proteins involved in replicative processes over the Cell Cycle. *PLoS Pathogens*, *5*, e1000565.

Armitage, E. G., & Barbas, C. (2014). Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis, 87,* 1–11.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics, 25*, 25–29.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., … Wang, H. (2010). TriTrypDB: A functional genomic resource for the *Trypanosomatidae*. *Nucleic Acids Research, 38*, D457–D462.

Avliyakulov, N. K., Hines, J. C., & Ray, D. S. (2003). Sequence elements in both the intergenic space and the 3′ untranslated region of the *Crithidia fasciculata* KAP3 gene are required for cell cycle regulation of KAP3 mRNA. *Eukaryotic Cell, 2*, 671–677.

Awuoche, E. O. (2012). Tsetse fly saliva: Could it be useful in fly infection when feeding in chronically aparasitemic mammalian hosts. *Open Veterinary Journal, 2*, 95–105.

Bader, G. D., Betel, D., & Hogue, C. W. V. (2003). BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Research, 31*, 248–250.

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews. Genetics, 5*, 101–113.

Bebek, G., & Yang, J. (2007). PathFinder: Mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics, 8*, 335.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., … El-Sayed, N. M. (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science (New York, N.Y.), 309*, 416–422.

Blais, A., & Dynlacht, B. D. (2005). Constructing transcriptional regulatory networks. *Genes & Development, 19*, 1499–1511.

Brightwell, R., Dransfield, R., Kyorku, C., Golder, T., Tarimo, S., & Mungai, D. (1987). A new trap for *Glossina pallidipes*. *International Journal of Pest Management*, *33*, 151–159.

Broadhead, R., Dawe, H. R., Farr, H., Griffiths, S., Hart, S. R., Portman, N., … Gull, K. (2006). Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature*, *440*, 224–227.

Bruce, D. (1897). *Further report on the tsetse fly disease or nagana, in Zululand*. Durban: Bennett and Davis.

Bullard, W., Kieft, R., Capewell, P., Veitch, N. J., Macleod, A., & Hajduk, S. L. (2012). Haptoglobin-hemoglobin receptor independent killing of African trypanosomes by human serum and trypanosome lytic factors. *Virulence*, *3*, 72–76.

Bursell, E. (1966). Aspects of the flight metabolism of tsetse flies (*Glossina*). *Comparative Biochemistry and Physiology*, *19*, 809–818.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. *BMC Genomics*, *7*, 40.

CDC. (2020, September 29). African trypanosomiasis. Retrieved December 15, 2020, from https://www.cdc.gov/parasites/sleepingsickness/index.html

Cecchi, G., Mattioli, R. C., Slingenbergh, J., & de la Rocque, S. (2008). Land cover and tsetse fly distributions in sub-Saharan Africa. *Medical and Veterinary Entomology*, *22*, 364–373.

Chakraborty, C., & Clayton, C. (2018). Stress susceptibility in *Trypanosoma brucei* lacking the RNA-binding protein ZC3H30. *PLoS Neglected Tropical Diseases*, *12*, e0006835.

Challier, A., & Laveissiere, C. (1973). A new trap for catching *Glossina*: Description and field trials. *Cahiers ORSTOM, Serie Entomologie Medicale et Parasitologie*, *11*, 251–62.

Cheng, Q., & Aksoy, S. (1999). Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. *Insect Molecular Biology*, *8*, 125–132.

Clayton, C. (2013). The regulation of Trypanosome gene expression by RNA-Binding proteins. *PLoS Pathogens*, *9*, e1003680.

Clayton, C. (2019). Regulation of gene expression in trypanosomatids: Living with polycistronic transcription. *Open Biology*, *9*.

Clayton, C. E. (2002). Life without transcriptional control? From fly to man and back again. *The EMBO Journal*, *21*, 1881–1888.

Clayton, C., & Shapira, M. (2007). Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Molecular and Biochemical Parasitology*, *156*, 93–101.

Cooper, J., & Dobson, H. (1993). *Aerial spraying for tsetse fly control: A handbook of aerial spray calibration and monitoring for the sequential aerosol technique*. Natural Resources Institute.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., … Boone, C. (2010). The genetic landscape of a cell. *Science (New York, N.Y.)*, *327*, 425–431.

Cottret, L., & Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. *Parasitology*, *137*, 1393–1407.

Coutinho-Abreu, I. V., Zhu, K. Y., & Ramalho-Ortigao, M. (2010). Transgenesis and paratransgenesis to control insect-borne diseases: Current status and future challenges. *Parasitology International*, *59*, 1–8.

De Las Rivas, J., & Fontanillo, C. (2012). Protein-protein interaction networks: Unraveling the wiring of molecular machines within the cell. *Briefings in Functional Genomics*, *11*, 489–496.

De Vooght, L., Caljon, G., De Ridder, K., & Van Den Abbeele, J. (2014). Delivery of a functional anti-trypanosome Nanobody in different tsetse fly tissues via a bacterial symbiont, *Sodalis glossinidius*. *Microbial Cell Factories*, *13*, 156.

De Vooght, L., Caljon, G., Stijlemans, B., De Baetselier, P., Coosemans, M., & Van Den Abbeele, J. (2012). Expression and extracellular release of a functional anti-trypanosome Nanobody® in *Sodalis glossinidius*, a bacterial symbiont of the tsetse fly. *Microbial Cell Factories*, *11*, 23.

De Vooght, L., Van Keer, S., & Van Den Abbeele, J. (2018). Towards improving tsetse fly paratransgenesis: Stable colonization of *Glossina morsitans morsitans* with genetically modified *Sodalis*. *BMC Microbiology*, *18*, 165.

Dean, S., Marchetti, R., Kirk, K., & Matthews, K. R. (2009). A surface transporter family conveys the trypanosome differentiation signal. *Nature*, *459*, 213–217.

Deken, R. D., & Bouyer, J. (2018). Can sequential aerosol technique be used against riverine tsetse? *PLOS Neglected Tropical Diseases*, *12*, e0006768.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology*, *23*, 1499–1501.

Dransfield, R. D., Williams, B. G., & Brightwell, R. (1991). Control of tsetse flies and trypanosomiasis: Myth or reality? *Parasitology Today (Personal Ed.)*, *7*, 287–291.

Dutton, J. E. (1902). Preliminary note upon a trypanosome occurring in the blood of man. *Thompson Yates Lab Rep*, *4*, 455–468.

Dyer, N. A., Rose, C., Ejeh, N. O., & Acosta-Serrano, A. (2013). Flying tryps: Survival and maturation of trypanosomes in tsetse flies. *Trends in Parasitology*, *29*, 188–196.

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

El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A.-N., … Andersson, B. (2005). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science (New York, N.Y.)*, *309*, 409–415.

Enserink, M. (2007). Welcome to Ethiopia's Fly Factory. *Science*, *317*, 310–313.

Farr, H., & Gull, K. (2012). Cytokinesis in trypanosomes. *Cytoskeleton*, *69*, 931–941.

Fèvre, E. M., Picozzi, K., Jannin, J., Welburn, S. C., & Maudlin, I. (2006). Human African trypanosomiasis: Epidemiology and control. *Advances in Parasitology*, *61*, 167–221.

Fèvre, E. M., Wissmann, B. v., Welburn, S. C., & Lutumba, P. (2008). The burden of Human African Trypanosomiasis. *PLoS Neglected Tropical Diseases*, 2.

Flint, S. (1985). A comparison of various traps for *Glossina spp.* (*Glossinidae*) and other *Diptera*. *Bulletin of Entomological Research*, *75*, 529–534.

Ford, J., & Blaser, E. (1971). Some aspects of cattle raising under prophylactic treatment against trypanosomiasis on the Mkwaja Ranch, Tanzania. *Acta Tropica*, *28*, 69–79.

Forde, R. M. (1902). Some clinical notes on a European patient in whose blood a trypanosome was observed. *J Trop Med*, *5*, 261–263.

Franco, J., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., … Argaw, D. (2018). Monitoring the elimination of human African trypanosomiasis: Update to 2016. *PLoS Neglected Tropical Diseases*, *12*, e0006890.

Franco, J. R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., … Argaw, D. (2020). Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018. *PLOS Neglected Tropical Diseases*, *14*, e0008261.

Franco, J., Simarro, P. P., Diarra, A., & Jannin, J. G. (2014). Epidemiology of human African trypanosomiasis. *Clinical Epidemiology*, *6*, 257–275.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., & Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference. *Nucleic Acids Research*, *40*, W147–W151.

Gazestani, V. H., Nikpour, N., Mehta, V., Najafabadi, H. S., Moshiri, H., Jardim, A., & Salavati, R. (2016). A protein complex map of *Trypanosoma brucei*. *PLoS Neglected Tropical Diseases*, *10*, e0004533.

Gerber, A. P., Luschnig, S., Krasnow, M. A., Brown, P. O., & Herschlag, D. (2006). Genome-wide identification of mRNAs associated with the translational

regulator PUMILIO in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 4487–4492.

Gibson, G., & Torr, S. J. (1999). Visual and olfactory responses of haematophagous *Diptera* to host stimuli. *Medical and Veterinary Entomology*, *13*, 2–23.

Gibson, W., & Peacock, L. (2019). Fluorescent proteins reveal what trypanosomes get up to inside the tsetse fly. *Parasites & Vectors*, *12*, 6.

Gomperts, B. D., Gomberg, E. S. L., Bastien, D. G., Tatham, P. E., & others. (2015). *Signal transduction* (2nd ed.). Academic Press.

Green, C., & Flint, S. (1986). An analysis of colour effects in the performance of the F2 trap against *Glossina pallidipes* Austen and *G. morsitans morsitans* Westwood (*Diptera: Glossinidae)*. *Bulletin of Entomological Research*, *76*, 409–418.

Guruharsha, K. G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., … Artavanis-Tsakonas, S. (2011). A protein complex network of *Drosophila melanogaster*. *Cell*, *147*, 690–703.

Haile, S., & Papadopoulou, B. (2007). Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Current Opinion in Microbiology*, *10*, 569–577.

Haines, L. R., Hancock, R. E. W., & Pearson, T. W. (2003). Cationic antimicrobial peptide killing of African trypanosomes and *Sodalis glossinidius*, a bacterial symbiont of the insect vector of sleeping sickness. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *3*, 175–186.

Hammond, D. J., & Gutteridge, W. E. (1984). Purine and pyrimidine metabolism in the *trypanosomatidae*. *Molecular and Biochemical Parasitology*, *13*, 243–261.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, *402*, C47-52.

Haury, A.-C., Mordelet, F., Vera-Licona, P., & Vert, J.-P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, *6*, 145.

Hay, S. I., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., … Murray, C. J. L. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: A systematic analysis for the global burden of disease study 2016. *The Lancet*, *390*, 1260–1344.

He, B., & Tan, K. (2016). Understanding transcriptional regulatory networks using computational models. *Current Opinion in Genetics & Development*, *37*, 101–108.

Hegedus, D., Erlandson, M., Gillott, C., & Toprak, U. (2009). New insights into peritrophic matrix synthesis, architecture, and function. *Annual Review of Entomology*, *54*, 285–302.

Hendriks, E. F., & Matthews, K. R. (2005). Disruption of the developmental programme of *Trypanosoma brucei* by genetic ablation of TbZFP1, a differentiation-enriched CCCH protein. *Molecular Microbiology*, *57*, 706–716.

Hendriks, E. F., Robinson, D. R., Hinkins, M., & Matthews, K. R. (2001). A novel CCCH protein which modulates differentiation of *Trypanosoma brucei* to its procyclic form. *The EMBO Journal*, *20*, 6700–6711.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., … Apweiler, R. (2004). IntAct: An open source molecular interaction database. *Nucleic Acids Research*, *32*, D452–D455.

Holmes, P. (2013). Tsetse-transmitted trypanosomes—Their biology, disease impact and control. *Journal of Invertebrate Pathology*, *112 Suppl*, S11-14.

Holmes, P. (2014). First WHO meeting of stakeholders on elimination of Gambiense Human African Trypanosomiasis. *PLOS Neglected Tropical Diseases*, *8*, e3244.

Horvath, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. New York: Springer-Verlag. Retrieved from https://www.springer.com/gp/book/9781441988188

Hu, H., An, T., Kurasawa, Y., Zhou, Q., & Li, Z. (2019). The trypanosome-specific proteins FPRC and CIF4 regulate cytokinesis initiation by recruiting CIF1 to the

cytokinesis initiation site. *The Journal of Biological Chemistry*, *294*, 16672–16683.

Hu, Y., & Aksoy, S. (2005). An antimicrobial peptide with trypanocidal activity characterized from *Glossina morsitans morsitans*. *Insect Biochemistry and Molecular Biology*, *35*, 105–115.

Imhof, S., Knüsel, S., Gunasekera, K., Vu, X. L., & Roditi, I. (2014). Social motility of African trypanosomes is a property of a distinct life-cycle stage that occurs early in Tsetse fly transmission. *PLOS Pathogens*, *10*, e1004493.

Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M., … Myler, P. J. (2005). The genome of the Kinetoplastid parasite, *Leishmania major*. *Science (New York, N.Y.)*, *309*, 436–442.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651–654.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*, 27–30.

Kappmeier, K. (2000). A newly developed odour-baited "H trap" for the live collection of *Glossina brevipalpis* and *Glossina austeni* (*Diptera: Glossinidae*) in South Africa. *The Onderstepoort Journal of Veterinary Research*, *67*, 15–26.

Kariithi, H. M., Meki, I. K., Schneider, D. I., De Vooght, L., Khamis, F. M., Geiger, A., … Abd-Alla, A. M. M. (2018). Enhancing vector refractoriness to trypanosome infection: Achievements, challenges and perspectives. *BMC Microbiology*, *18*, 179.

Kgori, P. M., Modo, S., & Torr, S. J. (2006). The use of aerial spraying to eliminate tsetse from the Okavango Delta of Botswana. *Acta Tropica*, *99*, 184–199.

Kharchenko, P., Church, G. M., & Vitkup, D. (2005). Expression dynamics of a cellular metabolic network. *Molecular Systems Biology*, *1*, 2005.0016.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*, 357–360.

Koh, G. C. K. W., Porras, P., Aranda, B., Hermjakob, H., & Orchard, S. E. (2012). Analyzing Protein–Protein Interaction Networks. *Journal of Proteome Research*, *11*, 2014–2031.

Kolch, W., Halasz, M., Granovskaya, M., & Kholodenko, B. N. (2015). The dynamic control of signal transduction networks in cancer cells. *Nature Reviews. Cancer*, *15*, 515–527.

Kolev, N. G., Ramey-Butler, K., Cross, G. A. M., Ullu, E., & Tschudi, C. (2012). Developmental progression to infectivity in *Trypanosoma brucei* triggered by an RNA-Binding protein. *Science (New York, N.Y.)*, *338*, 1352–1353.

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

Krafsur, E. S. (2009). Tsetse flies: Genetics, evolution, and role as vectors. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *9*, 124–141.

Kramer, S., Bannerman-Chukualim, B., Ellis, L., Boulden, E. A., Kelly, S., Field, M. C., & Carrington, M. (2013). Differential localization of the two *T. brucei* Poly(A) binding proteins to the nucleus and RNP granules suggests binding to distinct mRNA pools. *PLoS ONE*, *8*, e54004.

Krinsky, W. L. (2019). Tsetse flies (*Glossinidae*). In G. R. Mullen & L. A. Durden (Eds.), *Medical and Veterinary Entomology (Third Edition)* (pp. 369–382). Academic Press.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics*, *24*, 719–720.

Langley, P. A., Hargrove, J. W., & Wall, R. L. (1990). Maturation of the tsetse fly *Glossina pallidipes* (*Diptera: Glossinidae*) in relation to trap-orientated behaviour. *Physiological Entomology*, *15*, 179–186.

Langousis, G., & Hill, K. L. (2014). Motility and more: The flagellum of *Trypanosoma brucei*. *Nature Reviews. Microbiology*, *12*, 505–518.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., … Collins, F. H. (2007). VectorBase: A home for invertebrate vectors of human pathogens. *Nucleic Acids Research*, *35*, D503–D505.

Leak, S., Mulatu, W., Rowlands, G., & d'Ieteren, G. (1995). A trial of a cypermethrin 'pour-on'insecticide to control *Glossina pallidipes, G. fuscipes fuscipes* and *G. morsitans submorsitans* (*Diptera: Glossinidae*) in South-west Ethiopia. *Bulletin of Entomological Research*, *85*, 241–251.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*, 882–883.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*, D28–D31.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., … Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science*, *303*, 540–543.

Li, Z., Umeyama, T., & Wang, C. C. (2008). The chromosomal passenger complex and a mitotic kinesin interact with the Tousled-Like kinase in trypanosomes to regulate mitosis and cytokinesis. *PLOS ONE*, *3*, e3814.

Ling, A. S., Trotter, J. R., & Hendriks, E. F. (2011). A Zinc Finger Protein, TbZC3H20, stabilizes two developmentally regulated mRNAs in trypanosomes. *The Journal of Biological Chemistry*, *286*, 20152–20162.

Lutumba, P., Robays, J., Bilenge, C. M. mia, Mesu, V. K. B. K., Molisho, D., Declercq, J., … Boelaert, M. (2005). Trypanosomiasis control, Democratic Republic of Congo, 1993–2003. *Emerging Infectious Diseases*, *11*, 1382–1388.

MacGregor, P., Szöoõr, B., Savill, N. J., & Matthews, K. R. (2012). Trypanosomal immune evasion, chronicity and transmission: An elegant balancing act. *Nature Reviews. Microbiology*, *10*, 431–438.

MacLeod, E. T., Maudlin, I., Darby, A. C., & Welburn, S. C. (2007). Antioxidants promote establishment of trypanosome infections in tsetse. *Parasitology*, *134*, 827–831.

Mahmood, R., Hines, J. C., & Ray, D. S. (1999). Identification of *cis* and *trans* Elements involved in the cell cycle regulation of multiple genes in *Crithidia fasciculata*. *Molecular and Cellular Biology*, *19*, 6174–6182.

Mantilla, B. S., Marchese, L., Casas-Sánchez, A., Dyer, N. A., Ejeh, N., Biran, M., … Silber, A. M. (2017). Proline metabolism is essential for *Trypanosoma brucei brucei* survival in the tsetse vector. *PLoS Pathogens*, *13*, e1006158.

Mao, L., Van Hemert, J. L., Dash, S., & Dickerson, J. A. (2009). Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, *10*, 346.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7*, S7.

Matthews, K. R. (2005). The developmental cell biology of *Trypanosoma brucei*. *Journal of Cell Science*, *118*, 283–290.

Matthews, K. R., McCulloch, R., & Morrison, L. J. (2015). The within-host dynamics of African trypanosome infections. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*, 20140288.

Miller, N., & Lehane, M. J. (1990). In vitro perfusion studies on the peritrophic membrane of the tsetse fly *Glossina morsitans morsitans* (*Diptera, Glossinidae*). *Journal of Insect Physiology*, *36*, 813–818.

Mitchell, P. J., & Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (New York, N.Y.)*, *245*, 371–378.

Mugo, E., & Clayton, C. (2017). Expression of the RNA-binding protein RBP10 promotes the bloodstream-form differentiation state in *Trypanosoma brucei*. *PLoS Pathogens*, *13*, e1006560.

Murray, C. J. (1994). Quantifying the burden of disease: The technical basis for disability-adjusted life years. *Bulletin of the World Health Organization*, *72*, 429–445.

Murray, C. J., & Lopez, A. D. (1996). Evidence-based health policy—Lessons from the global burden of disease study. *Science (New York, N.Y.)*, *274*, 740–743.

Mwiinde, A. M., Simuunza, M., Namangala, B., Chama-Chiliba, C. M., Machila, N., Anderson, N., … Welburn, S. C. (2017). Estimating the economic and social consequences for patients diagnosed with human African trypanosomiasis in Muchinga, Lusaka and Eastern Provinces of Zambia (2004–2014). *Infectious Diseases of Poverty*, *6*, 150.

Najafabadi, H. S., 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 Research*, *41*, 8591–8600.

Oberholzer, M., Saada, E. A., & Hill, K. L. (2015). Cyclic AMP regulates social behavior in African trypanosomes. *MBio*, *6*, e01954-14.

Ooi, C.-P., & Bastin, P. (2013). More than meets the eye: Understanding *Trypanosoma brucei* morphology in the tsetse. *Frontiers in Cellular and Infection Microbiology*, *3*, 71.

Ouellette, M., & Papadopoulou, B. (2009). Coordinated gene expression by post-transcriptional regulons in African trypanosomes. *Journal of Biology*, *8*, 100.

Pays, E., & Nolan, D. P. (1998). Expression and function of surface proteins in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, *91*, 3–36.

Pham, K. T. M., Zhou, Q., Kurasawa, Y., & Li, Z. (2019). BOH1 cooperates with Polo-like kinase to regulate flagellum inheritance and cytokinesis initiation in *Trypanosoma brucei*. *Journal of Cell Science*, 132.

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., … Golshani, A. (2006). PIPE: A protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, *7*, 365.

Politano, G., Orso, F., Raimo, M., Benso, A., Savino, A., Taverna, D., & Di Carlo, S. (2016). CyTRANSFINDER: A Cytoscape 3.3 plugin for three-component (TF, gene, miRNA) signal transduction pathway construction. *BMC Bioinformatics*, 17.

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

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ralston, K. S., Lerner, A. G., Diener, D. R., & Hill, K. L. (2006). Flagellar Motility contributes to cytokinesis in *Trypanosoma brucei* and is modulated by an evolutionarily conserved dynein regulatory system. *Eukaryotic Cell*, *5*, 696–711.

Rico, E., Rojas, F., Mony, B. M., Szoor, B., MacGregor, P., & Matthews, K. R. (2013). Bloodstream form pre-adaptation to the tsetse fly in *Trypanosoma brucei*. *Frontiers in Cellular and Infection Microbiology*, *3*, 78.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140.

Rotureau, B., Ooi, C.-P., Huet, D., Perrot, S., & Bastin, P. (2014). Forward motility is essential for trypanosome infection in the tsetse fly. *Cellular Microbiology*, *16*, 425–433.

Rotureau, B., & Van Den Abbeele, J. (2013). Through the dark continent: African trypanosome development in the tsetse fly. *Frontiers in Cellular and Infection Microbiology*, *3*, 53.

Saada, E. A., Kabututu, Z. P., Lopez, M., Shimogawa, M. M., Langousis, G., Oberholzer, M., … Hill, K. L. (2014). Insect stage-specific receptor adenylate cyclases Are localized to distinct subdomains of the *Trypanosoma brucei* flagellar membrane. *Eukaryotic Cell*, *13*, 1064–1076.

Salmon, D., Bachmaier, S., Krumbholz, C., Kador, M., Gossmann, J. A., Uzureau, P., … Boshart, M. (2012a). Cytokinesis of *Trypanosoma brucei* bloodstream forms depends on expression of adenylyl cyclases of the ESAG4 or ESAG4-like subfamily. *Molecular Microbiology*, *84*, 225–242.

Salmon, D., Vanwalleghem, G., Morias, Y., Denoeud, J., Krumbholz, C., Lhommé, F., … Pays, E. (2012b). Adenylate cyclases of *Trypanosoma brucei* inhibit the innate immune response of the host. *Science*, *337*, 463–466.

Savage, A. F., Kolev, N. G., Franklin, J. B., Vigneron, A., Aksoy, S., & Tschudi, C. (2016). Transcriptome profiling of *Trypanosoma brucei* development in the tsetse fly vector *Glossina morsitans*. *PLoS One*, *11*, e0168877.

Sayols, S., Scherzinger, D., & Klein, H. (2016). dupRadar: A Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. *BMC Bioinformatics*, *17*, 428.

Schmid, C. (2004). *10-day melarsoprol treatment of Trypanosoma brucei gambiense sleeping sickness: From efficacy to effectiveness* (PhD Thesis). University of Basel.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., & Schomburg, D. (2004). BRENDA, the enzyme database: Updates and major new developments. *Nucleic Acids Research*, *32*, D431–D433.

Seebeck, T., Gong, K., Kunz, S., Schaub, R., Shalaby, T., & Zoraghi, R. (2001). CAMP signalling in *Trypanosoma brucei*. *International Journal for Parasitology*, *31*, 491–498.

Sévin, D. C., Kuehne, A., Zamboni, N., & Sauer, U. (2015). Biological insights through nontargeted metabolomics. *Current Opinion in Biotechnology*, *34*, 1–8.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504.

Sharma, R., Gluenz, E., Peacock, L., Gibson, W., Gull, K., & Carrington, M. (2009). The heart of darkness: Growth and form of *Trypanosoma brucei* in the tsetse fly. *Trends in Parasitology*, *25*, 57–524.

Shaw, A., Robays, J., Fèvre, E. M., Lutumba, P., & Boelaert, M. (2010). The burden of Human African Trypanosomiasis. In *Handbook of diseasebBurdens and quality of life measures* (pp. 1433–1442). Springer, New York, NY. Retrieved from https://doi.org/10.1007/978-0-387-78665-0_83

Shereni, W., Anderson, N. E., Nyakupinda, L., & Cecchi, G. (2016). Spatial distribution and trypanosome infection of tsetse flies in the sleeping sickness focus of Zimbabwe in Hurungwe District. *Parasites & Vectors*, *9*, 605.

Simarro, P. P., Cecchi, G., Franco, J. R., Paone, M., Diarra, A., Ruiz-Postigo, J. A., … Jannin, J. G. (2012). Estimating and mapping the population at risk of sleeping sickness. *PLoS Neglected Tropical Diseases*, *6*, e1859.

Simarro, P. P., Cecchi, G., Paone, M., Franco, J. R., Diarra, A., Ruiz, J. A., … Jannin, J. G. (2010). The Atlas of human African trypanosomiasis: A contribution to global mapping of neglected tropical diseases. *International Journal of Health Geographics*, *9*, 57.

Simarro, P. P., Jannin, J., & Cattand, P. (2008). Eliminating Human African Trypanosomiasis: Where do we stand and what comes next? *PLoS Medicine*, *5*, e55.

Smith, T. K., Bringaud, F., Nolan, D. P., & Figueiredo, L. M. (2017). Metabolic reprogramming during the *Trypanosoma brucei* life cycle. *F1000Research*, 6.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., & Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*, *3*, 34.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., & Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, *18* Suppl 2, S231-240.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*, e21800.

Supper, J., Spangenberg, L., Planatscher, H., Dräger, A., Schröder, A., & Zell, A. (2009). BowTieBuilder: Modeling signal transduction pathways. *BMC Systems Biology*, *3*, 67.

Sutherland, C. S., & Tediosi, F. (2019). Is the elimination of 'sleeping sickness' affordable? Who will pay the price? Assessing the financial burden for the elimination of human African trypanosomiasis *Trypanosoma brucei gambiense* in sub-Saharan Africa. *BMJ Global Health*, *4*, e001173.

Telleria, E. L., Benoit, J. B., Zhao, X., Savage, A. F., Regmi, S., e Silva, T. L. A., … Aksoy, S. (2014). Insights into the Trypanosome-Host interactions revealed through transcriptomic analysis of rarasitized tsetse fly salivary glands. *PLoS Neglected Tropical Diseases*, *8*, e2649.

Tetley, L., Turner, C. M., Barry, J. D., Crowe, J. S., & Vickerman, K. (1987). Onset of expression of the variant surface glycoproteins of *Trypanosoma brucei* in the tsetse fly studied using immunoelectron microscopy. *Journal of Cell Science*, *87* (Pt 2): 363–372.

Tuncbag, N., Gursoy, A., Nussinov, R., & Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols*, *6*, 1341–1354.

Uilenberg, G., & Boyt, W. (1998). *A field guide for the diagnosis, treatment and prevention of African animal trypanosomosis*. Food and Agriculture Organization.

Urwyler, S., Studer, E., Renggli, C. K., & Roditi, I. (2007). A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of *Trypanosoma brucei*. *Molecular Microbiology*, *63*, 218–228.

Vale, G. A. (1993). Development of baits for tsetse flies (*Diptera: Glossinidae*) in Zimbabwe. *Journal of Medical Entomology*, *30*, 831–842.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, *19*, 575–592.

van Weelden, S. W. H., van Hellemond, J. J., Opperdoes, F. R., & Tielens, A. G. M. (2005). New functions for parts of the Krebs cycle in procyclic *Trypanosoma brucei*, a cycle not operating as a cycle. *The Journal of Biological Chemistry*, *280*, 12451–12460.

Vassella, E., Acosta-Serrano, A., Studer, E., Lee, S. H., Englund, P. T., & Roditi, I. (2001). Multiple procyclin isoforms are expressed differentially during the development of insect forms of *Trypanosoma brucei*. *Journal of Molecular Biology*, *312*,597–607.

Vickerman, K. (1985). Developmental cycles and biology of pathogenic trypanosomes. *British Medical Bulletin*, *41*, 105–114.

Vreysen, M. J. (2001). Principles of area-wide integrated tsetse fly control using the sterile insect technique. *Medecine Tropicale: Revue Du Corps De Sante Colonial*, *61*, 397–411.

Vreysen, M. J. B., Saleh, K., Mramba, F., Parker, A., Feldmann, U., Dyck, V. A., … Bouyer, J. (2014). Sterile insects to enhance agricultural development: The case of sustainable tsetse eradication on Unguja Island, Zanzibar, using an area-wide integrated pest management approach. *PLOS Neglected Tropical Diseases*, *8*, e2857.

Vreysen, M. J. B., Seck, M. T., Sall, B., & Bouyer, J. (2013). Tsetse flies: Their biology and control using area-wide integrated pest management approaches. *Journal of Invertebrate Pathology*, 112 Suppl: S15-25.

Vreysen, M. J., Khamis, I. S., & Van der Vloedt, A. M. (1996). Evaluation of sticky panels to monitor populations of *Glossina austeni* (*Diptera: Glossinidae*) on Unguja island of Zanzibar. *Bulletin of Entomological Research*, *86*, 289–296.

Weiss, B., & Aksoy, S. (2011). Microbiome influences on insect host vector competence. *Trends in Parasitology*, *27*, 514–522.

Weiss, B. L., Mouchotte, R., Rio, R. V. M., Wu, Y., Wu, Z., Heddi, A., & Aksoy, S. (2006). Interspecific transfer of bacterial endosymbionts between tsetse fly species: Infection establishment and effect on host fitness. *Applied and Environmental Microbiology*, *72*, 7013–7021.

Weiss, B. L., Wang, J., Maltz, M. A., Wu, Y., & Aksoy, S. (2013). Trypanosome infection establishment in the tsetse fly gut is influenced by microbiome-regulated host immune barriers. *PLoS Pathogens*, 9.

Welburn, S. C., & Maudlin, I. (1997). Control of *Trypanosoma brucei brucei* infections in tsetse, *Glossina morsitans*. *Medical and Veterinary Entomology*, *11*, 286–289.

WHO. (1998). Control and surveillance of African trypanosomiasis: Report of a WHO expert committee. Retrieved March 17, 2019, from https://apps.who.int/iris/handle/10665/42087

WHO. (2002). WHO | WHO programme to eliminate sleeping sickness: Building a global alliance. Retrieved March 17, 2019, from http://www.who.int/trypanosomiasis_african/resources/who_cds_csr_eph_2002.13/en/

WHO. (2012). *Accelerating work to overcome the global impact of neglected tropical diseases: A roadmap for implementation : executive summary*. Retrieved from https://apps.who.int/iris/handle/10665/70809

WHO. (2013). *Report of a WHO meeting on elimination of African trypanosomiasis (Trypanosoma brucei gambiense)*. Geneva. Retrieved from https://www.who.int/neglected_diseases/resources/who_htm_ntd_idm_2013.4/en/

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics*, *6*, 227.

Wu, G., & Ji, H. (2013). ChIPXpress: Using publicly available gene expression data to improve ChIP-seq and ChIP-chip target gene ranking. *BMC Bioinformatics*, *14*, 188.

Wurst, M., Seliger, B., Jha, B. A., Klein, C., Queiroz, R., & Clayton, C. (2012). Expression of the RNA recognition motif protein RBP10 promotes a bloodstream-form transcript pattern in *Trypanosoma brucei*. *Molecular Microbiology*, *83*, 1048–1063.

Xenarios, I., Salwínski, Ł., Duan, X. J., Higney, P., Kim, S.-M., & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, *30*, 303–305.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biology*, *11*, R14.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article 17.

Zhou, Q., Dong, G., & Li, Z. (2018a). Flagellum inheritance in *Trypanosoma brucei* requires a kinetoplastid-specific protein phosphatase. *The Journal of Biological Chemistry*, *293*, 8508–8520.

Zhou, Q., Gu, J., Lun, Z.-R., Ayala, F. J., & Li, Z. (2016a). Two distinct cytokinesis pathways drive trypanosome cell division initiation from opposite cell ends. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 3287–3292.

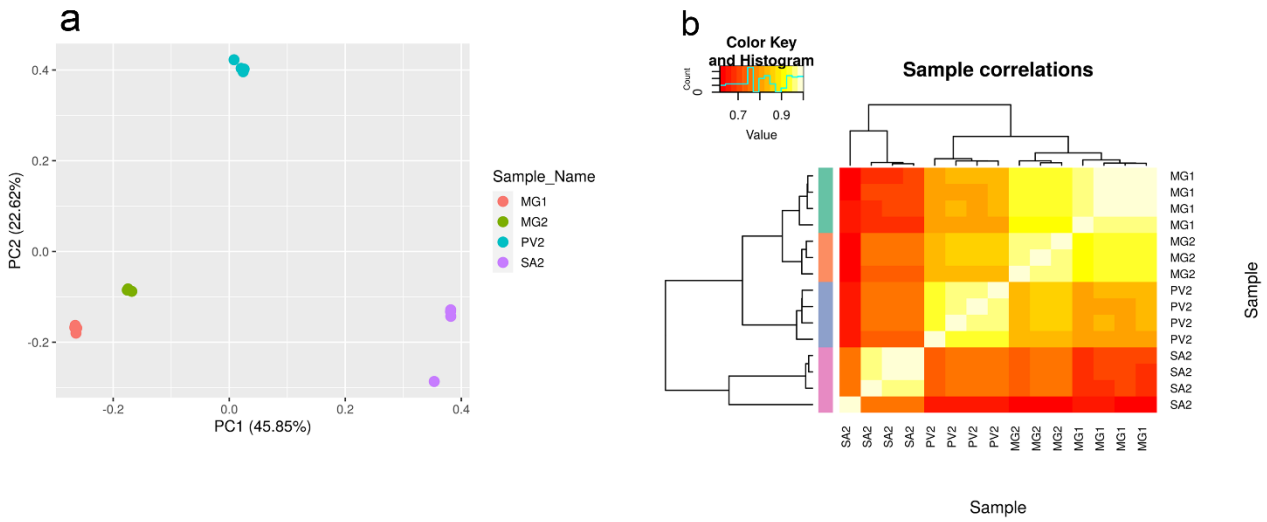
Zhou, Q., Hu, H., & Li, Z. (2014). New insights into the molecular mechanisms of mitosis and cytokinesis in trypanosomes. *International Review of Cell and Molecular Biology*, *308*, 127–166.

Zhou, Q., Hu, H., & Li, Z. (2016b). An EF-hand-containing protein in *Trypanosoma brucei* regulates cytokinesis initiation by maintaining the stability of the Cytokinesis Initiation Factor CIF1. *The Journal of Biological Chemistry*, *291*, 14395–14409.

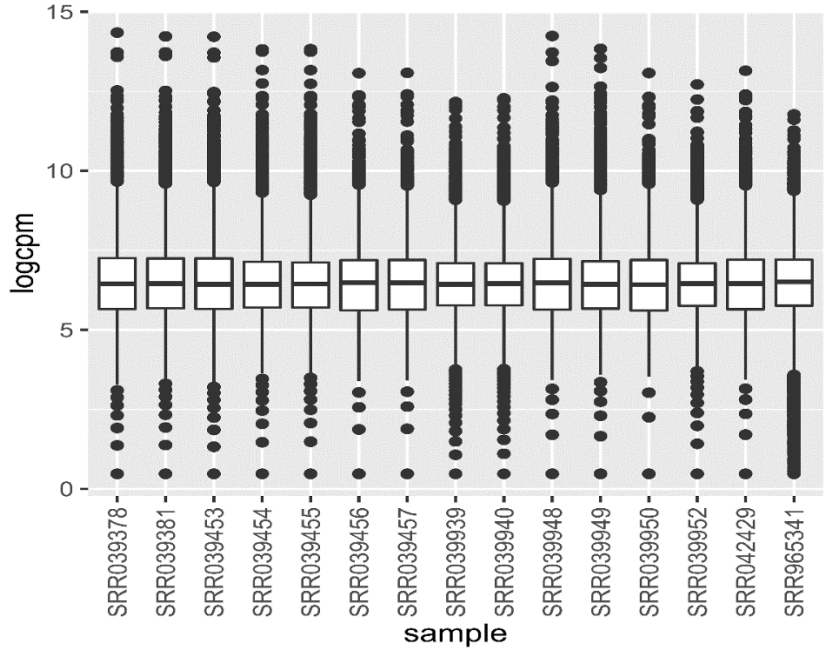
Zhou, Q., Lee, K. J., Kurasawa, Y., Hu, H., An, T., & Li, Z. (2018b). Faithful chromosome segregation in *Trypanosoma brucei* requires a cohort of divergent spindle-associated proteins with distinct functions. *Nucleic Acids Research*, *46*, 8216–8231.

Zubiaga, A. M., Belasco, J. G., & Greenberg, M. E. (1995). The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation. *Molecular and Cellular Biology*, *15*, 2219–2230.

APPENDICES



Appendix I: **Principal component analysis and sample correlation heatmap using hierarchical clustering before adjusting for batch effects. a) Principal component analysis plot. b) Sample correlation heatmap plot.**



Appendix II: **Assessment of per-gene read counts distribution per sample using boxplot before normalization and batch effect adjustment.**