ABSTRACT

Title of Document:	Simultaneous transcriptome profiling of *Trypanosoma cruzi* parasites and their human host cells. Yuan Li, Doctor of Philosophy, 2014
Directed By:	Dr. Najib M. El-Sayed, Associate Professor Department of Cell Biology and Molecular Genetics

The genome of the kinetoplastid parasite *Trypanosoma cruzi*, causative agent of Chagas disease, was published nine years ago, yet a systematic and comprehensive analysis of the transcriptomes of the parasite and the human host has not been conducted. The parasite responds rapidly to transmission between arthropod vectors and mammalian hosts by undergoing complex cellular differentiation processes that are not well understood. In this study, we generated the first transcriptome map for both *T. cruzi* and infected human host cells across the infection cycle including time points of 4, 6, 12, 24, 48 and 72 hours post invasion with the next generation RNA sequencing technology (RNA-Seq). We also captured the transcriptome of the parasite in its bloodstream form (trypomastigote) and its replicative form inside insect vector (epimastigote). We successfully mapped transcribed regions for the pathogen at single nucleotide

resolution on a genomic scale and characterized the RNA processing (*trans*-splicing and polyadenylation) events across its various developmental stages. Here we report the prevalent heterogeneity of RNA processing sites across the genome. We also note the preference of different primary sites in various developmental stages presenting as a potential and interesting approach of posttranscriptional regulation, which may hypothetically contribute to the survival of the parasite across different environments. Our work has significantly enhanced the current genome annotation of *T. cruzi*. In addition, using the *T. cruzi* and human genome sequence as reference, we explored these data with informatics tools to identify genes with significant regulation and successfully profiled gene expressions from both species simultaneously. We examined the subsets of differentially expressed genes both in the parasite and the host cell over the course of the infection to understand the mechanisms of invasion and intracellular survival strategy as well as host-pathogen interactions. *T. cruzi* genes that were significantly regulated during the infection process might present as new targets for drug development, whereas human genes that were significantly regulated might signal the immunoinflammatory response triggered by the manipulation of the parasite. Furthermore, we investigated the gene expression patterns of *T. cruzi* across its different developmental stages, clustered gene with similar patterns, and identified possible sequence motifs in coexpressed gene clusters.

SIMULTANEOUS TRANSCRIPTOME PROFILING OF *TRYPANOSOMA CRUZI* PARASITES AND THEIR HUMAN HOST CELLS

By

Yuan Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:

Associate Professor Najib M. El-Sayed, Chair

Professor David M. Mosser

Associate Professor Stephen Mount

Assistant Professor Hector Corrado Bravo

Associate Professor Carlos A. Machado, Dean's Representative

Dedication

To my parents, Haicong Li and Yiling Yang,

who have been there for me from day one.

Thank you for all of the love, support, encouragement and dedication.

Acknowledgements

I would like to express my deep appreciation and gratitude to my advisor, Dr. Najib M. El-Sayed, for the patient guidance and mentorship he provided to me, all the way from the start of my PhD. study, through the completion of this degree. Dr. El-Sayed's intellectual heft is matched only by his genuinely good nature, and his dedication to science, and I am truly fortunate to have had the opportunity to work with him.

I would also like to thank my committee members, Drs. David Mosser, Stephen Mount, Hector Corrada Bravo, and Carlos A. MacHado for their friendly guidance, thought-provoking suggestions and the general collegiality that each of them offered to me over the years. In a similar vein, I'd like to recognize Dr. Barbara Burleigh for the contributions that she made to my intellectual growth during my years of study at the University of Maryland. I would like to thank everyone in my lab – those has have moved on, those in the quagmire, and those just beginning: Drs. Gustavo Cerqueira, Michael Waisberg, Ramzi Temanni, Wanderson Da Rocha, Ashton Trey Belew, Jungmin Choi, Cecilia Fernandes as well as April Hussey, Ginger Houston-Ludlam, Laura Dillon, Keith Hughitt, Rondon Neto and Pablo Smircich - for their insightful feedback on various aspects of my projects and for their support and friendship. I also want to thank my boyfriend Wai Lim Ku for his help along the way.

Finally, I want to express my heart-felt gratitude to my parents Dr. Haicong Li and Yiling Yang, who has been a constant source of love, concern, support and strength for me all these years.

Table of Contents

ABSTRACT i

Dedication ii

Acknowledgements iii

Table of Contents iv

List of Tables vi

List of Figures vii

List of Abbreviations xi

Chapter 1 Introduction 1

1.1 Chagas disease 1

1.1.1 *Trypanosoma cruzi* 1

1.1.2 The life cycle of *T. cruzi* 1

1.1.3 Chagas disease 4

1.2 The Genome of *T. cruzi* 15

1.3 Genetic diversity of *T. cruzi* 20

1.4 Gene expression in trypanosomes 27

1.5 Comparative genomics of trypanosomes 30

1.6 RNA sequencing technology 34

1.7 Summary of dissertation work 39

1.8 Significance 43

Chapter 2: RNA Processing Events in *T. cruzi* Developmental Stages Pre- and Post-Infection of Human Host Cells 44

2.1 Objective of Study 44

2.2 Materials and Methods 44
2.2.1 Materials 44
2.2.2 Methods 45
2.3 Results 51
2.4 Conclusion and discussion 88
Chapter 3: Simultaneous Interrogation of the Transcriptomes of the Human Pathogen *Trypanosoma cruzi* and its Infected Host Cell 92
2.1 Objective of Study 92
2.2 Materials and Methods 92
2.2.1 Materials 92
2.2.2 Methods 93
2.4 Results 99
2.5 Conclusion and discussion 146
Chapter 4: Conclusion remarks and future perspective 150
Acknowledgement 160
Appendices 161
Appendix 1 161
Appendix 2 163
Appendix 3 166
Appendix 4: Supplementary table index 169
Bibliography 176

List of Tables

Table 1 Classifications of *T. cruzi* strains. 21

Table 2 Summary of mapping of reads and reads containing RNA processing features. 52

Table 3 Characterization of different *T. cruzi* gene structure components. 62

Table 4 GC content of different gene structure components in *T. cruzi*. 67

Table 5 Usage frequency of nucleotide proceeding splicing acceptor (AG) site. 76

Table 6 Summary of experiment design and treatments. 161

Table 7 Summary of total number of reads mapped to the reference genomes. 163

Table 8 Summary of mapping statistics to individual genomes 166

List of Figures

Figure 1 Life cycle of *T. cruzi* 2

Figure 2 Estimated number of infections with *T. cruzi* outside endemic countries. 5

Figure 3 Chromosome assembly of *T. cruzi*. 18

Figure 4 Two hybridization events define the population structure of *T. cruzi*. 23

Figure 5 Comparison of gene content for some of the largest gene families between Sylvio X10 and CL Brener strain of *T. cruzi*. 24

Figure 6 Gene expression in trypanosomatids. 29

Figure 7 Comparative genomics of Tritryps. 33

Figure 8 Overview of a typical RNA-Seq experiment. 35

Figure 9 Outline of an end-enriched RNA-Seq experiment for trypanosomes. 37

Figure 10 A general overview of the pipeline of RNA-Seq analysis for detecting differential expressed genes. 38

Figure 11 Length and position distribution of various gene structure components in *T. cruzi*. 56

Figure 12 Distribution of the lengths of 5′ UTRs in transcript expressed in different developmental stages of *T. cruzi*. 58

Figure 13 Distribution of the lengths of 3′ UTRs in transcripts expressed in different developmental stages of *T. cruzi*. 59

Figure 14 Correlations between CDS length and either UTR length or between corresponding UTR lengths. 61
Figure 15 Comparison of gene structure components between *T. cruzi* and *T. brucei*. 65
Figure 16 Distribution of UTR lengths for *T. cruzi* and *T. brucei* genes in 3-way COGs. 66
Figure 17 An example of alternative *trans*-splicing events. 70
Figure 18 Sequence composition at RNA processing sites and distance between primary and alternative RNA processing sites. 71
Figure 19 Analysis of sequence composition at alternative spliced leader acceptor sites. 73
Figure 20 Analysis of sequence composition near the minor polyadenylation sites. 73
Figure 21 Distribution of distances between primary and alternative polyadenylation sites in transcripts from different developmental stages of *T. cruzi*. 79
Figure 22 Alternative splicing profiles of *T. cruzi* trypomastigote, amastigote and epimastigote stages. 81
Figure 23 Analysis of the usage of primary and secondary trans-splicing sites in epimastigotes and amastigotes (72 hpi) in context of expression across developmental stages. 83
Figure 24 Examples of alternative *trans*-splicing events across developmental stages. 86

Figure 25 Examples of alternative *trans*-splicing events across different developmental stages. 87
Figure 26 Simultaneous interrogation of the host and pathogen transcriptomes - Experimental design. 100
Figure 27 Overview of the pipeline for differential expression analysis. 102
Figure 28 Examples of quality scores across all bases from sequenced samples by FastQC. 104
Figure 29 An example of nucleotide composition from sequenced reads. 105
Figure 30 Number of non-zero genes detected in samples with various sequencing depth. 107
Figure 31 Number of mapped reads for *T. cruzi* and human samples. 109
Figure 32 Heatmap of Pearson correlation between samples. 110
Figure 33 Mean-variance trend of *T. cruzi* and human samples. 112
Figure 34 Distribution of global gene expression levels in *T. cruzi* and human samples. 114
Figure 35 Standardized median Pearson correlation between *T. cruzi* and human samples. 116
Figure 36 Principal component analysis of global transcriptome profiles in *T. cruzi* and human host cells at various stages of the infection. 117

Figure 37 Hierarchical clustering of *T. cruzi* samples. 119
Figure 38 Hierarchical clustering of human samples. 120
Figure 39 Pairwise Pearson correlation between *T. cruzi* samples. 121
Figure 40 Pairwise Pearson correlation between human samples. 122
Figure 41 Differentially expressed (DE) genes from human at various stages of the infection. 125
Figure 42 Differentially expressed (DE) genes from *T. cruzi* at various stages of the infection. 126
Figure 43 Heatmap of the top 200 *T. cruzi* genes significantly regulated pre- and post- replication at the intracellular stages. 128
Figure 44 Heatmap of the top 200 human genes significantly regulated pre- and post- intracellular parasite replication. 129
Figure 45 K-means clustering of *T. cruzi* transcriptome based on the dynamic progression of gene profiles across different developmental stages..... 135
Figure 46 K-means clustering of human transcriptome based on the dynamic progression of gene profiles across different developmental stages..... 136
Figure 47 Overrepresented motifs detected in the 3' UTR of cluster 1 in *T. cruzi.* 142

List of Abbreviations

A	adenine
Ama	amastigote
ARE	AU-rich elements
ATP	adenosine triphosphate
AU	adenylate/Uridylate
BAC	bacterial artificial chromosome
BH_4	tetrahydrobiopterin
BLAST	basic local alignment search tool
bp	base pair
Bz	benznidazole
C	cytosine
cAMP	cyclic adenosine monophosphate
cDNA	complementary DNA
CDS	coding DNA Sequence
ChIP-Seq	chromatin immunoprecipitation sequencing
CO_2	carbon dioxide
COG	cluster of orthologous genes

CS	cycling sequence
CSBP	cycling sequence binding protein
DE	differential expression
DGF	dispersed gene family
DMEM	Dulbecco’s modified Eagle medium
DNA	deoxyribonucleic acid
DTU	discrete typing units
ECG	electrocardiography
ELISA	enzyme-linked immune sorbent assay
Epi	epimastigote
EST	expressed sequence tag
FBS	fetal bovine serum
G	guanine
GCH1	GTP-cyclohydrolase 1
GO	gene ontology
GP 63	glycoprotein 63
GPI	glycophosphatidylinositol
gRNA	guide RNA

GSEA	gene set enrichment analysis
GTP	guanosine triphosphate
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic
HFF	human foreskin fibroblasts
HPGL	Host-Pathogen Genomics Laboratory
hpi	hour post infection
IFA	immunofluorescence assay
IFN	interferon
IHA	indirect hemagglutination assay
kDNA	kinetoplastid DNA
KEGG	Kyoto Encyclopedia of Genes and Genomes
MASP	mucin-associated surface protein
MLGP	mucin-like glycoprotein
mtDNA	mitochondrial DNA
NCBI	National Center for Biotechnology Information
Nfx	nifurtimox
nt	nucleotide
ORF	open reading frame

PBS	phosphate-buffered saline
PCA	principal component analysis
PCR	polymerase chain reaction
PKA	cAMP-dependent protein kinase
pol II	RNA polymerase II
PTU	polycistronic unit
PWM	position weight matrix
qPCR	real-time polymerase chain reaction
RBP	RNA binding protein
RHS	recombination hot spot
Ribo-Seq	ribosome profiling
RNA	ribonucleic acid
RNA-Seq	RNA Sequencing
siRNA	small interfering RNA
SL	splice leader
SMC	smooth muscle cell
SSR	strand switching region
T	thymine

Tb	*Trypanosoma brucei*
Tc	*Trypanosoma cruzi*
TCT	tissue culture-derived trypomastigote
tRNA	transfer RNA
Trypo	trypomastigote
TS	*trans*-splicing
TSS	transcripton start sites
U	uradine
UTR	untranslated region
WHO	World Health Organization

Chapter 1 Introduction

1.1 Chagas disease

1.1.1 *Trypanosoma cruzi*

Trypanosoma cruzi is an intracellular protozoan parasite and the etiological agent of Chagas disease, or American trypanosomiasis. It was first discovered in 1909 by Brazilian physician Dr. Carlos Chagas. He named the pathogen after his mentor Oswaldo Cruz. Chagas' original report described in great detail both the cycle of transmission and the acute clinical manifestation of the first human case. Findings from paleoparasitology studies have recovered *T. cruzi* DNA from human mummies and showed that the presence of Chagas disease can be dated as early as 9000 years ago (Aufderheide, Salo et al. 2004). Notably, it has been hypothesized that Charles Darwin might have suffered from Chagas disease as the result of the bite of triatomine during his expedition to South America in 1835, suggested by his vivid description of contact with the kissing bug and by some of his symptoms in later life (Bernstein 1984).

1.1.2 The life cycle of *T. cruzi*

The *T. cruzi* life cycle is complex and consists of several distinct life stages, morphological states and hosts with three main developmental forms: epimastigote, trypomastigote and amastigote (Rassi, Rassi et al. 2010) (**Figure 1**). Epimastigotes proliferate in the midgut of Triatomine insects that feed on infected mammals. In the hindgut, they differentiate into infective metacyclic

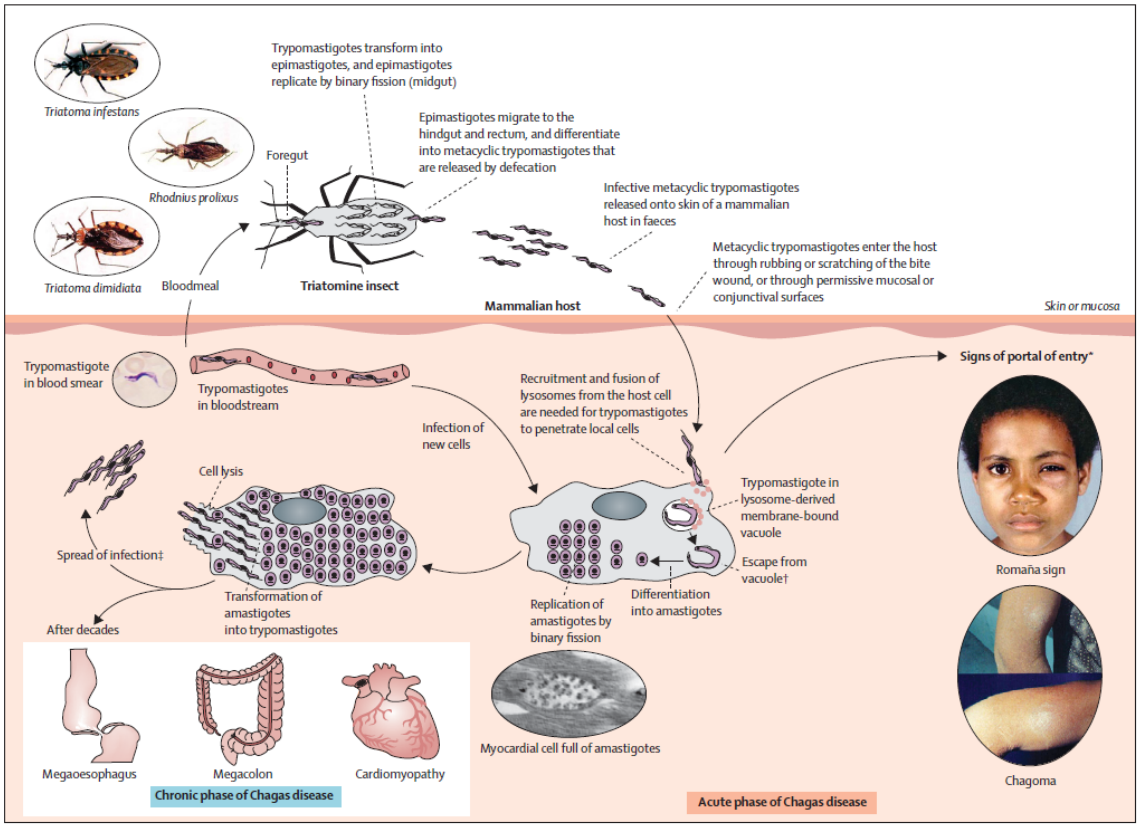

Figure 1 Life cycle of *T. cruzi*

Cell invasion by infective trypomastigotes after host scratch near the bite site on the skin is followed by transformation into replicative amastigotes and multiple binary divisions in the cell cytoplasm. Upon the rupture of the host cell, released trypomastigotes can either start another infection cycle or get picked up by the insect vector (Rassi, Rassi et al. 2010) .

trypomastigotes that are released by defecation. These trypomastigotes enter mammalian cells and transform into proliferative amastigote forms. When the local cells are swollen with amastigotes, they transform again into non-dividing and infective trypomastigotes. Trypomastigotes lyse infected cells, invade adjacent tissues, and disseminate via the lymphatics and bloodstream to distant sites, mainly muscle cells and ganglion cells, where they undergo further cycles of intracellular multiplication (Brener 1971). To achieve successful completion of the life cycle, the parasite must rapidly adapt to different environments by regulating its gene expression.

1.1.3 Chagas disease

1.1.3.1 Epidemiology

Chagas disease is ranked as one of the most important pathogens throughout Central America and South America. According to countrywide survey done in the 1980s, more than 100 million people were exposed to the risk of infection and 17.4 million were infected in 18 countries resulted in more than 50000 death per year (WHO 2002; Rassi, Rassi et al. 2010; Hotez, Dumonteil et al. 2013). The disease was correlated with poor living conditions and usually occurred in rural areas, in which vector borne transmission to man occurred. In the past 20 years, great efforts, including improved vector control programs and compulsory blood-bank screening, have been made to control new cases of infection and reduce disease prevalence (Moncayo 2003). Current estimates indicate that the number of infection cases decreased to 7.7 million, with the number of death per year dropped to 12500 per year (Moncayo 2003). However, the recent influx of immigrants from countries endemic for disease has suggested that Chagas disease is becoming an important health issue in the USA and Canada and in many parts of Europe and the western Pacific, where an increasing number of infected cases has been reported (Schmunis 1991; Martinez-Perez, Norman et al. 2014) (**Figure 2**). In USA, an estimated 300,167 individuals are infected with *T. cruzi* (Bern and Montgomery 2009).

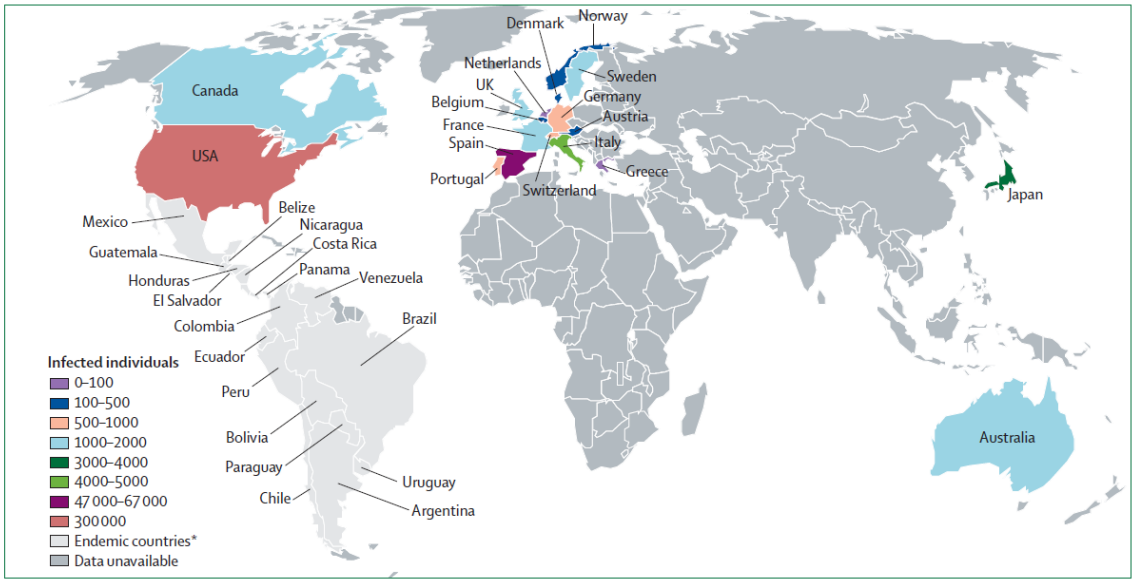

Figure 2 Estimated number of infections with *T. cruzi* outside endemic countries.

Infected population in non-endemic countries is indicated in different colors and other infected populations in endemic regions are colored in grey. The recent influx of immigrants from countries endemic for disease has implied that Chagas disease is becoming an important health issue in the USA and Canada and in many parts of Europe and the western Pacific, where an increasing number of infected cases has been reported (Rassi, Rassi et al. 2010).

1.1.3.2 Pathogenesis

In the acute phase of Chagas disease, organ and tissue damage is caused by the invasion of the parasite and the immunoinflammatory response triggered by thepresence of the parasite (Tarleton and Zhang 1999; Soares, Pontes-De-Carvalho et al. 2001; Marin-Neto, Cunha-Neto et al. 2007; Munoz-Saravia, Haberland et al. 2012). Several studies with experimental models of *T. cruzi* infection have reported that a strong T-helper-1 immune response with both CD4 and CD8 cells and production of some specific cytokines, including interferon γ, tumor necrosis factor α, and interleukin 12 , plays key roles in the control of parasitism (Abrahamsohn and Coffman 1996; Aliberti, Cardoso et al. 1996; Silva, Aliberti et al. 1998; Martins, Vieira et al. 1999). The pathogenesis of chronic Chagas disease is controversial. Two main hypotheses have been developed to explain the main cause of tissue damage: one focuses on the direct parasite factors, and the other is based on the immunopathology or even the autoimmune mechanisms (Soares, Pontes-De-Carvalho et al. 2001; Marin-Neto, Cunha-Neto et al. 2007). However, a growing consensus indicates that the presence of pathogen is required for the development of the disease (Tarleton and Zhang 1999; Tarleton 2003). The emergence of more sensitive diagnostic tests, such as polymerase chain reactions (PCRs), confirmed the persistence of the parasite deep inside myocardium (Britto, Cardoso et al. 1999; Brasil, De Castro et al. 2010). The balance between immune-mediated parasite containment and inflammation of the host tissue is important in the progression of the disease: if the immunological response is ineffective, or generates tissue damage, both

parasite burden and immune-mediated inflammation increased. On the contrary, a well-executed immune response, in which parasite load is lowered and inflammatory consequences are kept to a minimum, results in a better outcome (Tarleton 2003).

The severe heart symptoms shown at late stage of Chagas disease is a result of slowly progressive and incessant myocarditis, which leads to the impairment of contractile function and dilatation of all four chambers. Histological examination shows widespread damage of myocardial cells, diffuse fibrosis, oedema, mononuclear cell infiltration in the myocardium and scarring of the conduction system (Andrade 1983; Rassi, Rassi et al. 2010). The progressive destruction of cardiac fibers and the intense fibrosis from replacement of dead myocytes can cause heart failure and ventricular arrhythmias (Rassi, Rassi et al. 2000). Chronic Chagas gastrointestinal disease is caused by the destruction of intramural autonomic ganglia, which can have impact on esophagus, colon, or both (Rassi, Rassi et al. 2000; Rassi, Rassi et al. 2010).

1.1.3.3 Progression and Clinical manifestations

1.1.3.3.1 Acute phase

The progression of Chagas disease has been characterized into three distinct developmental stages: acute phase, indeterminate phase and chronic phase (Dias, Laranja et al. 1956; WHO 2002). The acute phase of infection with *T. cruzi*

lasts for 4-8 weeks. Patients at acute stage are usually asymptomatic and most of them are not aware of the infection (Rassi, Rassi et al. 2010). When symptoms occur they include: malaise, fever, swelling of lymph nodes and tissues. In the case of vector-borne transmission, the signs of portal of entry of *T .cruzi* include subcutaneous oedema through the skin or via the ocular mucous membranes, which is referred to as Romaña's sign. Death occurs in less than 5-10% of symptomatic cases at this stage, most of which result from myocarditis or meningoencephalitis or both(Rassi, Rassi et al. 2010). At acute stage, high blood parasitaemia is detected and every nucleated cell in the host is a target of infection. Antiparasitic drugs, such as benznidazole and nifurtimox, can usually cure acute infection and prevent chronic manifestations (Voigt, Bock et al. 1972; Masana, de Toranzo et al. 1983; Pinto, Ferreira et al. 2009). Even without the intervention of trypanocidal drugs, the acute infection can spontaneously resolves in about 90% of infected individuals.

1.1.3.3.2 Indeterminate phase

Approximately 60-70% of patients with acute Chagas disease will never develop clinically apparent disease and remain in the indeterminate phase for life. This phase is characterized by positive serology, low level of parasitaemia, normal electrocardiogram (ECG) and no abnormality of organs (Macedo 1999). Sanchez-Gullen's work has shown that the concentration of superoxide dismutase (SOD) were significantly elevated in patients in the indeterminate stage compared to chronic stage indicating the role of host's anti-oxidant defense

response in the inhibition of inflammation after infection (Perez-Fuentes, Torres-Rasgado et al. 2008). A recent study has reported that cardiac mitochondria are involved in the genesis and progression to chronic chagasic cardiopathy despite the 'silent' state of the phase when the host- pathogen equilibrium is altered (Baez, Lo Presti et al. 2013).

1.1.3.3.3 Chronic phase

About 30-40% of patients with infection of *T. cruzi* will develop a determinate form of chronic Chagas disease (Lescure, Le Loup et al. 2010; Rassi, Rassi et al. 2010). Typical clinical manifestations are associated with heart, colon, esophagus, or a combination and can be grouped into three categories: cardiac, digestive and cardiodigestive (Rassi, Rassi et al. 2000; Tarleton 2003). The cardiac form is the most serious manifestation of the chronic phase and about 20-30% of the chronic patients will develop into this form. Common symptoms include enlargement of heart, bradyarrhythmias and tachyarrhythmias, abnormalities of the conduction system, apical aneurysms, and thromboembolism (Marin Neto, Simoes et al. 1999; Ribeiro, Nunes et al. 2012). Cardiac failure is often a late manifestation and chagasic cardiomyopathy is the most common cause of death during the chronic phase of Chagas in South America. Accounting for nearly 70% of all deaths, sudden death is the main reason of death, followed by refractory heat failure and thromboembolism (Rassi, Rassi et al. 2001). Patients who were asymptomatic may suffer from sudden death as well. The digestive manifestation of chronic Chagas disease is often

referred to as “megasyndromes”, in the form of dilation of the gastrointestinal tract (megaesophagus and megacolon), combined with weight loss and swallowing difficulties. Patients with megacolon can also present chronic constipation, abdominal pain, volvulus, obstructions, and intestinal perforations(Garcia, Aranha et al. 2003). Amastigotic cysts can damage ganglionic neurons in the digestive system thus reducing the ability of smooth muscle to dilate and contract and causing a loss of muscle control. The parasympathetic denervation is another characteristic of chronic Chagas disease (Tostes, Bertulucci Rocha-Rodrigues et al. 2005).

1.1.3.4 Prevention, diagnostics and treatment

1.1.3.4.1 Prevention

Currently prevention of Chagas disease only relies on vector control and prevention of transmission from non-vectorial mechanisms (Rassi, Rassi et al. 2010; Quijano-Hernandez and Dumonteil 2011). Vector eradication programs, including chemical control of the vector by conventional spray or by community participation during the surveillance phase, have been launched in a number of Central and South American countries such as Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Honduras, Nicaragua, Panama, Peru, Uruguay and Venezuela (Schofield and Dujardin 1997). Combined with compulsory screening of blood donors, continuous health education, improved housing conditions and epidemiological surveillance, a

considerably reduced disease incidence and prevalence has been observed in most Latin America countries participated in the program (Rojas de Arias, Ferro et al. 1999; Moncayo 2003; Gurtler, Kitron et al. 2007). Nonetheless, control of the disease still requires adequate attention and efforts. Several challenges remains as a number of insect vectors have shown difficulty to control and the possibility of resurgence of vector-borne transmission in regions once successfully controlled still exists. The effect of insecticides can last no more than 1 year, so control interventions have to be repeated at a regular basis (Oliveira Filho 1999; Vazquez-Prokopec, Spillmann et al. 2009).

Currently there is no vaccine available for Chagas disease, although efforts have been made in the development of *T. cruzi* vaccines (Quijano-Hernandez and Dumonteil 2011). A very limited variety of *T. cruzi* antigens have been evaluated so far, and some studies have focused on the identification of novel antigens. A growing consensus has pointed out that a successful vaccine will need to induce a strong immune response, which requires the stimulation of Th1 immune profile with the activation of $CD8^+$ cytotoxic T cells in order to effectively control *T. cruzi* parasites (Zapata-Estrella, Hummel-Newell et al. 2006). Recent studies have shown that several vaccines types, including recombinant proteins, DNA and viral vectors, as well as heterologous prime-boost combinations, can be immunogenic and protective in mouse models, which provides first-hand data on the feasibility of a preventive and/or therapeutic vaccine to control the parasite infection (Araujo, de Alencar et al. 2005; Crampton and Vanniasinkam 2007;

Lorena, Lorena et al. 2010). However, many challenges remain to be addressed, including technical, ethical and logistical issues. A major concern with Chagas disease has been the damage caused by vaccination, which may stimulate autoimmunity and induce exacerbation rather than protection (Kierszenbaum 1999; Tarleton and Zhang 1999). This has considerably slowed down vaccine research. Another major issue is the design of clinical trials. The course of Chagas disease can be 20-30 years. The trials would need to take place for at least 20 years to observe any significant effect (Camargo 2009). In addition, given the low incidence of disease in some countries, randomized controlled trials may need a large number of volunteers (Quijano-Hernandez and Dumonteil 2011). The lack of widely accepted biomarkers to evaluate vaccine efficacy can be another problem in the development of *T. cruzi* vaccines (Ndao, Spithill et al. 2010).

1.1.3.4.2 Diagnostics and treatment

At the early stage of infection with *T. cruzi*, trypomastigotes may be detected in the bloodstream by direct microscopic observation in blood or by various culture techniques (WHO 2002; Gomes, Lorena et al. 2009). Unfortunately, acute Chagas disease are often undetected because most of the patients are asymptomatic or only have mild and unspecific symptoms. Once the immune response initiates, parasitaemia is scarce and detection of parasites becomes very difficult (Dubner, Schapachnik et al. 2008). At the chronic phase, the presence of IgG antibodies against parasite antigens needs to be confirmed by at

least two different serological techniques based on different principles and detecting different antigens to conclude diagnosis, according to the recommendation from World Health Organization (WHO) (WHO 2002; 2005; Rassi, Rassi et al. 2010). In the case of ambiguous and inconsistent results, a third technique should be used. Conventional serological methods include primarily immunofluorescence assays (IFA), enzyme-linked immune sorbent assays (ELISA) and indirect hemagglutination assays (IHA) (Afonso, Ebell et al. 2012). A recent review of eighteen studies and 61 assays of the diagnostic serological assays of Chagas disease pointed out that the sensitivity and specificity of the serological diagnostic assays appear less accurate than previously thought, only 90% and 98% respectively (Afonso, Ebell et al. 2012).

The aim of treatment is to eradicate the parasite and alleviate the symptoms of the disease (Rassi, Rassi et al. 2010). Since the 1960s, the only drugs with proven efficacy against Chagas diseases are, nifurtimox (Nfx) and benznidazole (Bz). Nfx is a nitrofurane and Bz is a nitroimidazole compound. Nfx targets the nitroreductase in the parasite, which generates a nitroanion leading to the formation of reactive nitrogen species that *T. cruzi* fails to remove (Wegner and Rohwedder 1972). Bz can create a covalent bond with various parasite components and inhibits the reduction-oxidation processes in *T. cruzi* (Masana, de Toranzo et al. 1983; Perez-Molina, Perez-Ayala et al. 2009). The use of these drugs to treat the acute phase of the disease is widely accepted, but not for the chronic phase (Rassi, Rassi et al. 2010). Both drugs can display serious side

effects, frequently forcing the physician to stop treatment. The most common adverse effects of Nfx are: anorexia, loss of weight, psychic alterations, excitability, sleepiness, digestive manifestations such as nausea and diarrhea. In the case of Bz, skin manifestations are the most frequent, including hypersensitivity, dermatitis, and generalized oedema, with depression of bone marrow, thrombocytopenic purpura and agranulocytosis being the more severe manifestations. Toxicity tests with Nfx indicated neurotoxicity, testicular damage, ovarian toxicity, and deleterious effects in adrenal, colon, esophageal and mammary tissue. Deleterious effects of Bz were observed in adrenals, colon and esophagus. In addition, the metabolism of several xenobiotics biotransformed by the cytochrome P450 system can be inhibited by Bz. Studies have pointed out that both drugs exhibited mutagenic effects and can be tumorigenic or carcinogenic (Castro, de Mecca et al. 2006). Bz is usually used as first line option because of its better safety and efficacy profile. Antitrypanosomal treatment is strongly recommended for all cases of acute, and reactivated infection and for both children and adult patients. By contrast, it should not be offered to patients with severe Chagas heart symptoms or megaesophagus with swallowing difficulties. Treatment for patients at the chronic phase of Chagas disease usually focus on the cardiac and/or digestive symptoms.

1.2 The Genome of *T. cruzi*

The *T. cruzi* genome project started in the 1990s when a group of laboratories began to collect clone libraries, EST sequences and a few initial shorter genome sequences for the selected reference strain CL Brener. The strain was selected based on its clinical and laboratory characteristics: (1) it was isolated from domiciliary vector *Triatoma infestans*; (2) its infectivity in mice was well studied; (3) it prefers to invade smooth muscle cells; (4)it demonstrated a clear acute phase post invasion; (5) it was susceptible to current antiparasitic drugs (Zingales, Pereira et al. 1997; Teixeira, de Paiva et al. 2012). In 2005, the genome sequence of *T. cruzi* CL Brener was published, in parallel with the sequences of *Trypanosoma brucei* and *Leishmania major* (Berriman, Ghedin et al. 2005; El-Sayed, Myler et al. 2005; Ivens, Peacock et al. 2005). The project applied both a clone-by-clone strategy and a whole genome shotgun strategy, because of the high repeat content and hybrid nature of the genome. Based on the comparison of contigs with a low-coverage genome sequence that produced from the Esmeraldo genome, a parental strain of CL Brener belonging to *T. cruzi* IIb subgroup, they successfully distinguished reads of the two distinct haplotypes, named Esmeraldo-like and non-Esmeraldo-like. The genome was established by 5489 scaffolds (containing 8740 contigs), totaling 67Mb (El-Sayed, Myler et al. 2005). Based on the assembly results, the diploid genome was estimated to be approximately 110 Mbp in size. The G+C content of the entire genome is 51%. A total of 22,570 open reading frames (ORFs) were identified for both haplotypes, of which 6159 were from Esmeraldo-like haplotype, 6043 from the other and

10368 from either one. 3590 pseudogenes were also identified. The mean length of coding DNA sequence (CDS) is 1513 bp with a median value of 1152bp. The G+C content of the CDS region (53.4%) is significantly higher, compared to that of the inter-CDS region (47%), consistent with what has been reported to other systems as well, such as *C. elegans* and *D. melanogaster* (Zhang, Kasif et al. 2004; El-Sayed, Myler et al. 2005). Protein-coding genes were arranged as long polycistronic unit on the same DNA strand. Approximately half of the CDSs identified, no function could be assigned on the basis of significant similarity to previously characterized proteins or functional domains. One interesting finding from the genome analysis is the high degree of synteny between the two haplotypes with differences in intergenic and subtelomeric regions or amplification of repetitive sequences.

More than half of the *T. cruzi* genome consists of repetitive regions, consisting large gene families of surface antigens, retrotransposons and subtelomeric regions. Most chromosomes consist of core regions containing multiple polycistronic transcription units (PTU)(Liang, Haritan et al. 2003). The core regions are surrounded by the large repeated regions. Occasionally subtelomeric-like regions can occur inside core regions in a chromosome, probably caused by fusion of chromosomes. The regions between polycistronic units are referred to as strand-switch regions (SSR). Depending on the transcriptional orientation, the units can be convergent (transcriptional operons on opposite strands are converging towards the SSRs) or divergent

(transcriptional operons start on opposite strands of the SSRs and diverge from one another) (Riou and Yot 1977; Liang, Haritan et al. 2003; Ouellette and Papadopoulou 2009). The length of SSR varies from 100bp to 10kb. Convergent SSRs are found to be potential transcription termination sites, while divergent SSRs are usually associated with transcription initiation.

In 2009, the majority of the contigs and scaffolds were assembled into pairs of homologous chromosomes on the basis of predicted parental haplotype, inference from TriTryp synteny maps and the end sequences from *T. cruzi* bacterial artificial chromosome (BAC) libraries (Weatherly, Boehlke et al. 2009). A total of 41 pairs of chromosomes were assembled, in agreement with the predicted number of *T. cruzi* chromosomes based on pulse field gel analysis (**Figure 3**). Over 90% of annotated genes were included in the assembly of chromosomes. The majority of genes excluded from the chromosomes belong to multi-gene families, as the repetitive nature of those genes made accurate position almost impossible (Weatherly, Boehlke et al. 2009).

In addition to investigations of the nuclear genome, several studies have examined the mitochondrial genome of kinetoplastid which contains interlocked DNA rings known as kinetoplast DNA (kDNA) (Brener 1973; Simpson 1973). In *T. cruzi*, a kDNA network consists of thousands of minicircles 0.5-10 kb in size and dozens of maxicircles 20-40 kb in size. *In vivo*, kDNAs condense into a disk-shaped structure. Maxicircles, like mitochondrial DNAs (mtDNAs) from other

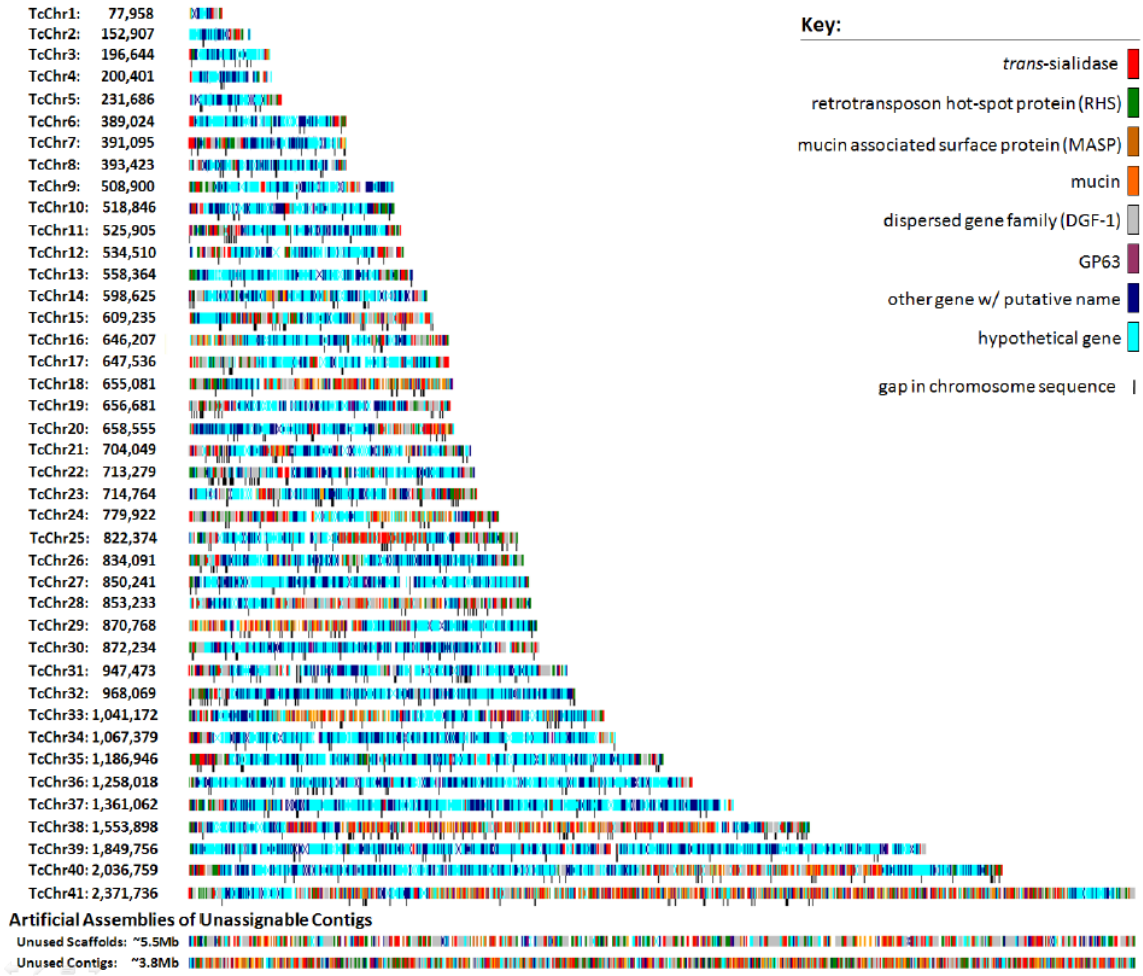

Figure 3 Chromosome assembly of *T. cruzi*.

A Model of assembled chromosomes represent the consensus view of both "Esmeraldo-like" and "non-Esmeraldo-like" haplotypes, since CL Brener reference strain is a hybrid of both lineages. Gene family members are depicted as non-blue colors (Weatherly, Boehlke et al. 2009).

eukaryotes, encode rRNAs and several proteins involved in energy transduction. As with other trypanosomatid mitochondrial genes, sequence analysis has reported that frameshift errors were observed in most genes on *T. cruzi* maxicircles and can be corrected at the RNA level by a complex uridine (U)-insertion/deletion process known as RNA editing (Hajduk, Harris et al. 1993). The information necessary for the proper insertion/deletion is present as small mitochondrial guide RNAs (gRNAs) encoded mainly by minicircles with a few exceptions encoded by maxicircles. The gRNAs covalently joined to the 3' end of an mRNA and undertake U insertion or deletion by editosome machinery (Stuart and Panigrahi 2002). The sequence of the 25kb *T. cruzi* maxicircles have 18 tightly clustered mitochondrial protein-coding genes and two rRNA genes, which are syntenic with those in *T. brucei* and *Leishmania tarentolae*. Comparative analyses of the mitochondrial genomes of different strains of *T. cruzi* revealed that outside the coding region existed strain-specific repetitive regions which enable researchers to examine phylogenetic relationships between different strains based on mitochondrial DNAs.

1.3 Genetic diversity of *T. cruzi*

The population of *T. cruzi* is highly heterogeneous and consists of a large number of strains with distinctive characteristics with regards to morphology, growth, virulence, pathogenesis, antigenic profile, metacyclogenesis, tissue tropism, and responses to drugs. For example, the digestive form of chronic Chagas disease is observed almost exclusively south of the Amazon basin (including Argentina, Brazil, Chile and Bolivia) and is only sporadic in northern South America, Central America, and Mexico. This geographical difference was contributed to the strain difference of *T. cruzi* in these areas (Miles, Feliciangeli et al. 2003; Campbell, Westenberger et al. 2004; Rassi, Rassi et al. 2010). In 1999, a consensus regarding the strains of *T. cruzi* was reached and two major lineages were identified: *T. cruzi* I and *T. cruzi* II. The *T. cruzi* I lineage predominates in the sylvatic transmission cycle and is less resistant to trypanocidal drugs. The *T. cruzi* II lineage predominates in the domestic environment and are characterized by their drug resistance. The *T. cruzi* II was further subdivided into five discrete typing units (DTUs): IIa, IIb, IIc, IId, and IIe. Phylogenetic analysis of the *T. cruzi* strains indicated the existence of the third major group of *T. cruzi* population. In 2009, six distinct DTUs were reported based on an international consensus: *T. cruzi* I-VI (**Table 1**). Commonly used experimental strains are scattered among the DTUs: Sylvio X10 (I), Y (IIb), CL Brener (IIe), and Tulahuen (IIe) (Teixeira, de Paiva et al. 2012).

Table 1 Classifications of *T. cruzi* strains.

Current designation[a]	Equivalence to former classifications[b]	Examples of representative strains
T. cruzi I	*T. cruzi* I/DTU I	Sylvio X-10, Dm28c
T. cruzi II	*T. cruzi* II/DTU IIb	Esmeraldo, Y
T. cruzi III	*T. cruzi* III/DTU IIc	CM17
T. cruzi IV	DTU IIa	CanIII
T. cruzi V[c]	DTU IId	SO3
T. cruzi VI[c]	DTU IIe	CL Brener

(Teixeira, de Paiva et al. 2012)

Two hybridization events define the population structure of *T. cruzi* (**Figure 4**). In the first one, a fusion between ancestral DTU I and IIb strains gave rise to a heterozygous hybrid that later homogenized its genome to become the homozygous progenitor of DTUs IIa and IIc. In the second event, the hybridization between DTU IIb and IIc strains and generated DTUs IId and IIe introduced extensive heterozygosity with subsequent recombination of parental genotypes (Westenberger, Barnabe et al. 2005).

Comparative genomics between different *T. cruzi* strains has been done between the Sylvio X10 clone from TcI DTU and CL Brener (Franzen, Ochaya et al. 2011). The Sylvio X10 genome (~ 44 Mbp per haploid genome) is considerably smaller than that of CL Brener (~55 Mbp per haploid genome) (Andersson 2011; Franzen, Ochaya et al. 2011). The genome synteny and gene identity are highly conserved with very limited differences noted in the core gene content between the two, however, 6 open reading frames were reported missing in Sylvio X10 (Franzen, Ochaya et al. 2011). The genetic diversity between the two stains is significant: 77,349 nucleotide differences and 1,861 insertion-deletion were identified in CDSs, 19% of which were associated with critical amino acid changes. Many large multigene families in Sylvio have fewer copies, compared to CL Brener, including MASP, glycoprotein 63 (GP 63), DGF, and mucin, which account for 50% of the size difference between the two strains. The repertoires of surface molecule genes are very divergent as well (Franzen, Ochaya et al. 2011)(**Figure 5**).

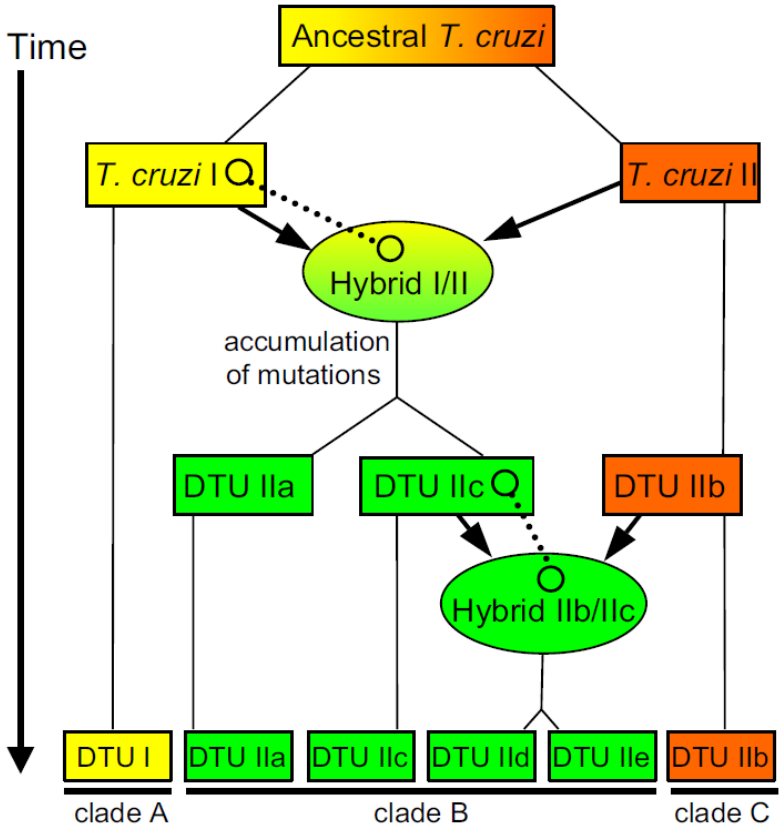

Figure 4 Two hybridization events define the population structure of *T. cruzi*.

The schema of the evolutionary history of *T. cruzi* subgroups highlighted two hybridization events: one between ancient Type 1 and Type 2; the other between DTU IIc and DTU IIb. The inheritance of maxicircles from parental donors is indicated by open circles and dotted lines. Based on the maxicircles, *T. cruzi* strains were categorized into three clades, which were color coded by yellow, green and orange (Westenberger, Cerqueira et al. 2006).

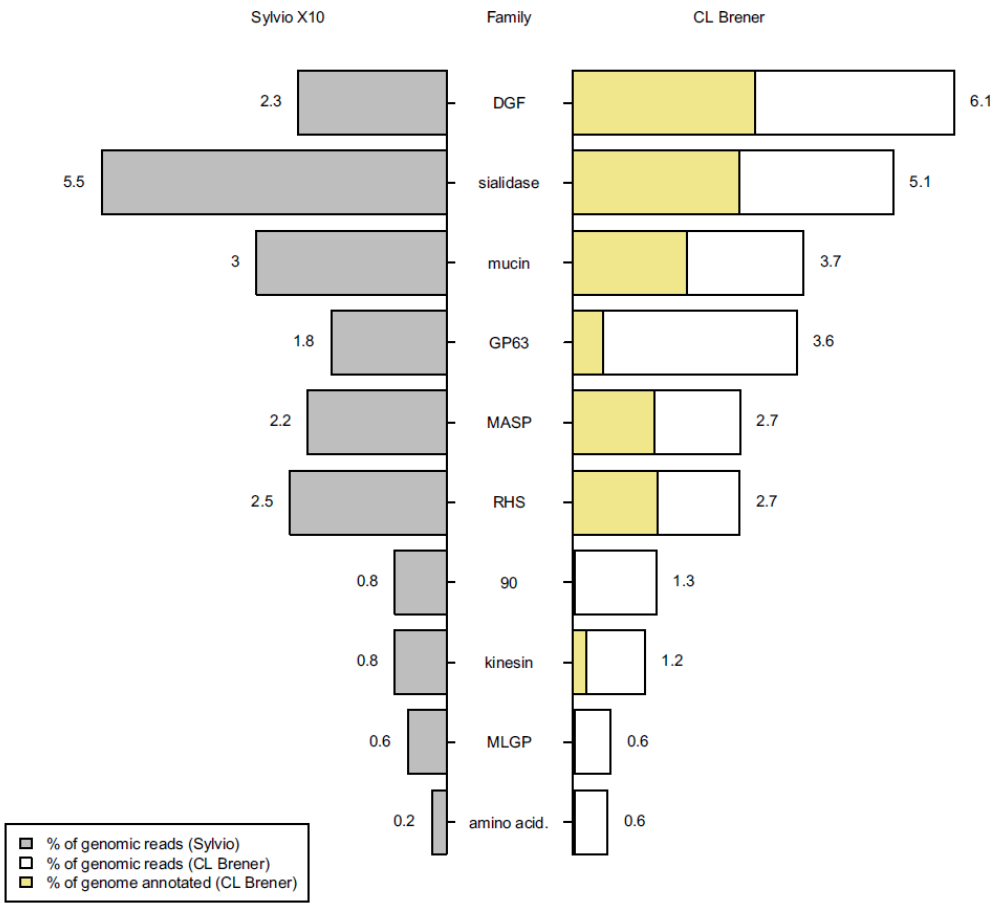

Figure 5 Comparison of gene content for some of the largest gene families between Sylvio X10 and CL Brener strain of *T. cruzi*.

Gene content of selected multigene faimilies were expressed as the percentage of genome and were label next to the bars. The yellow markings indicated the number of genes for each family in the annotated CL Brener. The gene families are: dispersed gene family 1 (DGF), trans-sialidases, mucins, glycoprotein 63 (GP 63), mucin-associated surface proteins (MASP), recombination hot spot (RHS), 90 kDa protein family (90), kinesin (one subfamily), mucin-like glycoprotein (MLGP), amino acid transporter (amino acid) (Andersson 2011).

It has been reported that the non-coding portion of maxicircles have strain-specific repetitive areas and a variable region that is unique in each strain with the exception of a conserved sequence motif that may serve as the start site of replication (Westenberger, Cerqueira et al. 2006). Previous studies have noticed uniparental maxicircle inheritance with *in vitro* crosses of *T. cruzi* strains (Gibson, Crow et al. 1997; Gaunt, Yeo et al. 2003). So it is not surprising that maxicircle sequences have been used as taxonomic markers in the population genetics and evolution analysis of *T. cruzi*. Three major clades of maxicircles have been identified from 45 *T. cruzi* strains including six defined DTUs (Machado and Ayala 2001; Flores-Lopez and Machado 2011). *T. cruzi* Y and Esmeraldo strains represent clade C, while CL Brener belongs to clade B. Integrating with the model of *T. cruzi* hybridization events, it has been proposed that clade B evolved from a clade A maxicircle passed by the common ancestor of DTU IIa/IIc strains after the DTU I and DTU IIb hybridization event (Brisse, Henriksson et al. 2003) (**Figure 4**). The clade B maxicircle was then inherited by DTU IId/IIe hybrids from their DTU IIc ancestor. Phylogenetic analysis of several loci on the maxicircles have confirmed the closer association of clade B with clade A than clade C, supporting the uniparental inheritance of DTU IIa/IIc maxicircles from DTU I strain (Westenberger, Cerqueira et al. 2006).

Since Chagas disease presents in a diversity of clinical forms, these studies can contribute to the understanding of genetic variation and their correlation with differences in disease pathogenesis, host preferences and provide a framework

for the selection of potential drug targets and antigenic candidates for better diagnosis and vaccine development. For instance, strains belonging to *T. cruzi* II, and the hybrid V and VI are the most virulent ones. Comparisons between the above strains and non-virulent strains may help identify an interesting gene list that are disease-causing (Lima, Lenzi et al. 1995). However, the degree to which the genomic variations are associated with strain differences in host preference and the ability to cause Chagas disease remains to be determined.

1.4 Gene expression in trypanosomes

Unlike other eukaryotes, *T. cruzi* genes in coding for proteins with unrelated functions are organized into co-directional clusters that undergo polycistronic transcription by RNA polymerase II (Pol II) (El-Sayed, Myler et al. 2005; Siegel, Tan et al. 2005). Most chromosomes contain at least two polycistronic gene clusters, which can be either divergently or convergently transcribed (Weatherly, Boehlke et al. 2009). Transcription initiates from divergent strand-switch regions (SSRs) and terminates at convergent SSRs, where tRNA genes are often located (although they can be present at non-SSRs) (Ouellette and Papadopoulou 2009; Siegel, Hekstra et al. 2009). Recent chromatin immunoprecipitation and sequencing (ChIP-seq) experiments examining the distribution of chromatin components in other trypanosomes revealed that histone post-translational modifications can direct the regulation of transcription and play a crucial role in polycistronic transcription initiation and termination (Siegel, Hekstra et al. 2009). Destabilization of nucleosomes by histone variants is an evolutionarily ancient and general mechanism of transcription initiation.

Trans-splicing, together with polyadenylation, allows polycistronic transcripts to be processed into monocistronic units ready for translation (El-Sayed, Myler et al. 2005; Daniels, Gull et al. 2010)(**Figure 6**). In the mini exon region, or the splice leader locus, each gene possesses a Pol II promoter region. During *trans*-splicing event, a 39-nt splice leader sequence is transferred from SL RNA to the 5′ end of every mRNAs, providing the cap structure needed (Agabian 1990;

Liang, Haritan et al. 2003). The signal of *trans*-splicing events has been reported as an AG dinucleotide and a polypyrimidine tract of varying length upstream of it (Michaeli 2011). *Trans*-splicing events are spatially and temporally coordinated with the polyadenylation events (LeBowitz, Smith et al. 1993; Matthews, Tschudi et al. 1994). There is no consensus polyadenylation signal in the 3′ UTR. Instead, evidence from a small number of loci in related trypanosome species *T. brucei* has suggested a preference usage of polyadenylation sites around adenines (Benz, Nilsson et al. 2005). However, no such analysis has been done for *T. cruzi*. Current knowledge suggests that trypanosomatids lack precise transcriptional control because no classical promoters have been identified (Siegel, Gunasekera et al. 2011). Regulation of gene expression is mainly at the post-transcriptional level, and it has been proposed to occur through pre-mRNA processing, RNA degradation, or translational repression (Martinez-Calvillo, Vizuet-de-Rueda et al. 2010). Both the 5′ UTR and 3′ UTR can be involved in stabilization-destabilization mechanisms, up-regulating and down-regulating mRNA levels in a developmentally regulated manner. Heterogeneity of RNA processing sites, present as alternative *trans*-splicing and polyadenylation sites, have been detected at high frequency across the genome in *T. brucei* and *T. vivax*, but in *C. elegans*, of which the *trans*-splicing events are common, the frequency of alternative *trans*-splicing is much lower (Hirsh and Huang 1990; Graber, Salisbury et al. 2007). The significance of this phenomena and their possible roles in the posttranscriptional regulation has not been fully investigated (Kolev, Franklin et al. 2010; Greif, de Leon et al. 2013).

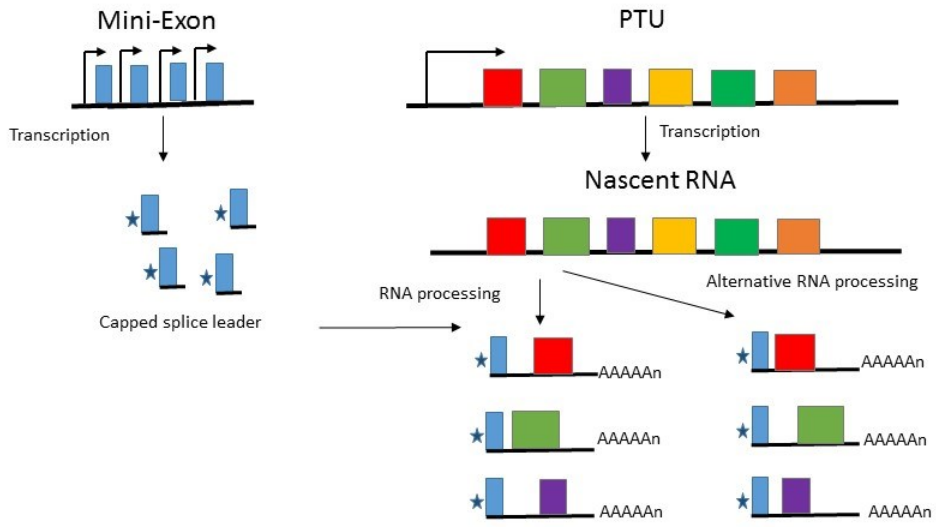

Figure 6 Gene expression in trypanosomatids.

Large clusters of genes coding unrelated functions are organized into polycistronic units (PTU), designated by rectangles in different color. Transcription start sites (TSS), indicated by the arrow on transcript, are usually located upstream of the first gene of the PTU. The premature polycistronic RNA is transcribed by RNA pol II, and is cleaved into individual RNAs by the coupled trans-splicing and polyadenylation reactions. A 39-nt splice leader sequence will be added to the 5' end of each individual RNA (blue rectangular), whereas a poly(A) tail will be added to the 3' end, represented by AAAA. The cap in the SLRNA is indicated by an asterisk at the 5' end of the RNA.

1.5 Comparative genomics of trypanosomes

The publication of genomic sequences of *Trypanosoma brucei* (Berriman, Ghedin et al. 2005), *Trypanosoma cruzi* (El-Sayed, Myler et al. 2005) and *Leishmania major (Ivens, Peacock et al. 2005)*provided a foundation for researchers to investigate the similarity and differences of these parasites at the genomic scale and understand their distinctive adaptation strategies in different host environments. *T. brucei* and *L. major* are the etiologic agents of sleeping sickness (African trypanosomiasis) and leishmaniasis, respectively. These trypanosomes are hemoflagellates of the family Trypanosomatidae belonging to the order Kinetoplastida and are characterized by the presence of a single flagellum and kinetoplast, which contains mitochondrion. Each parasite has a complex life cycle involving both human host and insect vectors: *T. cruzi*, a triatomine bug, also known as the “kissing bug”; *T. brucei*, a blood-sucking fly, known as the tsetse fly; *L. major*, a phlebotomine sandfly. Both *T. cruzi* and *L. major* are intracellular parasites, but *T. cruzi* have the capacity to invade various nucleated host cells while *L. major* usually targets macrophages. *T. brucei* is an extracellular parasite and can rapidly manipulate their membrane proteins and escape the attack from host immune responses.

Comparative genomic analysis provided interesting insights in the genetic and evolutionary facts that may explain the distinct and shared characteristics of the parasites. Strikingly, large-scale highly conserved synteny between the three trypanosomes were observed despite the fact that the three parasites diverged

200 to 500 million years ago, indicating strong selective pressure of the gene order and orientation. An ‘‘all-versus-all’’ basic local alignment search tool (BLASTP) comparison of the predicted protein sequences within each of the three genomes was made to cluster closely related paralogous genes. The mutual best BLASTP hits between the three proteomes were grouped as clusters of orthologous genes (COGs). Approximately 6,200 genes were identified as three-way COGs, also known as the Tritryp core, presenting an interesting gene list of potential drug targets against the three diseases at the same time (**Figure 7**). About 2,000 out of the 6200 genes do not have any orthologs in human host, making chemotherapeutic intervention even more feasible. Almost all genes in the Tritryp core are located in regions of conserved synteny. Approximately 40% of the synteny breakpoints are associated with expansions of multi-gene families, retroelements, and/or structural RNAs, which also has been observed in mammalian systems (Gimelli, Pujana et al. 2003; Bailey, Baertsch et al. 2004). Marked difference in gene size and density between the three trypanosomes is significant. The average length of CDSs in *L. major* is substantially longer than that of *T. cruzi* and *T. brucei*. Since the boundaries of transcripts were not defined by the time this comparison analysis took place, the authors pointed out the different inter-CDS length between Tritryps: *L. major* has the longest one, twice that in *T. brucei* and three times that in *T. cruzi*. Consequently, *T. cruzi* has the highest gene density of the three. Amino acid sequence alignment of the three-way COGs revealed an average 57% identity between *T. cruzi* and *T. brucei* and 44% identity between *L. major* and the other two trypanosomes,

consistent with expected phylogenetic relationships. In addition, the comparative genomic analysis identified species-specific gene clusters encoding proteins responsible for unique form of each disease and different survival strategies applied by each parasite (**Figure 7**).

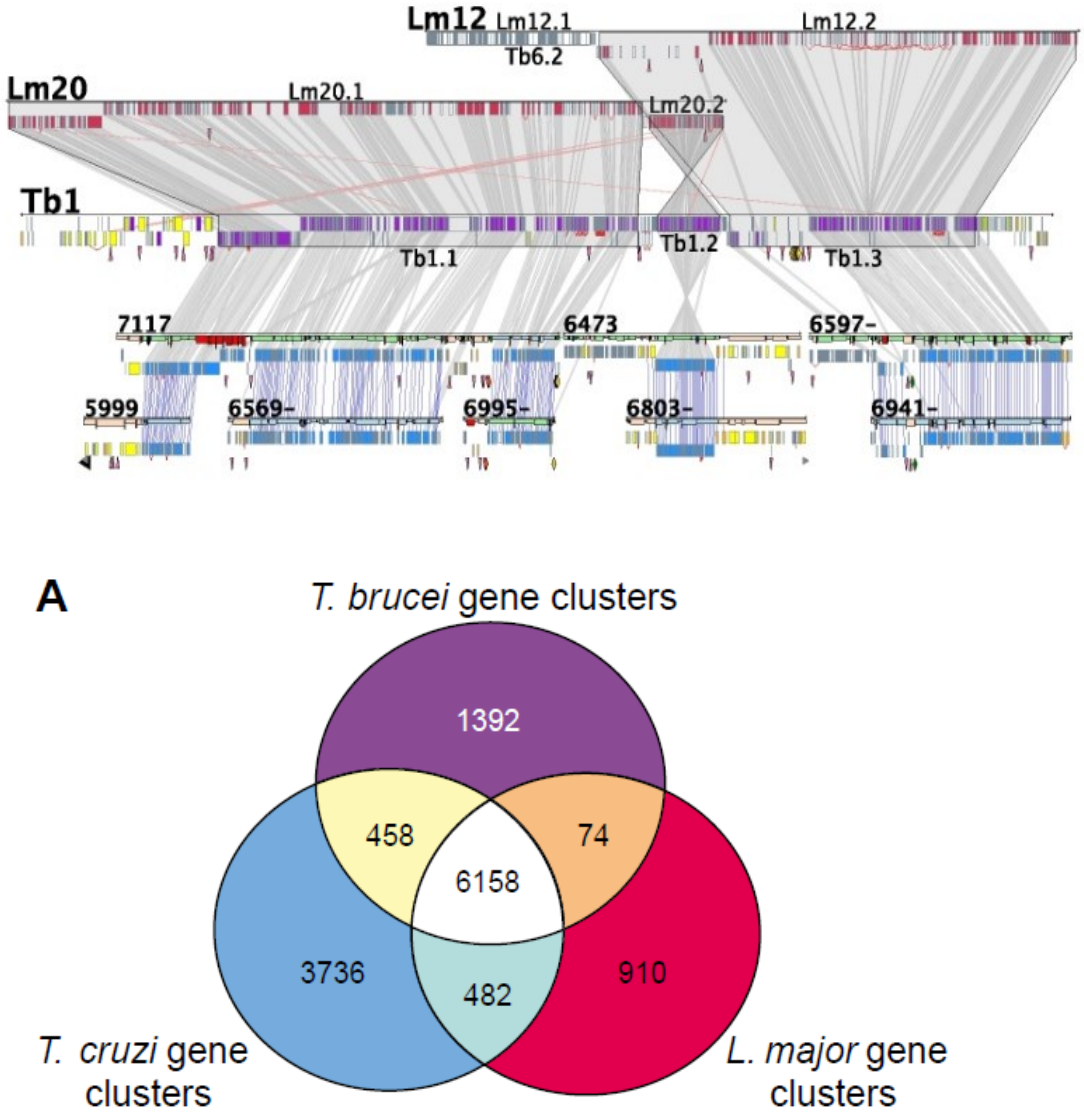

Figure 7 Comparative genomics of Tritryps.

(A) Conserved synteny was observed between the genomes of Tritryps. Synteny breaks are located mainly in subtelomeric region. (Red: *Leishmania major*, Purple: *Trypanosome brucei*, Blue: *Trypanosome cruzi*) (B) A Venn diagram of distribution of genes among the kinetoplastid parasites, calculated with the use of Jaccard-filtered COGs. (El-Sayed, Myler et al. 2005)

1.6 RNA sequencing technology

Over the past decade, expression microarray is the most widely used methodology for transcriptome analysis, but it is approaching its technical limits including dye-based hybridization and cross hybridization artifacts. In addition, it cannot detect any new genes besides the predicted ones. Microarray technology can only measure expression levels within a certain range, due to noises produced by non-specific cross-hybridization, saturation, spot density and quality. Also, comparison between different experiments can be very difficult and may require complex statistical methods. (Euskirchen, Rozowsky et al. 2007; Bloom, Khan et al. 2009; Fu, Fu et al. 2009; Bradford, Hey et al. 2010; Griffith, Griffith et al. 2010). The above issues have made it difficult for standard array designs to provide an accurate and comprehensive detection for transcriptome, especially for *T. cruzi* of which genome has relatively low variation and less steady state mRNA (Minning, Weatherly et al. 2009). RNA-Seq as a recently developed technology is an optimum solution since large scale high-throughput mRNA sequencing can reveal exclusive details of expression profile and transcription pattern for *T. cruzi* in its life stages. It can also be used in reconstructing full length transcripts, correcting 3′ and 5′ UTRs, detecting novel genes and genomic variation at the nucleotide or transcript level. (Cloonan and Grimmond 2008; Marioni, Mason et al. 2008; Mortazavi, Williams et al. 2008; Pepke, Wold et al. 2009; Wang, Gerstein et al. 2009; Griffith, Griffith et al. 2010; Wilhelm, Marguerat et al. 2010) (**Figure 8**).

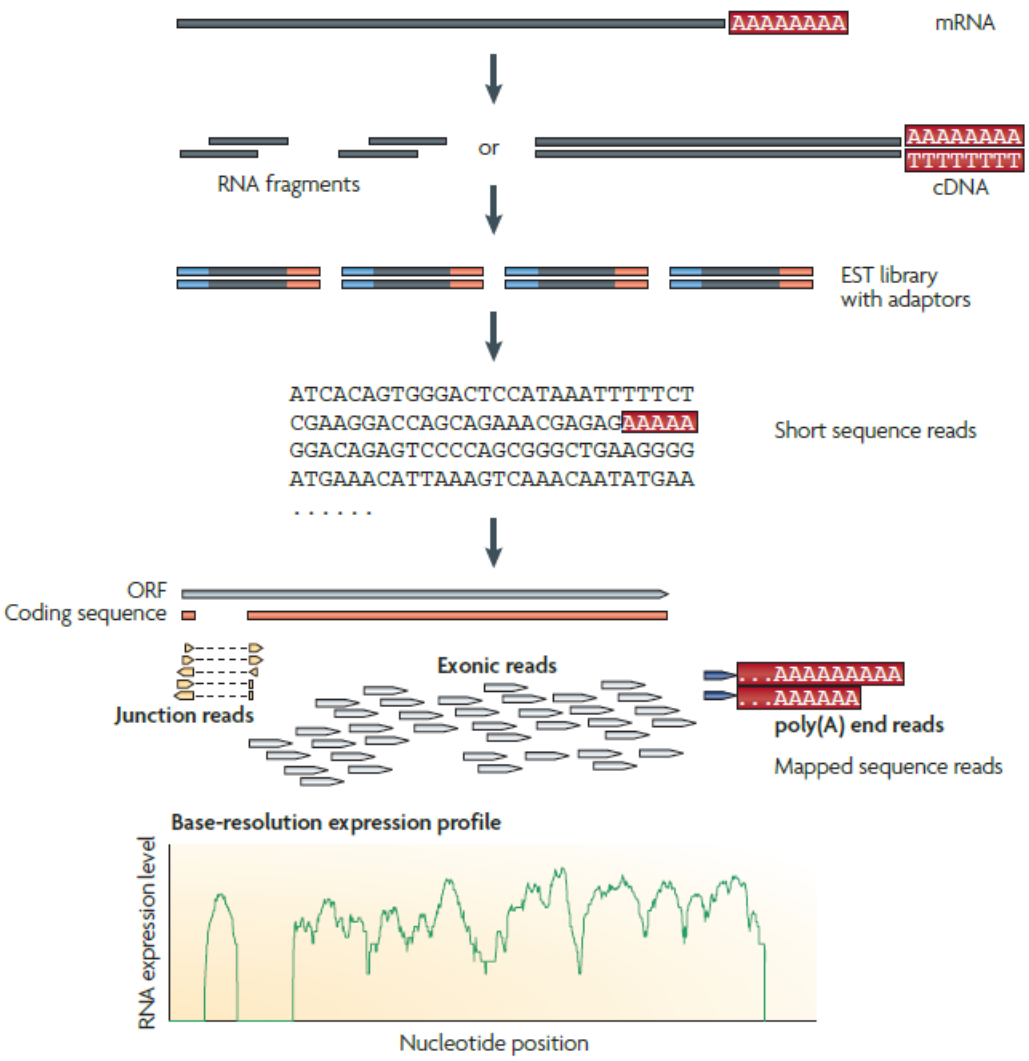

Figure 8 Overview of a typical RNA-Seq experiment.

Long RNAs are converted into cDNA libraries by either RNA fragmentation or DNA fragmentation. Sequencing adaptors are added to the ends of each cDNA fragments. One or both ends of the DNA fragment are sequenced on high-throughput parallel sequencing platform. Subsequently, the obtained reads are aligned to the reference genome or transcriptome (Wang, Gerstein et al. 2009).

The library construction protocol has been optimized for RNA-Seq experiments in trypanosomes when the aim of study is to capture the RNA processing events (**Figure 9**). Both ends of the transcripts can be enriched by taking advantage of the unique splice leader sequence at the 5' end of trypanosome RNA or the poly(A) tail at the 3' end (Kolev, Franklin et al. 2010; Mulindwa, Fadda et al. 2014). First-strand cDNA synthesis is generated with random hexamers or oligo(dT) primers, then second strand cDNA can be reverse transcribed with SL-specific primers. A clear preference for the 5' end and 3' end of transcripts were noted. Similar method has also been proposed for the enrichment of parasite RNA from a pool of RNAs from both host and pathogen. However, distortion of transcriptomic profiles have been reported in the latter case (Mulindwa, Fadda et al. 2014). So unless the main purpose of RNA-Seq is the characterization of RNA processing sites, random priming should be applied in library construction, instead of SL-specific priming.

The pipeline for a typical RNA-Seq differential expression (DE) analysis is outlined in **Figure 10**. First, reads are aligned to the reference genome or transcriptome. Second, the number of reads mapped to each feature (gene, exon, junction or transcript) are summarized into a count table. Next, the gene table of counts will be normalized followed by statistical testing of DE. The resulting output is usually a ranked list of genes with P value or fold changes. Finally, biological insight from the differentially expressed genes can be obtained by Gene Ontology, or network analysis (Oshlack, Robinson et al. 2010).

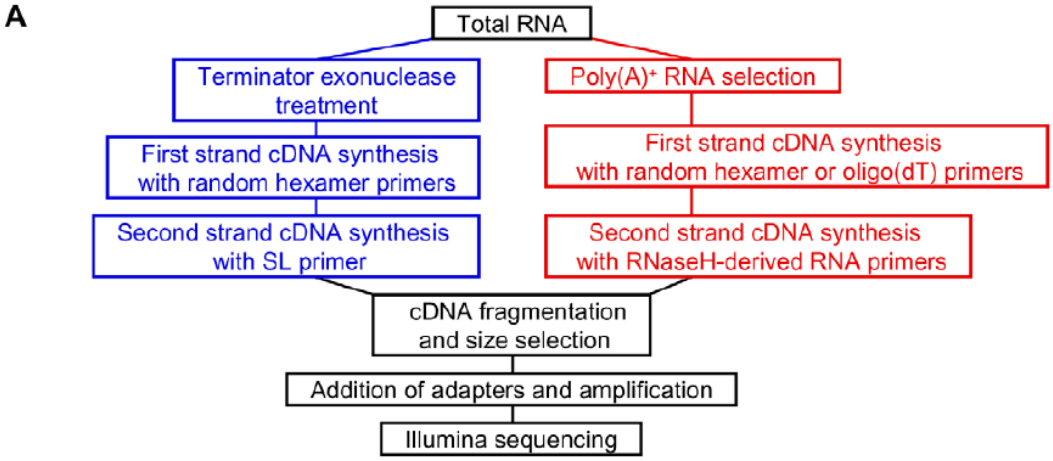

Figure 9 Outline of an end-enriched RNA-Seq experiment for trypanosomes.

To capture more reads that are generated from the 5' end of the mRNAs, an SL-specific primer can be used for the synthesis of the second strand of cDNA (blue texts). To capture more reads that generated from the 3' end of the mRNAs, an oligo(dT) primer can be used for the synthesis of the first strand of cDNA (red texts). (Kolev, Franklin et al. 2010)

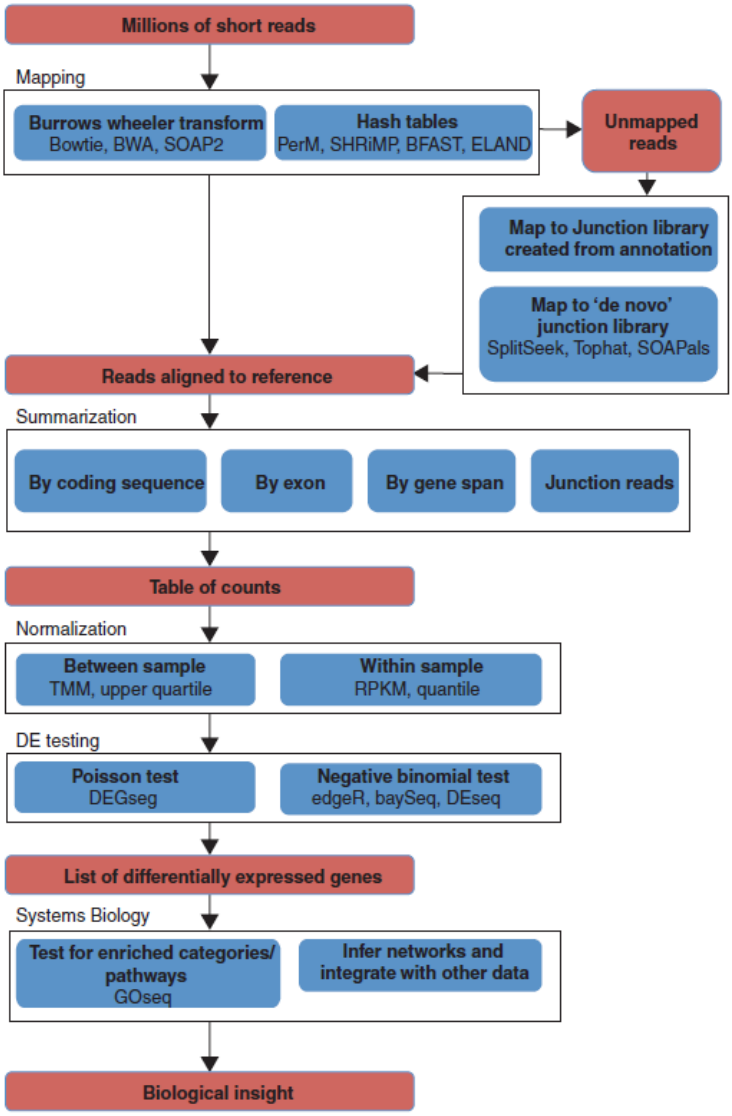

Figure 10 A general overview of the pipeline of RNA-Seq analysis for detecting differential expressed genes.

Red boxes describe the steps applied in the pipeline; blue boxes display the methodological components and software examples of the pipeline. Major processes of differential expression analysis include mapping of reads to reference genome, summarization, normalization, DE statistical testing, and function analysis. (Oshlack, Robinson et al. 2010)

1.7 Summary of dissertation work

The long-term goal of this project was to understand the biology of the human parasite *Trypanosoma cruzi* across different developmental stages of its life cycle, characterize the host-pathogen interaction, and contribute to the identification of potential drug targets. *T. cruzi* is a kinetoplastid parasite with a very complex life cycle involving both human host and insect vector. Our work here included two major components: the first part is the characterization of RNA processing events in *T. cruzi* developmental stages pre- and post-infection of human host cells; the second is simultaneous interrogation of the transcriptome of the human pathogen *Trypanosoma cruzi* and its infected host cell.

The publication of the genome sequence of *T. cruzi* CL Brener in 2005 represents a major advance that has contributed to the understanding of the biology of the Chagas disease parasite (El-Sayed, Myler et al. 2005). Yet a systematic genome-wide identification of transcripts has not been conducted for *T. cruzi* and a major challenge remains to identify the *bona fide* transcripts and the exact boundaries of the transcripts. Unlike other eukaryotes, trypanosome genes coding for proteins with unrelated functions are organized into co-directional clusters that undergo polycistronic transcription by RNA polymerase II (Pol II)(El-Sayed, Myler et al. 2005; Siegel, Tan et al. 2005; Siegel, Gunasekera et al. 2011). *Trans*-splicing, together with polyadenylation, allows polycistronic transcripts to be processed into monocistronic units ready for translation (Sutton and Boothroyd 1986; Muhich and Boothroyd 1988; El-Sayed, Myler et al. 2005;

Daniels, Gull et al. 2010). The lack of identifiable RNA pol II promoters in trypanosomatids suggests that these organisms lack precise transcriptional control over the majority of their genes (Siegel, Gunasekera et al. 2011). Thus, regulation of gene expression occurs mainly at the post-transcriptional level, through pre-mRNA processing, RNA degradation, or translational repression (Martinez-Calvillo, Vizuet-de-Rueda et al. 2010). Both the 5′ UTR and 3′ UTR can be involved in stabilization-destabilization mechanisms, up-regulating and down-regulating mRNA levels in a developmentally regulated manner. Heterogeneity of RNA processing sites, present as alternative *trans*-splicing and polyadenylation sites, have been detected at high frequency across the genome in *T. brucei* and *T. vivax*. The significance of this phenomena and its potential role in posttranscriptional regulation has not been fully investigated (Kolev, Franklin et al. 2010; Greif, de Leon et al. 2013). In our current study, we have generated the first complete transcriptome map for *Trypanosoma cruzi* with the next generation RNA sequencing technology (RNA-Seq). The analysis documented in Chapter 2 focused on the characterization of RNA processing events across different developmental stages of *T. cruzi*. We mapped transcribed regions at single nucleotide resolution on a genomic scale, retrieved *trans*-splicing and polyadenylation sites for *T. cruzi* across various developmental stages, and enhanced and curated the current genome annotation. In addition, with the unprecedented resolution of transcriptome, we discovered the prevalent heterogeneity of RNA processing sites across the genome. The preference of different primary sites were noted in various developmental stages, which

presents as a potential and interesting approach for posttranscriptional regulation.

Great efforts have been made in the understanding of interactions between the Chagas disease parasite and human host. However, little is known about the strategies exploited by the pathogen during the infection process. In this study, we present the transcriptomes of three main *T. cruzi* life cycle stages including the simultaneous capture of host and parasite transcriptomes in an intracellular infection time course *in vitro*. Using the *T. cruzi* and human genome sequence as scaffolds, we explored these data with informatics tools to identify genes with significant regulation and successfully profiled gene expressions from both species simultaneously. The present study compares the steady state transcriptomes of three main life cycle stages of *T. cruzi*: (1) tissue culture-derived trypomastigotes: non-dividing, tissue penetrating forms; (2) intracellular stages up to and including replicating amastigotes; and (3) axenically-grown epimastigotes which correspond to the replicating insect vector stage of the parasite. A comparison of steady state transcriptomes revealed a significant number of differentially expressed genes between distinct life cycle stages of the parasite, reflective of their divergent biology. Distinct signatures of gene expression are observed as the motile invasive forms of *T. cruzi* transition into non-motile replicative forms in mammalian host cells. Gene Ontology (GO) enrichment of K-means clusters revealed the function associations between genes with similar expression patterns, while motif analysis of UTRs of clustered

genes identified consensus regulatory elements as potential binding sites for *trans*-acting factors. In addition, we simultaneously captured changes in the steady state transcriptome of infected human host cells and identified many significantly regulated genes detected in human host throughout the infection cycle contributing to the understanding of the host responses after invasion.

1.8 Significance

Trypanosoma cruzi is the etiological agent of Chagas disease, which is ranked as one of the most important pathogens throughout Central and South America. In spite of the intensive investigation and rigorous medical control initiatives over the past decades, Chagas disease remains a significant danger to human health. Currently no vaccine is available and treatments are limited and toxic. In this study, we generated the first complete transcriptome map for *Trypanosoma cruzi* and infected human host cells with the next generation RNA sequencing technology (RNA-Seq). We conducted a systematic characterization of *bona fide* transcripts and genomic structures, which can provide a fundamental and important framework for the better understanding of both transcriptional and posttranscriptional regulation mechanisms in the pathogen. The identification of differentially expressed genes from both the parasite and host across different time points of the infection process as well as distinctively regulated genes in *T. cruzi* at its extracellular stages will help elucidate the survival strategy of the parasite applied at various environments, the defense response from the human host post invasion, as well as the biology of host-pathogen interaction. This knowledge can eventually contribute to the development of effective medical intervention, the improvement of public health in the developing countries where Chagas disease posts as great burden and danger.

Chapter 2: RNA Processing Events in *T. cruzi* Developmental Stages Pre- and Post-Infection of Human Host Cells

2.1 Objective of Study

Unlike the majority of eukaryotes, typical transcription unit of the parasite *T. cruzi* does not have a classical promoter region and protein-coding genes with unrelated functions are organized into co-directional clusters that undergo polycistronic transcription by RNA polymerase II (Pol II). Despite the remarkable progress in the sequencing of trypanosomatid genomes, we still do not have a good understanding of the RNA processing events. Various genomic structure components, including UTRs, polypyrimidine tract, acceptor sequence, as well as intergenic regions, were not annotated in *T. cruzi*. In this study, we aim to construct the transcriptome of *T. cruzi* at single nucleotide resolution, characterize the RNA processing events across its different developmental stages (trypomastigote, epimastigote, and amastigote), as well as enhance and curate the current genome structure annotation.

2.2 Materials and Methods

2.2.1 Materials

Trypanosoma cruzi Y strain (Silva 1953) was cultivated by weekly passage in LLcMK2 cells (ATCC®CCL-7) in Dulbecco's modified Eagle medium (DMEM) with 2% fetal bovine serum (FBS), 2 mM L-glutamine, 10 mM HEPES and penicillin-streptomycin. Greater than 95% pure trypomastigotes were obtained

from the pelleted culture supernatants of infected LLcMK2 (1,000g for 10 minutes) after a 2-4 hour incubation at 37^{o}C, 5% CO_2 to allow the motile trypomastigotes to swim away from the pellet. Human foreskin fibroblasts (HFF) (ATCC®CRL-2522) were seeded in complete DMEM (as above, with 10% FBS) in 10 cm^2 plates or T-25 flasks with a glass coverslip and grown for 48 hours prior to infection with *T. cruzi* Y strain trypomastigotes for 2h, washed 5 times with PBS, and incubated in complete media. At the indicated time points (4-72 hpi), cells were rinsed with PBS and the coverslip removed and fixed in 4% paraformaldehyde/PBS. Epimastigote forms of *T. cruzi* were grown axenically in liver infusion tryptose medium at 27^{o}C.

2.2.2 Methods

Alignment of Spliced Leader (SL) reads and 5' UTR analysis

Sequences were pooled from biological replicates for each time point (4 hr, 6 hr, 12hr, 24 hr, 48 hr, 72 hr, trypomastigote and epimastigote). The 5' end of all reads were searched for the presence of at least 10 nucleotides from the 3' end of the SL sequence using the 'grep' function in Linux. The identified SL-containing reads and their respective paired mates were extracted for further analysis. Detected SL nucleotides (ranging from 10-39) were removed from the SL-containing read. Tophat (v2.0.6) was used to map the remaining portion of the SL-containing reads and their respective paired reads to the *T. cruzi* CL Brener reference genome Esmeraldo haplotype (El-Sayed, Myler et al. 2005; Trapnell, Pachter et al. 2009). A maximum of two mismatches per read were

allowed and multireads (reads that can be mapped to multiple loci) were mapped to the site with best alignment score. The coordinates and sequences of coding regions (CDS) were downloaded from TriTrypDB, version 4.1 (www.tritryps.org). For each instance where the SL-containing read occurred between two consecutive annotated CDSs, the alignment coordinates of the non-SL portion of the SL-containing read was used to retrieve the exact location of the *trans*-splicing site. The length of the 5' UTR was defined as the distance between the *trans*-splicing site and the start of the CDS. All sequence manipulations, *trans*-splicing and 5' UTR site analysis were performed using in-house Python or Linux shell scripts. For instances where the SL-containing read occurred inside annotated CDSs, the alignment coordinates of the non-SL portion of the SL-containing read were also collected as an additional dataset. Data visualization was carried out on R platform (Team 2008) with plotting packages ggplot2 installed (Wickham 2009).

Alignment of poly(T)-containing reads and 3' UTR analysis

Sequences were pooled from biological replicates for each time point (4 hr, 6 hr, 12hr, 24 hr, 48 hr, 72 hr, trypomastigote and epimastigote). The 5' end of all reads were searched for the presence of at least 5 nucleotides of thymidine using the 'egrep' function in Linux. The reason we chose to focus on poly(T)-containing reads, instead of poly(A)-containing reads was due to the fragment size of the cDNA library and the orientation of sequencing. Because the average fragment size was around 300bp and the reads were 101 bp, sequenced from 5' to 3' of the fragment. So the probability of a read reaching the poly(A) tail from the 5' end

of the fragment is very small unless the fragment size is much smaller 200 nt. On the contrary, its paired mate can be detected containing the poly(T) sequence at its 5' end. The identified poly(T)-containing reads and their respective paired mates were extracted for further analysis. Detected thymidine nucleotides were removed from the poly(T)-containing read. Tophat (v. 2.0.6) was used to map the remaining portion of the poly(T)-containing reads and their respective paired reads to the *T. cruzi* CL Brener reference genome Esmeraldo haplotype .A maximum of two mismatches per read were allowed and multireads (reads that can be mapped to multiple loci) were allowed to map to up to 20 locus. We applied a relatively less stringent strategy when mapping poly(T) reads in the polyadenylation analysis because we want to capture as many polyadenylation sites as possible at the initial steps as the number of poly(T) reads were much lower than SL-containing reads and depended on criteria described below to ensure the confidence of sites detected . For each instance where the poly(T)-containing read occurred between two consecutive annotated CDSs and the number of thymidines removed from the poly(T)-containing read exceeds the number of adenines in the genome at the same loci, the alignment coordinates of the non-poly(T) portion of the poly(T)-containing read was used to retrieve the exact location of the polyadenylation site. The length of the 3' UTR was defined as the distance between the polyadenylation site and the end of the CDS. All sequence manipulation, polyadenylation 3' UTR site analysis were performed using in-house Python or Linux shell scripts. Data visualization was carried out on R platform with plotting packages ggplot2 installed.

Alternative Start Site and Missing ORFs analysis

All 5' UTRs longer than 20 nt were scanned to identify potentially extended open reading frames. Alternative start codon(s) inside the 5' UTRs were detected if they were upstream and in-frame with the currently annotated start codon. The scanning and detection steps were performed using in-house Python scripts.

If 1) a 5' UTR or 3' UTR contained both a start codon and an in-frame stop codon and 2) the newly identified ORF was greater than 180 nt, then a potential missing ORFs was tagged. Novel ORFs were further selected from the pool by meeting the minimum coverage standard: the sum of reads across all samples mapped to the region had to be more than the number of samples n (here n=10). Novel ORFs with both *trans*-splicing and polyadenylation sites detected were further selected. All ORF-containing UTRs were removed from our gene structure analysis. The function and orthology of newly identified ORFs were further investigated by querying the RefSeq protein database [The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 18, The Reference Sequence (RefSeq) Project] with NCBI BLASTX algorithm (Gish and States 1993). The top hit of each query were outputted. Of all the novel ORFs, potential GPI-anchored proteins were predicted by FragAnchor software (Poisson, Chauve et al. 2007) [website: http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html], proteins with potential trans-membrane domains were predicted by software TMHMM v2.0 (Moller, Croning et al. 2001) and proteins with signal peptide were

predicted by SignalP v4.1 (Petersen, Brunak et al. 2011). Putative domains were identified by searches against the Conserved Domain Database (CDD) (Marchler-Bauer, Zheng et al. 2013).

Alternative splicing analysis

The *trans*-splicing site with the largest number of SL-containing reads mapped was defined as primary. Alternative splicing events were identified for each of the developmental stages and pooled together for some of the analyses. We focused our analysis on the developmental stages of trypomastigote, epimastigote, and amastigotes 72 hr post invasion because the number of SL-containing reads were significantly higher in these three stages. Primary *trans*-splicing sites with less than three reads mapped were filtered out in the downstream analysis. In order to capture genes with dominant usage of the primary site, we selected genes with a ratio of number of reads mapped to the primary site over the number of reads mapped to the secondary site greater than 1.6 (P/S > 1.6) at all three stages. We generated next a gene list by selecting genes that displayed a switch in the primary *trans*-splicing site between any two of the three developmental stages and with secondary sites at least 20 nt away from the primary site. Similar analysis were also conducted for secondary sites at least 50, 100, and 200 nt away from primary *trans*-splicing sites.

Splice acceptor site analysis

The splice acceptor site was identified for each gene by extracting the dinucleotide sequence in the genome upstream of the detected *trans*-splicing site. The sequence composition of the region spanning 90 nt upstream and 10 nt downstream of the *trans*-splicing sites, including both primary and minor sites, was plotted using WebLogo, version 3.3 (Crooks, Hon et al. 2004).

Polypyrimidine tract characterization

A window of 250 nt upstream of the primary *trans*-splicing site was scanned to identify polypyrimidine (polyPy) tracts. A polyPy tract was defined as the longest stretch of sequence consisting of pyrimidines, allowing interruption by no more than a single purine. The length of the polyPy tract, the distance between the 3' end of the polyPy tract and the *trans*-splicing site and the distance between the 3' end of the pyrimidine tract and the polyadenylation site were computed. The nucleotide composition of the polyPy tract was visualized by WebLogo, version.3.3 (Crooks, Hon et al. 2004).

Alternative polyadenylation analysis

The polyadenylation site with the largest number of poly(T)-containing reads mapped was defined as primary. Alternative polyadenylation events were identified for each of the developmental stages and pooled together for the analyses.

2.3 Results

Experimental samples interrogated using RNA-Seq

We applied an RNA sequencing approach to characterize the global transcriptome of the *T. cruzi* (Y strain) parasite across various time points in its life cycle. Those included two extracellular forms (epimastigote and trypomastigote) and the intracellular forms (amastigotes) at 4, 6, 12, 24, 48 and 72 hrs post-invasion of human foreskin fibroblasts (HFFs) cells. *In vitro* infection experiments were repeated on different dates and for each of the developmental stages, we collected sequence data from two to four independent biological replicates, generating a total of 1.35 billion pairs of 100 bp reads from 34 samples (**Table 7, see appendix**). Because the goal of this component of our study was to characterize the gene structure in *T. cruzi*, only ~400 million reads mapping to the parasite genome were used (**Table 2**)(**Tables 7, 8, S1, see appendix**).

Table 2 Summary of mapping of reads and reads containing RNA processing features.

Parasite Stage	Human (millions)	*T. cruzi* (millions)	SL-containing reads	Poly(A) containing reads
Extracellular	0	276	1.07%	0.27%
Intracellular	1,963	122	1.38%	0.28%
	1,963	398	1.16%	0.27%

Characterization of transcript boundaries and gene structure elements related to RNA processing

The significance of defining distinct transcript boundaries and gene structure regulatory elements for each of the 10,343 *T. cruzi* genes is of particular relevance to the biology of this pathogen which, like other trypanosomatids, lacks transcriptional control of its polycistronic gene clusters. We exploited two mRNA sequence features (*trans*-splicing of a mini-exon sequence and polyadenylation) to accurately map the 5′ and 3′ boundaries of genes. We selected subsets of reads that ended with at least a 10-nt match to the 3′ portion of the spliced leader (SL) sequence or at least five thymine residues. A total of 4,631,345 SL-containing and 1,055,377 poly(T)-containing reads were identified, respectively 1.16% and 0.27% of the reads that can be mapped to *T. cruzi* reference genome (**Table 2**)(**Tables 6, and 7, see appendix**). Trimming of the SL and poly(T) sequences and mapping the remaining tags back to the genome allowed us to identify at least one SL-addition site for 7,869 distinct genes (76% of annotated total) and at least one polyadenylation site for 6,311 distinct genes (61%) in one or more of the developmental stages. When we increase our mapping stringency by requiring at least three SL-containing or poly(T)-containing reads to assign the primary site for each feature, the number SL-addition site detected decreased to 7,203 distinct genes (70% of annotated total) and the number of polyadenylation sites detected decreased to 3,108 distinct genes (30%). This is consistent with the observation above showing a lower proportion of *T. cruzi* poly(T)-containing reads compared to SL-containing reads and can be explained, at in least in part,

by documented library construction artifacts causing under-representation of poly(A) segments. We report the coordinates of all putative SL-addition and polyadenylation sites detected in this study in relation to the existing *T. cruzi* gene structure annotation in TriTrypDB (**Tables S2, S3, see appendix**).

Using the coordinates of the SL-addition sites and existing start codon annotations for coding sequences (CDS), we defined the boundaries of all 5′ UTRs in the *T. cruzi* genome. The median 5′ UTR length was 68 nt with a range from 1 nt, right abutting the initiation codon, to 2029 nt (**Figure 11A**).The 39-nt SL sequence was not included in the 5′ UTR length calculation. The distribution of the 5′ UTR lengths was similar across different life stages of the parasite (**Figure 12**). A similar analysis using poly(A) addition sites and existing stop codon annotations allowed us to determine the boundaries of the 3′ UTRs. The 3' UTR length distribution was tighter than observed for the 5' UTRs, with a median length of 197 and a range from 1 to 4362 nt (**Figures 11B, 13**). A careful examination of the annotation of genes in the upper and lower quartiles of both 5' and 3' UTR length distributions revealed no enrichment in any particular functional category (**Table S4**). The median mRNA length was 1402 nt, of which the 5′ UTR, CDS, and 3′ UTR accounted for 5%, 81%, and 14%, respectively. No significant correlation was noted between CDS length and either UTR length or between corresponding UTR lengths (**Figure 14**). With the gene boundaries accurately mapped for the first time in *T. cruzi*, we were able to define the intergenic regions. The median value for intergenic spacing was 166 nt (**Figure 11G**) and the area under the curve indicated that more than 65% of the *trans*-

splicing sites were located within the range of 100 to 200nt downstream of the polyadenylation site. This may point to an optimal distance for the concerted action of the *trans*-splicing and polyadenylation machineries (LeBowitz, Smith et al. 1993; Matthews, Tschudi et al. 1994).

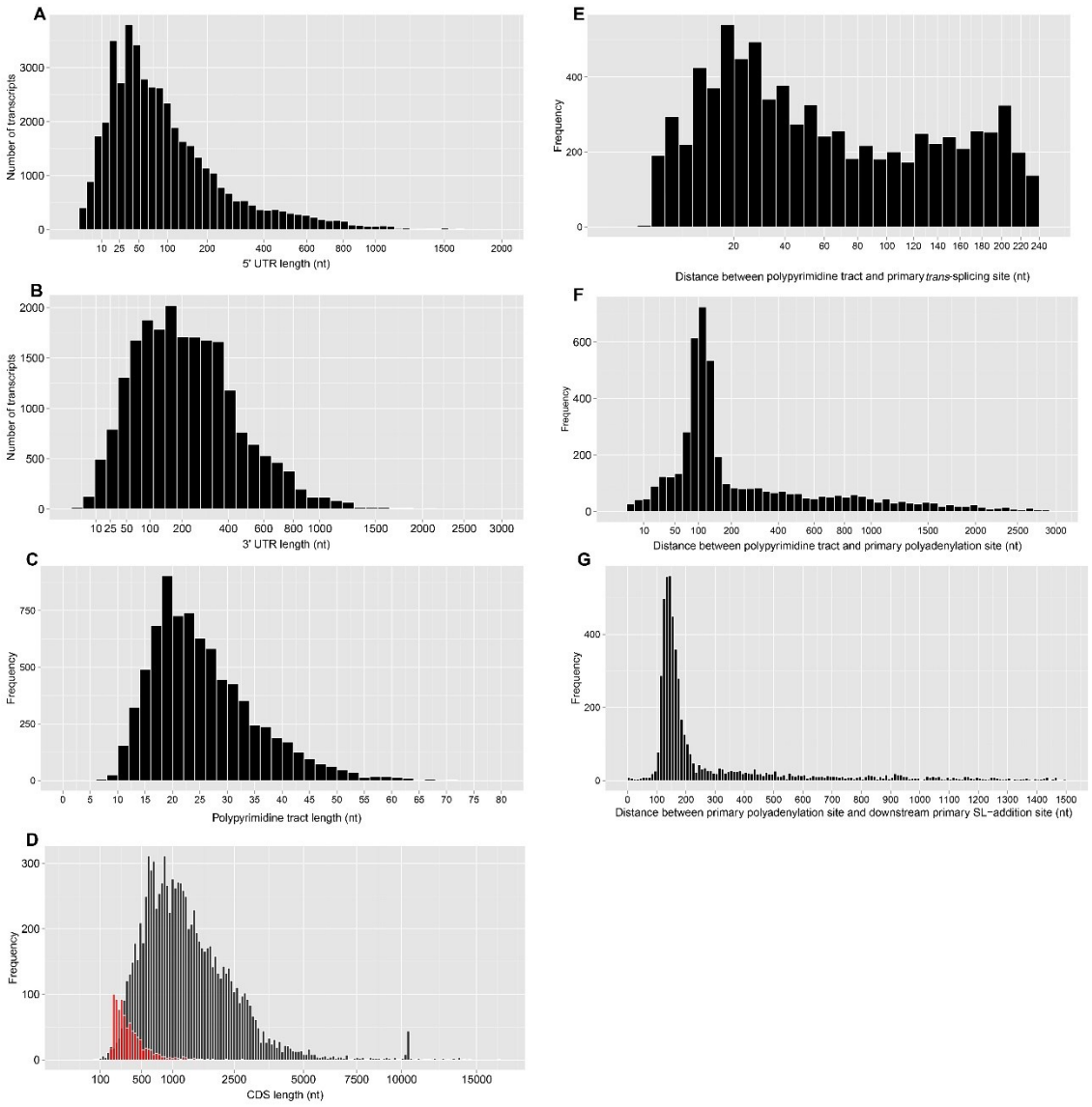

Figure 11 Length and position distribution of various gene structure components in *T. cruzi*.

(A) Distribution of the lengths of 5′ UTRs. A Spliced Leader (SL) analysis of reads obtained from all developmental stages was performed to map the exact *trans*-splicing site associated with each CDS and to identify the coordinates and lengths of the 5′ UTRs. Alternative splicing events were included in the analysis and 5′ UTRs containing newly identified ORFs were not. Two genes with 5′ UTR

length greater than 2 kbs were not included in the plot. (B).Distribution of the lengths of 3′ UTRs. A polyadenylation site analysis of sequences from all developmental stages was performed to map the poly(A) addition site associated with each CDS and to identify the coordinates and lengths of the 3′ UTRs as described in Methods. Alternative polyadenylation events were included in this analysis. All 3′ UTRs containing newly identified ORFs were excluded. Three genes with 3′ UTR length greater than 3 kbases were not included in the plot. (D).Distribution of novel and annotated CDS lengths. Start and stop coordinates for coding sequences were retrieved for the *T. cruzi* CL Brener reference genome - Esmeraldo haplotype (TriTrypDB, version 4.1), and CDS lengths computed. Black indicated the distribution of CDS length computed from the current annotation. Red bars are CDS length from the novel ORFs identified in this study. Partial CDSs were excluded along with a small percentage (3.78%) of ORFs in the current annotation from TriTrypDB (v4.1) with no clear translation frames defined. A window of 250 nt upstream of the primary *trans*-splicing site was scanned to identify polypyrimidine (polyPy) tracts. The lengths of the polyPy tracts (C), distance between the 3′ end of the polyPy tract and the *trans*-splicing site (E) and the distance between the 3′ end of the polyPy and the polyadenylation site (F) were binned in length intervals and plotted. (G) Size distribution of intergenic regions. Intergenic regions were mapped between the primary polyadenylation site and the primary *trans*-splicing site of the downstream gene and lengths computed for the pooled data from all developmental stages.

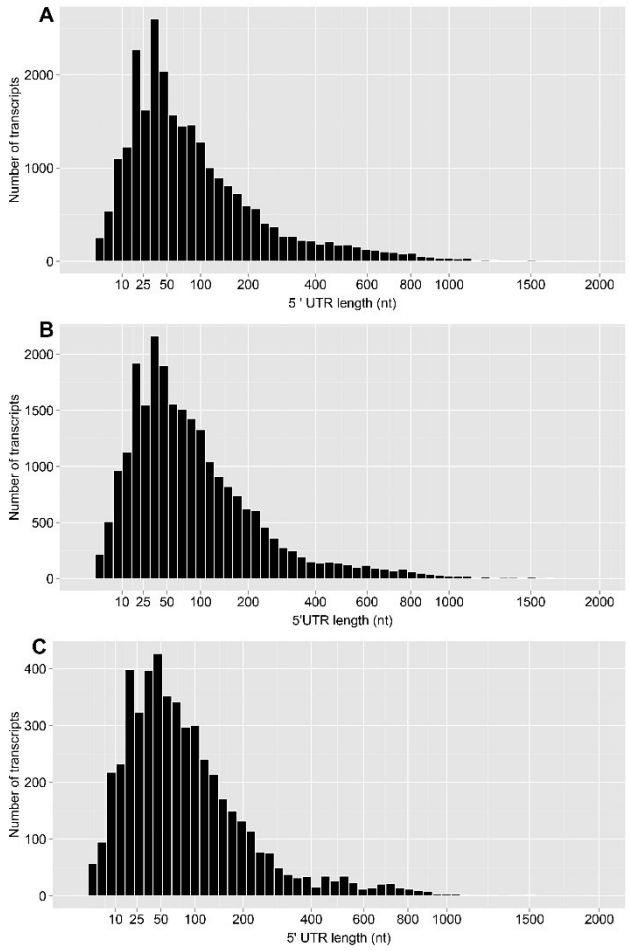

Figure 12 Distribution of the lengths of 5′ UTRs in transcript expressed in different developmental stages of *T. cruzi*.

A Spliced Leader (SL) analysis of reads obtained from each of the three developmental stages was performed to map the exact *trans*-splicing site associated with each CDS and to identify the coordinates and lengths of the 5′ UTR. Results are shown for (A) trypomastigotes, (B) epimastigotes, and (C) intracellular amastigotes (72 hpi). Alternative splicing events were included in the analysis and 5′ UTRs containing newly identified ORFs were not. A total of three genes with 5′ UTR length greater than 2 kbases were not included in the plots.

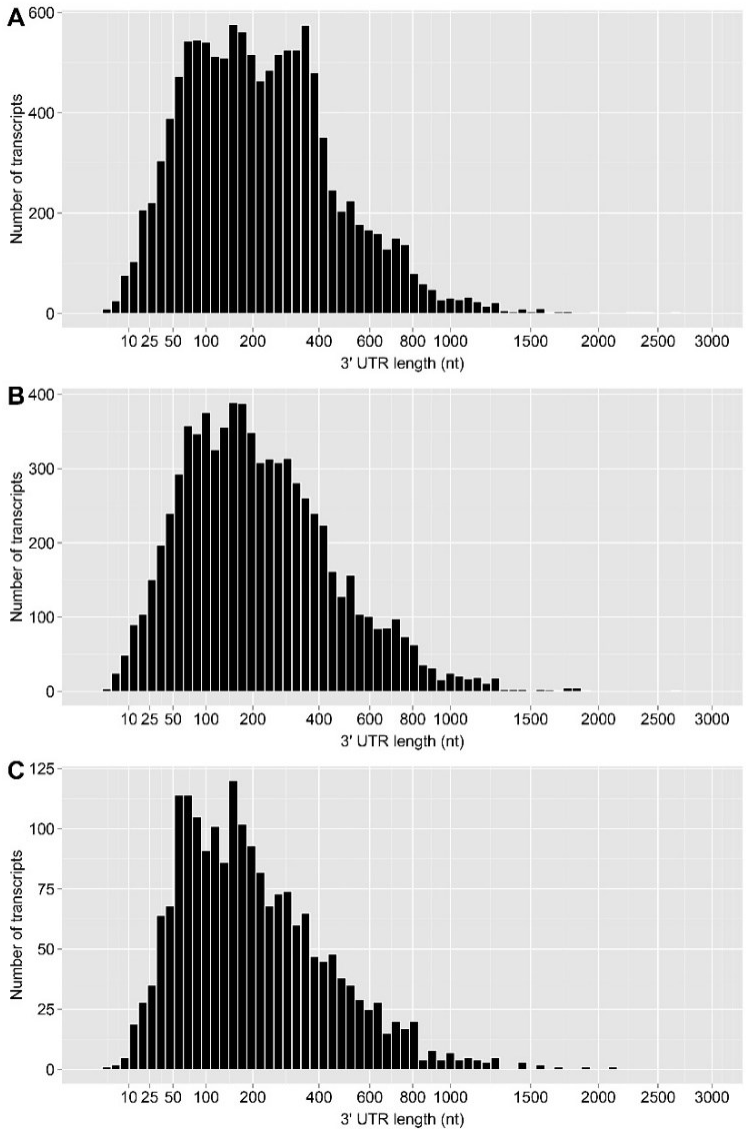

Figure 13 Distribution of the lengths of 3′ UTRs in transcripts expressed in different developmental stages of *T. cruzi*.

A polyadenylation site analysis of sequences from each of the three developmental stages was performed to map the poly(A) addition site associated with each CDS and identify the coordinates and lengths of the 3′ UTR coordinates as described in Methods. Results are shown for (A)

Trypomastigotes, (B) Epimastigotes, and (C) amasitgotes 72hpi. Alternative polyadenylation events were included in this analysis. All 3′ UTRs containing newly identified ORFs were excluded. A total of three genes with 3′ UTR length greater than 3 kb were not included in the plot.

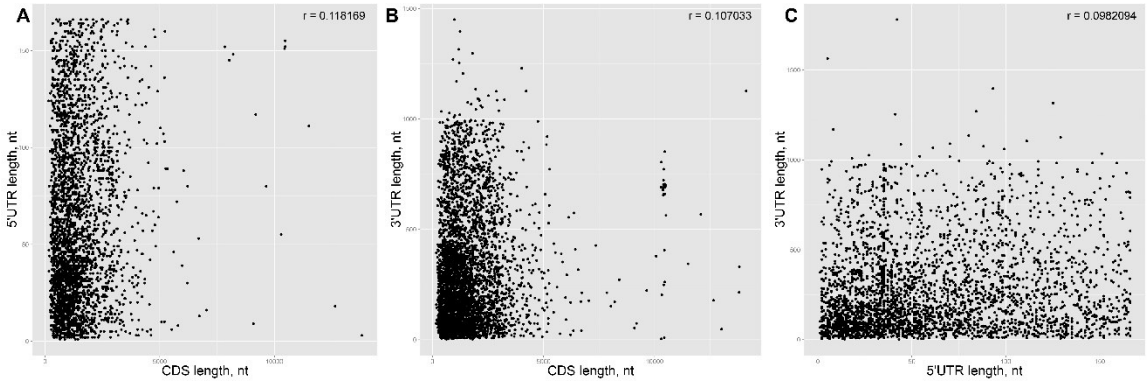

Figure 14 Correlations between CDS length and either UTR length or between corresponding UTR lengths.

(A) The correlation between the length of 5′ UTR and CDSs. (B) The correlation between the length of 3′ UTR and CDSs. (C) The correlation between the length of 5′ UTR and 3′ UTR. The value of Pearson correlation (r) was labeled in the upper right corner of each panel.

Table 3 Characterization of different *T. cruzi* gene structure components.

Feature	Median	Mean	75% percentile	Min	Max
5' UTR	68	131	152	1	2029
PolyPy tract	23	25	30	3	93
Distance between TS and polyPy	47	75	127	1	242
3' UTR	197	263	355	1	4362
Distance between poly(A) and polyPy	127	206	430	1	5887
CDS	1137	1503	1884	69	16764
Distance between downstream splice addition site and upstream poly(A)	166	240	409	1	5840
CDS in 3-way COG	1059	1476	1848	69	13746

We extended our analysis to include the examination of features known to be involved in the regulation of RNA processing events. Polypyrimidine (polyPy) tracts are located upstream of the 3' splice acceptor sites and provide a required signal for the *trans*-splicing machinery (Huang and Van der Ploeg 1991; Siegel, Tan et al. 2005; Gunzl 2010). We searched for the longest stretch of pyrimidine residues, interrupted by no more than two contiguous purines and located upstream of each of the primary *trans*-splicing sites. PolyPy tracts ranged from 3 to 93 nt in length, with a median value of 23 nt (**Figure 11C**) and a clear usage preference for thymine (72.47 %) over cytosine (23.7%) residues. This observation is consistent with the findings in *T. brucei* and *T. vivax* (Kolev, Franklin et al. 2010; Greif, de Leon et al. 2013) and our logo analysis shown in **Figure 18F**. The distance between the polyPy tract and the upstream primary polyadenylation site exhibited a relatively tight distribution with a median value of 127 nt (**Figure 11F**), which was twice the median distance (47 nt) between polyPy tract and the downstream primary SL-addition site (**Figure 11E**). This spacing of the polyPy tract about two thirds of intergenic regions can also be noted in *T. brucei* (Kolev, Franklin et al. 2010). The investigation of extreme values in each genomic components identified in this analysis did not reveal an obvious function or gene copy bias (**Table S4, see appendix**).

With the canonical features for the *T. cruzi* genes resolved at the single nucleotide level, we compared them to their orthologous elements reported in *T. brucei* (Kolev, Franklin et al. 2010). A striking aspect we observed was the

shorter median length of the 5′ and 3′ UTR regions in *T. cruzi* genes when compared to *T. brucei* (**Figure 15**). To account for possible biases from species-specific multigene families, we further restricted our analysis to the subset of syntenic three-way clusters of orthologous genes (COGs) in the reference trypanosomatid (TriTryp) genomes (El-Sayed, Myler et al. 2005) and obtained similar results (**Figure 16, Table 3**). This relative compaction of the *T. cruzi* UTRs is congruous with earlier observations we made at the level of coding sequences, whereby the mean length of CDS regions in *T. cruzi* was markedly shorter than in *T. brucei* (Kolev, Franklin et al. 2010). We have updated here our computation of mean CDS lengths in *T. cruzi* using a more recent version of the genome annotation and report the results in **Figure 11D**. Despite the fact that *T. cruzi* has a notably smaller genes (CDS + UTRs), our characterization here of the true intergenic region (and not the inter-CDS region) reveals longer intergenic regions in *T. cruzi* when compared to *T. brucei*. We infer that the higher gene density in the *T. cruzi* genome is directly the result of shorter UTR lengths.
We noted that the GC content of CDS is the highest (53.02%), followed by 5′ UTR (45.83%) and intergenic regions (42.44%), with the lowest level detected in 3′ UTR (41.26%) (**Table 4**).

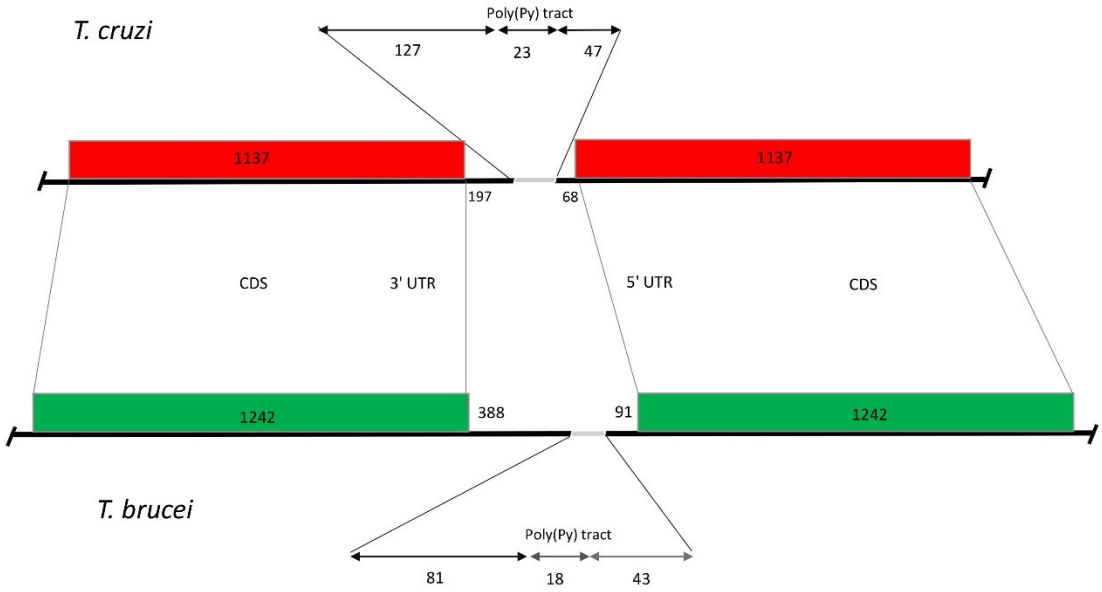

Figure 15 Comparison of gene structure components between *T. cruzi* and *T. brucei*.

The median length of each gene structure components of *T. cruzi* were compared with what has been reported in *T. brucei* (Kolev, Franklin et al. 2010). The median value of each component was labeled.

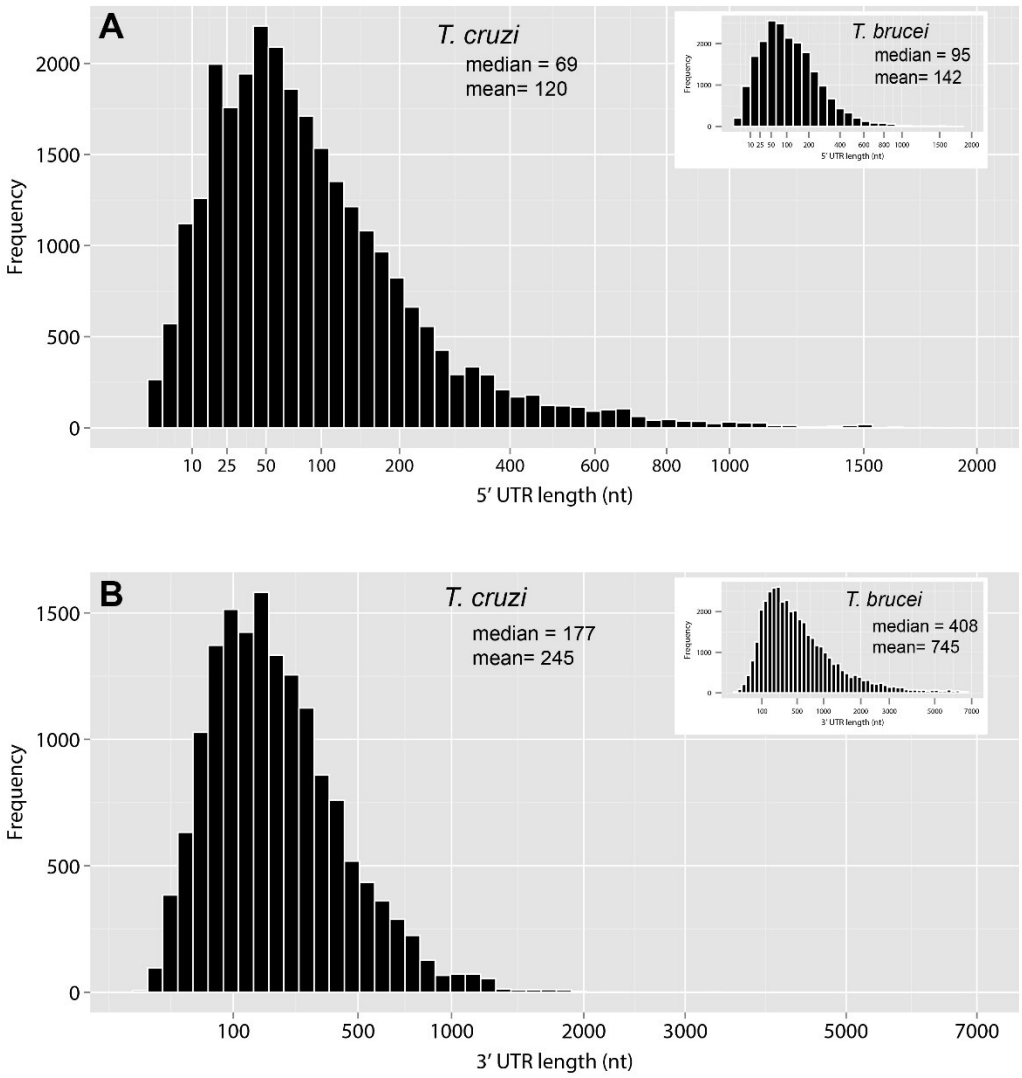

Figure 16 Distribution of UTR lengths for *T. cruzi* and *T. brucei* genes in 3-way COGs.

(A)The length of 5' UTRs for genes belonging to the three-way COGs were included in the analysis. In comparison, the lengths of 5' UTRs from *T. brucei* transcripts (Kolev, Franklin et al. 2010) that also belong to the three-way COGs were plotted in the inset. (B) The length distribution of 3' UTRs for *T. cruzi* and *T. brucei* (inset) genes belonging to the three-way COGs.

Table 4 GC content of different gene structure components in *T. cruzi*.

Nucleotide	5′ UTR	CDS	3′ UTR	Intergenic region	Genome
G	26.97%	29.29%	23.71%	23.24%	25.19%
C	18.86%	23.73%	17.55%	19.20%	25.19%
A	29.61%	23.55%	26.53%	20.75%	24.77%
T	24.56%	23.43%	32.10%	34.60%	24.85%
GC content	**45.83%**	**53.02%**	**41.26%**	**42.44%**	**50.38%**

Novel ORFs and revised ORF boundaries

The transcriptional evidence dataset we have generated here provided us with an opportunity to revisit our gene calls. We identified 834 novel ORFs present in *trans*-spliced and polyadenylated transcripts and not originating within previously annotated CDSs (**Tables S5, S6, and S7**). The median size of the newly identified ORFs was small (294 nt, **Figure 11D**) and expected since small ORFs with no evidence of expression were systematically excluded during the early gene predictions (El-Sayed, Myler et al. 2005). We examined the novel ORFs for sequence homology in RefSeq and the Conserved Domain Databases, and annotated them for features such as predicted signal peptides, transmembrane domains and GPI anchors. The novel ORFs were not particularly enriched for any of the attributes examined: 7.4% contained a sequence with homology to a known protein domain/superfamily, 9.7% were predicted to contain a signal peptide, 4.7% may represent GPI-anchored proteins and 36.1% contained putative transmembrane domains (**Table S7**). Of the 399 novel ORFs with database matches, a great majority (89%) had high levels of identity (E-value $<10^{-5}$) to *T. cruzi* hypothetical proteins.

We also conducted a genome-wide search across all 5′ UTRs to identify the potential for extended CDS regions. We detected 5,480 putative alternative in-frame start codons upstream of the existing start codon coordinate (**Table S8**). Validation of the start site(s) used for translation initiation will likely emerge soon, as ribosome profiling studies unravel (Vasquez, Hon et al. 2014).

Alternative RNA processing sites

The sequencing depth of our transcriptome profiling experiments allowed not only the identification of the SL-addition and polyadenylation sites at a single-base resolution, but also the characterization of widespread alternative RNA processing events in *T. cruzi*. Because our study design was mainly aimed at a quantitatively unbiased profiling of the transcriptome of the parasite at various developmental stages, we did not enrich for SL- or polyA-containing reads during library construction and relied on deep coverage to collect relatively large numbers of reads from both ends of transcripts (6.4 million SL- and 1.1 million polyA-containing reads). This approach permitted us to identify as well as quantitate differential RNA processing events (**Figure 17**).

Of the 7,869 genes with SL-addition sites detected, 88% used more than one trans-splicing site in at least one developmental stage (42% used two to four trans-splicing sites and 46% had greater than 5 sites). This observation is similar to what has been reported for *T. brucei* where 89% of genes showed evidence of alternative trans-splicing events (Kolev, Franklin et al. 2010). An examination of the *trans*-splicing sites revealed a propensity for usage of the canonical acceptor sequence (AG) both at the primary (95%) and minor (33%) splicing sites (**Figures 18A, 18B, 19; Table S9, see appendix**). The distribution of the distances between the primary and minor trans-splicing sites using the AG

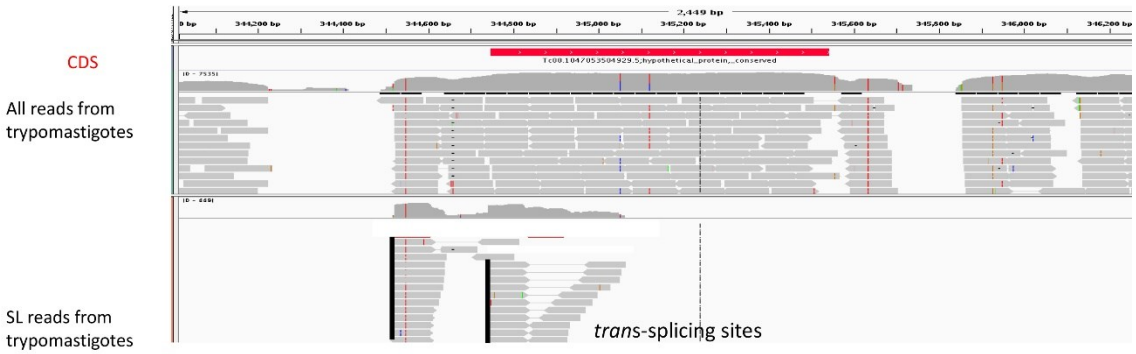

Figure 17 An example of alternative *trans*-splicing events.

The mapping coverage from all reads and SL-containing reads of gene Tc00.1047053504929.5 are visualized as tracks on Integrative Genomics Viewer (IGV). The red region indicates the CDS region. Two distinctive *trans*-splicing sites were detected for this particular gene.

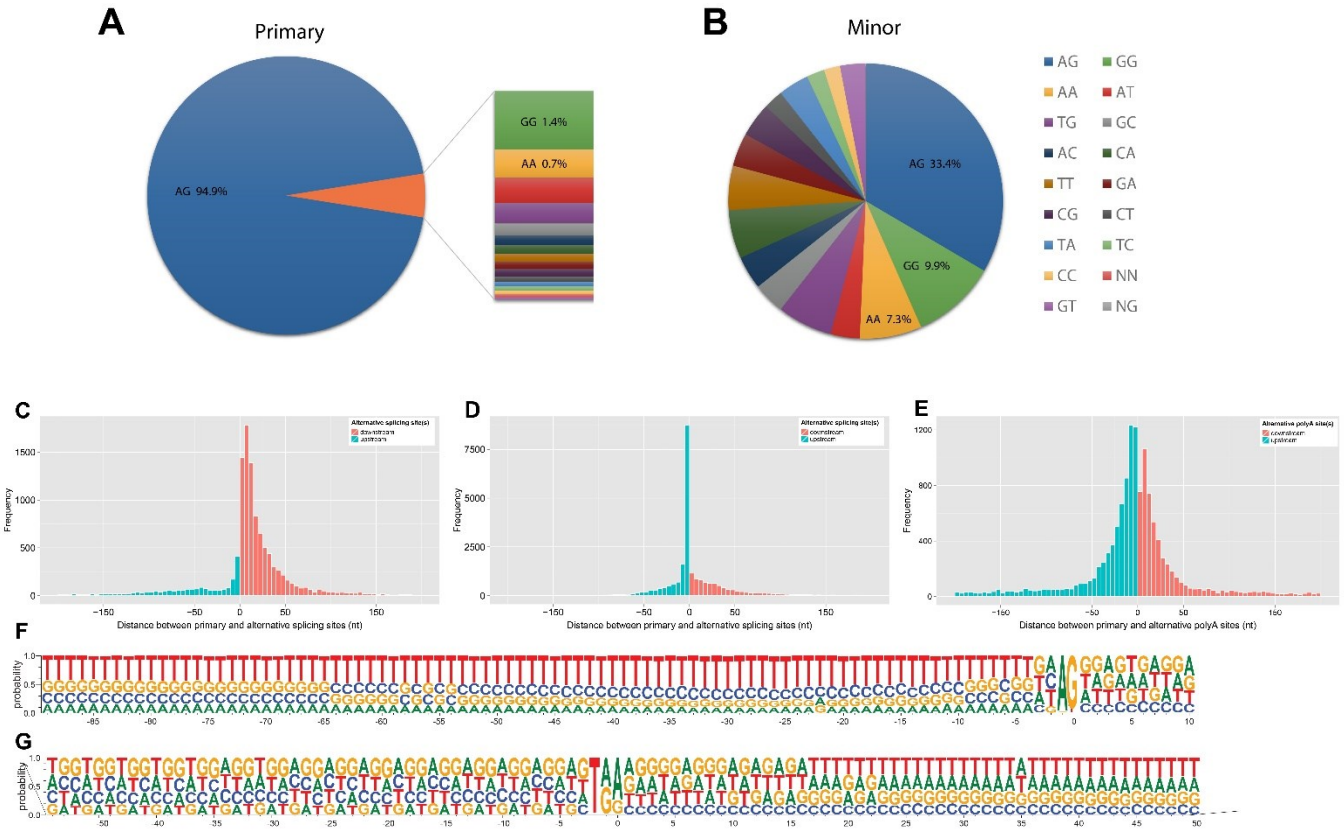

Figure 18 Sequence composition at RNA processing sites and distance between primary and alternative RNA processing sites.

Spliced leader acceptor sites were retrieved by extracting the dinucleotide upstream of each detected *trans*-splicing site. The frequency of dinucleotide usage was determined for primary (A) and alternative (B) acceptor sites. The sequence composition of the region encompassing 90 nt upstream and 10 nt downstream of the *trans*-splicing sites was plotted for the primary (F) and alternative (**Figure 19**) splice sites using WebLogo (Crooks, Hon et al. 2004). Distribution of distances between primary and alternative SL *trans*-splicing sites

for splicing sites applying AG (C) and non-AG (D). Alternative SL *trans*-splicing events were identified for each of the developmental stages and pooled for this analysis. The distance between the primary *trans*-splicing site and alternative *trans*-splicing sites is positive in value when the alternative *trans*-splicing site(s) are located upstream of the primary site and negative when the alternative *trans*-splicing site(s) are located downstream of the primary site. (E) Distribution of the distances between primary and alternative polyadenylation sites. Alternative polyadenylation events were identified for each of the developmental stages and pooled together for this analysis. The distance between the primary and the alternative poly(A) addition site(s) is positive in value when the alternative poly(A) addition site(s) was located upstream of the primary site and negative when the alternative poly(A) addition site(s) was located downstream of the primary site. A similar analysis was performed with transcripts derived from trypomastigotes, epimastigotes, and amastigotes 72hpi. The sequence composition of the region encompassing 50 nt upstream and 50 nt downstream of the primary poly(A) addition site was plotted for the primary (G) and alternative (**Figure 20**) poly(A) addition sites using WebLogo (Crooks, Hon et al. 2004).

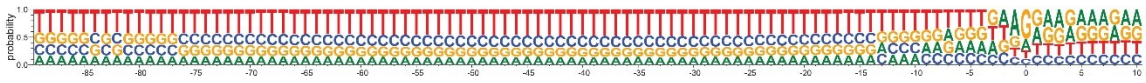

Figure 19 Analysis of sequence composition at alternative spliced leader acceptor sites.

Spliced leader acceptor sites were retrieved by extracting the dinucleotide upstream of each detected *trans*-splicing site. The sequence composition of the region encompassing 90 nt upstream and 10 nt downstream of the alternative *trans*-splicing sites was plotted using WebLogo (Crooks, Hon et al. 2004).

Figure 20 Analysis of sequence composition near the minor polyadenylation sites.

The sequence composition of the region encompassing 50 nt upstream and 50 nt downstream of the minor poly(A) addition site was using WebLogo (Crooks, Hon et al. 2004).

acceptor sequence revealed that a great majority (84%) of the alternative splice sites are located downstream of the primary site (**Figure 18C**). This observation is consistent with a model that proposes that the 3' splice site in mammalian introns is located by a scanning process that recognizes the first AG downstream of the branch point in a sequence-specific context (Mount 1983; Smith, Porro et al. 1989). In fact, we note a significant effect based on the nucleotide preceding the AG, whereby CAG and UAG account for 50% at the primary splice site in contrast to 37% at the downstream alternative sites (37%), further supporting a previous model that proposes that the combination of proximity and competition influences the selection of the primary splice site (**Table 5**) (Mount 1983).

The usage of non-canonical acceptor sequences was much more pronounced in the alternative (67%) than in the primary (5%) splice sites (**Figures 18A, 18B**). Unlike for the canonical acceptor sequences, the distribution of the distances between the primary and the alternative non-AG splicing sites was heterogeneous and showed no distinct pattern, with the exception of a peak of alternative *trans*-splicing events that occurred within the 10 nt that precede the primary site (**Figure 18D**). A similar peak can be observed for AG splicing sites, albeit less pronounced. Those peaks may reflect sloppy splicing events near the primary SL-addition site as has been suggested previously for *T. brucei* (Kolev, Franklin et al. 2010; Siegel, Hekstra et al. 2010) and/or slippage due to the presence of tandem NAGNAG acceptors. We also noted a slight preference for GG and AA among both the canonical and non-canonical acceptor sequences.

We are unable to rule out polymorphisms between the strain used in this study (strain Y) and the reference genome (CL Brener) as a source, at least in part, of this bias as well as some of the non-canonical sites we observe. A sequence composition analysis of the region immediately upstream of the SL-addition events allowed us to visualize the intervening polypyrimidine tract between the branch point and the AG 3' splice site (**Figure 18F, 18G**).

Table 5 Usage frequency of nucleotide proceeding splicing acceptor (AG) site.

Tri-nucleotide	Primary	Alternative downstream	Alternative upstream
C(AG)	21%	20%	20%
U(AG)	29%	17%	19%
A(AG)	31%	42%	22%
G(AG)	12%	21%	39%

A significant proportion (28%; 16,911) of alternative SL-addition sites we have identified fall within the annotated CDS regions, reflecting two possible scenarios: misannotated initiation start sites or events that have the potential to impact the resulting protein product(s) (**Table S10, see appendix**). An example of the latter is the gene encoding LYT1 (Tc00.1047053503829.50), for which we detected two alternative SL-addition sites at positions -46 and +10 relative to the start codon (+1) (**Figure 25A**). *T. cruzi LYT1* has been reported to generate two protein products, a process mediated through stage-regulated alternative *trans*-splicing events (Manning-Cela, Gonzalez et al. 2002). The shorter product (28 amino acid truncation) localizes to the mitochondrial kinetoflagellar zone, whereas the longer product localizes on the plasma membrane (Benabdellah, Gonzalez-Rey et al. 2007).

Of the 6,311 genes for which polyadenylation sites were identified, 3,988 use an alternative polyadenylation site in one or more developmental stages, 62% of which used two to four polyadenylation sites and 38% used more than four sites. The heterogeneity observed is similar to what we have calculated for *T. brucei* where 92% of genes display evidence for alternative polyadenylation events (Kolev, Franklin et al. 2010). The distribution of the distances between primary and alternative polyadenylation sites revealed that ~75% of the minor sites were located within a 200 nt window centered around the primary site, whereas 27% were located within a 20 nt window (**Figures 18E, 21**). This abundant

heterogeneity of polyadenylation sites at several closely spaced positions has been observed not only in *T. brucei* (Matthews, Tschudi et al. 1994; Benz, Nilsson et al. 2005; Kolev, Franklin et al. 2010) but also in other systems, including plants, animals and human (Elkon, Ugalde et al. 2013; Ji, Guan et al. 2014). While no consensus motif such as the AAUAAA required for cleavage and polyadenylation in higher eukaryotes was present upstream of the polyadenylation site, we noted a strong U(A/G)(A/G) motif abutting the Poly(A)+ addition site for the primary and alternative polyadenylation sites alike (**Figures 18G , 20**).

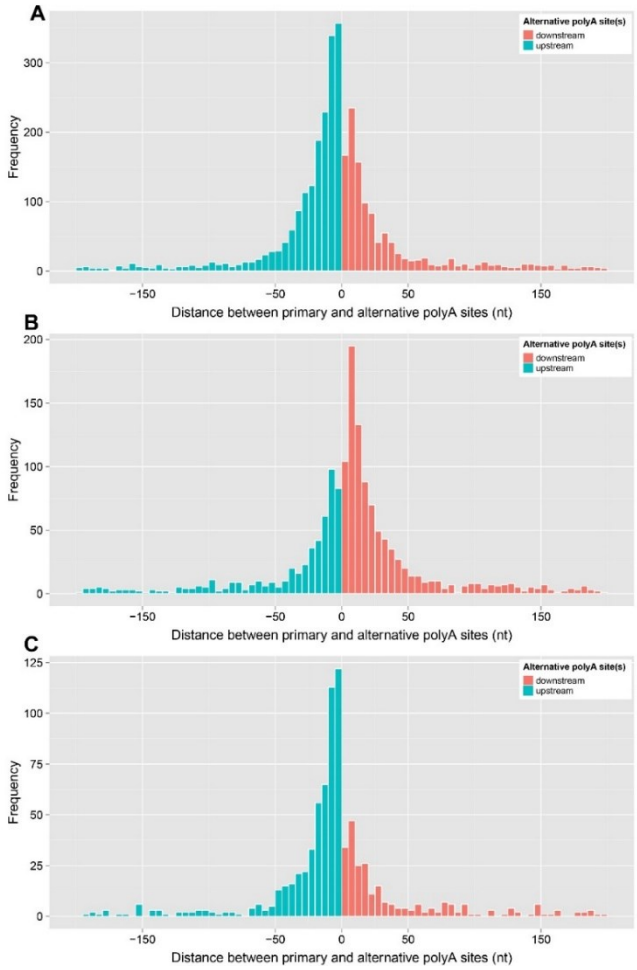

Figure 21 Distribution of distances between primary and alternative polyadenylation sites in transcripts from different developmental stages of *T. cruzi*.

Alternative polyadenylation events were identified for each of the developmental stages. The distance between the primary polyadenylation site and alternative polyadenylation sites is positive in value when the alternative polyadenylation site(s) was located upstream of the primary site and negative when the alternative polyadenylation site(s) was located downstream of the primary site. Results are shown for (A) trypomastigotes, (B) epimastigotes, and (C) intracellular amastigotes (72 hpi).

Alternative *trans*-splicing events across parasite development

The heterogeneity of RNA processing that we detected at high frequency in the form of alternative *trans*-splicing and polyadenylation events may be the manifestation of another level of gene expression regulation. To examine the role and dynamics of RNA processing events at different stages of the development of the parasite, we examined the usage of alternative SL-addition sites in the three stages where our coverage of the transcriptome was deep. Those included reads from the trypomastigote, epimastigote and amastigote 72 hpi samples, which in total, constituted 78% of the SL-containing reads. The remaining 22% included the combined reads from amastigotes at several time points in the range of 4-72 hpi.

In order to identify a subset of genes that switch the use of their primary *trans*-splicing site across different stages, we calculated the ratio of number of reads mapping to the primary site over the number of reads mapping to the secondary site (P/S) and plotted that ratio against the number of reads mapped to the primary site. The P/S ratio provided a measure of the preferential usage of the primary site, with increasing values denoting dominance of the primary site, and the number of reads mapping to the primary site was used as a surrogate for expression level (and indirectly confidence in the significance of the ratios) (**Figures 22A, 23**). This approach allowed us to visualize potentially interesting subsets of genes with high expression levels in trypomastigotes and a dominance of one primary site (generally localizing in quadrant IV), and further

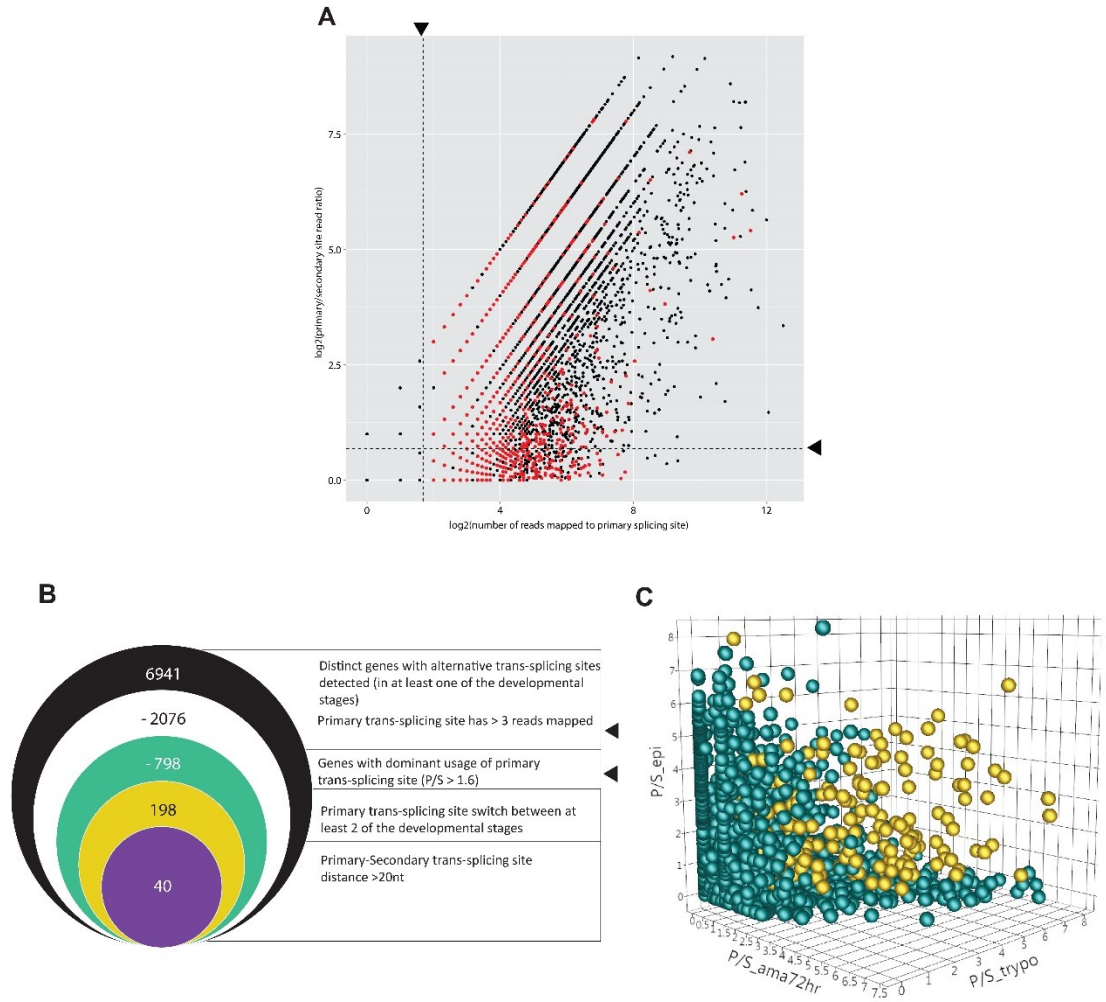

Figure 22 Alternative splicing profiles of *T. cruzi* trypomastigote, amastigote and epimastigote stages.

(A) Analysis of the usage of primary and secondary *trans*-splicing sites in trypomastigotes in context of expression across developmental stages. For each gene where alternative splicing events were identified, the ratio of SL-containing reads mapping to the primary site over the secondary site (P/S) was plotted against the number of SL-containing reads mapping to the primary site (P) in the trypomastigote stage (shown here) and in the epimastigote and amastigote (72 hpi) stages (**Figure 23**). Pink dots represent genes that display a primary site

switch between the reference stage shown and at least one of the other 2 stages. (B) Selection of genes with primary *trans*-splicing sites switching across developmental stages with coverage, primary site usage, and distance between alternative sites successive constraints. (C) Switching of primary spliced leader addition site across developmental stages. The ratio of SL-containing reads mapping to the primary site over the secondary site (P/S) was plotted for genes that switched primary site between any two of the three developmental stages and with more than three reads mapped to the primary sites in all three developmental stages. Red dots represent genes with P/S ratio greater than 1.6 at all three stages.

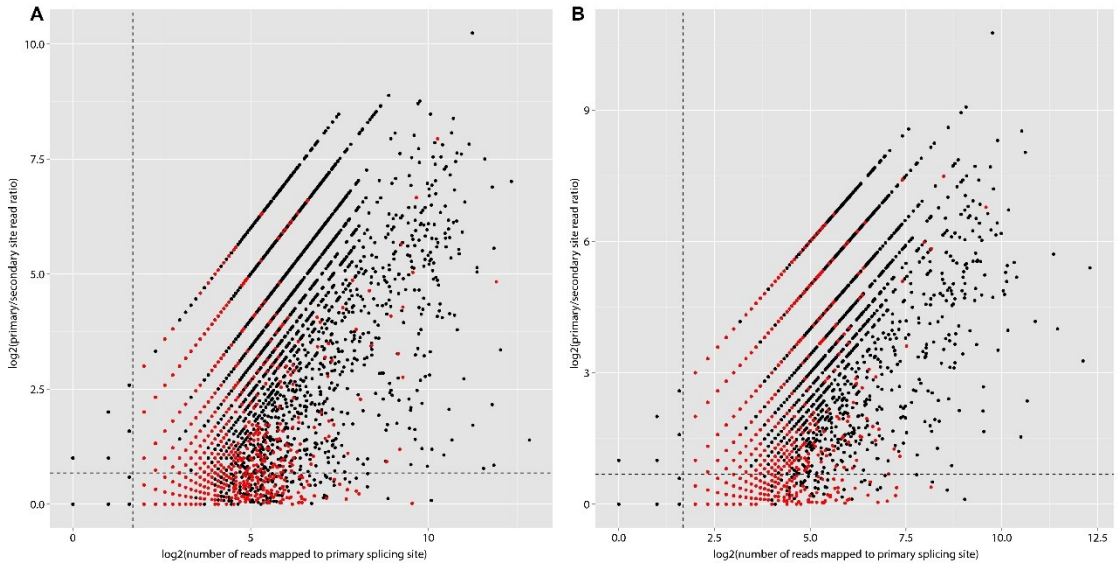

Figure 23 Analysis of the usage of primary and secondary trans-splicing sites in epimastigotes and amastigotes (72 hpi) in context of expression across developmental stages.

For each gene where alternative splicing events were identified, the ratio of SL-containing reads mapping to the primary site over the secondary site (P/S) was plotted against the number of SL-containing reads mapping to the primary site normalized by the total number of SL-containing reads at that stage (P) for (A) epimastigote and (B) amastigote (72 hpi) stages. Red dots represent genes that display a primary site switch between the reference stage shown and at least one of the other 2 stages.

highlight genes that underwent primary site switching in the epimastigote and amastigote (72 hpi) stages (labeled in red in **Figure 22A**). We repeated this analysis using each of the two other stages (epimastigotes and 72 hpi amastigotes) (**Figure 23**).

We further refined the list by carrying out sub-selections based on multiple criteria. We first reduced the 6,941 distinct genes which display alternative *trans*-splicing sites in at least one of the developmental stages to 4,874 by retaining genes that had a minimum of 4 reads mapping to the primary *trans*-splicing site in each of the three stages and therefore relatively weaker evidence for switching events (**Figure 22B**). Limiting our subset to genes with a P/S ratio > 1.6 in all three life stages further reduced the number to 4,076 genes. For genes that underwent primary site switching across developmental stages, we further investigated the dominance of the primary sites by visualizing the P/S ratio in a three dimensional space with each axis representing one of the three stages (**Figure 22C**). This illustration identifies a set of 198 genes (yellow spheres) that switched SL-addition sites between any two life stages, and highlights a subset locating in the outer periphery because of the high P/S value. Of those 198 genes that switched SL-addition sites between any two life stages, 40 were separated by a distance greater than 20 nt (**Figure 22B and Tables S11, S12, see appendix**). Among this subset of 40 genes, 27 and 11 displayed alternative SL-addition events separated by a distance greater than 50 nt and 200 nt, respectively. Such examples provide strong evidence of alternative *trans*-splicing

events that seem to be under tight developmental regulation and not the result of a sloppy splicing machinery (**Tables S11, S13**). Although a Gene Set Enrichment Analysis (GSEA) revealed no significant overrepresentation of gene functions in genes with stage-regulated alternative splicing of the SL, we did note some striking examples of alternative splicing events across stages (**Figure 24, 25**). For example, we noted *T. cruzi* gene Tc00.1047053511545.130, currently annotated as a conserved hypothetical protein, exclusively *trans*-spliced at a site 220 nt away from the start of the CDS at trypomastigote and amastigote stages but applied another site located approximately 150 nt away at epimastigote stage. Interestingly, this gene was significantly differentially regulated at epimastigote stage, compared to the other two developmental stages. Although the detailed mechanisms behind post transcriptional regulation were still not clear, the preference of different primary *trans*-splicing sites across developmental stages, resulting in the inclusion and exclusion of certain sequence elements, may influence the stability or translational efficiency of transcripts and present as a potential approach in the regulation of transcripts

We performed similar analyses aimed at characterizing alternative polyadenylation events across the three developmental stages. We were unable to detect clear events of stage-regulated alternative polyadenylation. This may be due to a combination of the extensive heterogeneity of polyadenylation sites and the relative low coverage of poly(A)-containing reads retrieved that could be mapped to unique sites.

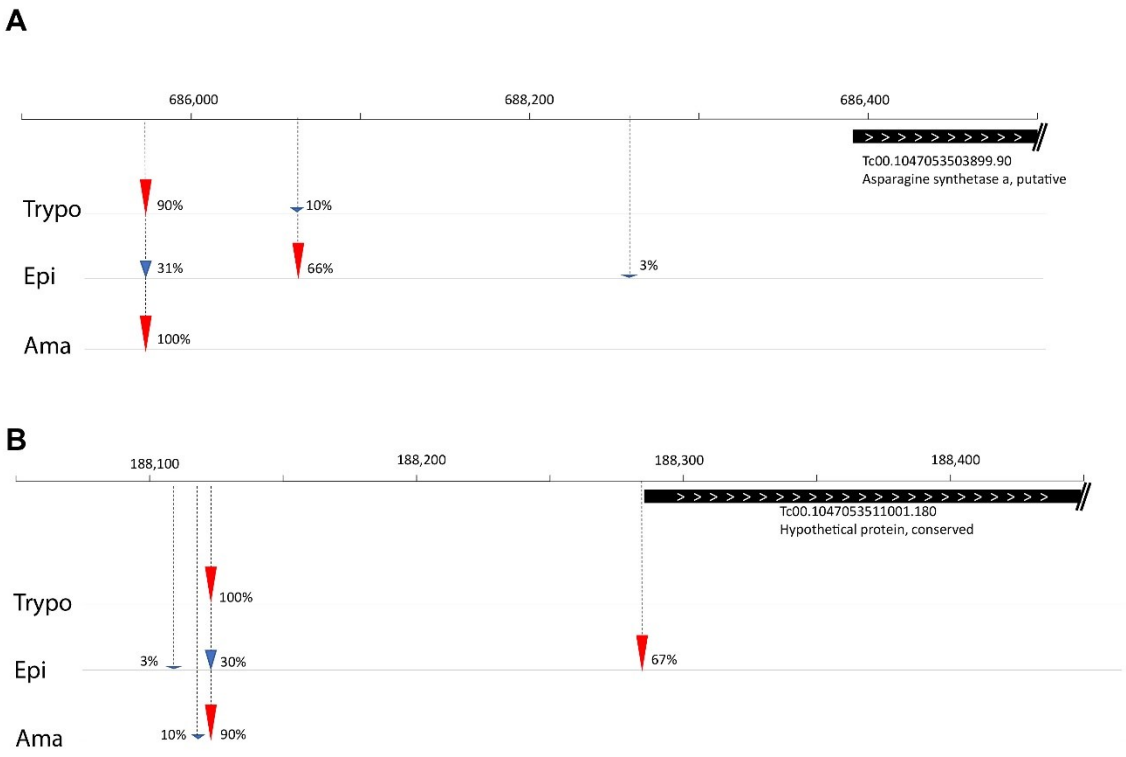

Figure 24 Examples of alternative *trans*-splicing events across developmental stages.

Usage of *trans*-splicing sites in the trypomastigote, epimastigote and amastigote (72 hpi) stages is shown for (A) Tc00.1047053503899.90 (asparagine synthetase a, putative) and (B) Tc00.1047053511001.180 (hypothetical protein, conserved). Primary sites are depicted as red arrowheads. Percentages indicate site usage frequency.

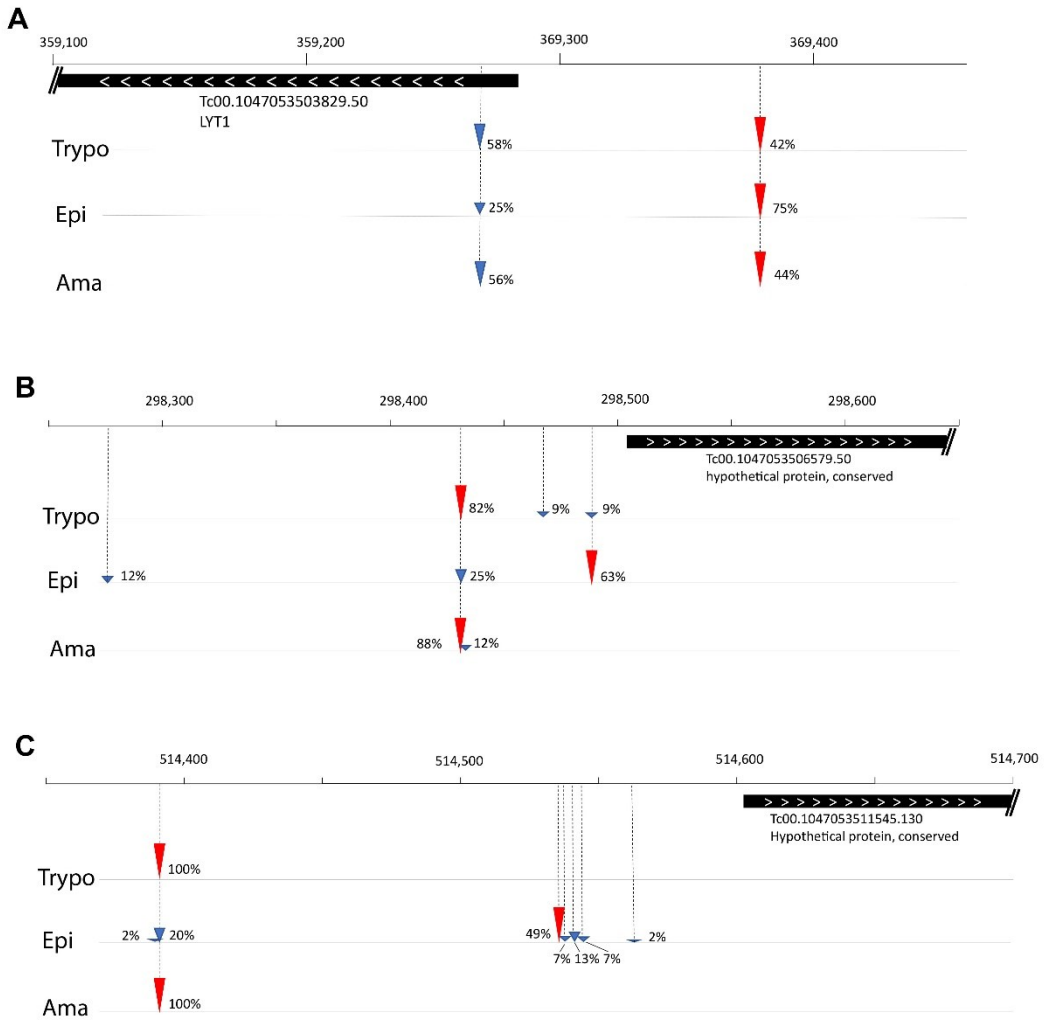

Figure 25 Examples of alternative *trans*-splicing events across different developmental stages.

Usage of *trans*-splicing sites in the trypomastigote, epimastigote and amastigote (72 hpi) stages is shown for (A) Tc00.1047053503829.50 (LYT1), (B) Tc00.1047053506579.50 (hypothetical protein, conserved), and (C) Tc00.1047053511545.130 (hypothetical protein, conserved). Primary sites are depicted as red arrowheads. Percentages indicate site usage frequency.

2.4 Conclusion and discussion

In this study we performed deep transcriptomic surveys at eight time points in *T. cruzi* involving both intracellular and extracellular life stages. Through the use of ultra high-throughput sequencing technology, we mapped in the detail the transcriptome structure throughout its development. Unlike higher eukaryotes, trypanosomes do not have classic promoter regions and genes with unrelated functions are clustered and transcribed as long polycistronic units. Most of the gene expression are regulated subtly at the post-transcriptional level through mRNA degradation, mRNA stability and translational efficiency. However, the mechanisms and regulatory elements involved in these processes are still unclear, but in many other eukaryotes, the 5' and 3' UTRs contain sequence elements that are responsible for regulating RNA or protein synthesis. So characterizing the RNA processing events in *T. cruzi* is of great significance in the efforts of understanding the biology of this pathogen. Genome-wide investigations of gene structure have been carried out in other trypanosomes but not in *T. cruzi* (Kolev, Franklin et al. 2010; Greif, de Leon et al. 2013). In this study, for the first time, we systematically defined the boundaries of UTRs and other gene structure elements for this parasite at the genomic scale. Our work not only provides fundamental insights into the organization of the *T. cruzi* transcriptome, but also a framework for the integration of additional physiological and functional datasets, as well as for a systematic investigation of regulatory elements.

By generating a highly resolved and calibrated transcriptome map of the pathogen, we characterized the mRNA processing events for more than 80% of the genes. A very high degree of heterogeneity of RNA processing sites were noted in *T. cruzi*. Recently, alternative *trans*-splicing has been discovered in drosophila and mammals, and in some cases has been shown to be physiologically significant (Horiuchi and Aigaki 2006). The degree of heterogeneity observed in *T. cruzi* was much higher than what was reported in *C. elegans*, but was comparable to other trypanosomes (Kolev, Franklin et al. 2010; Ramani, Calarco et al. 2011; Greif, de Leon et al. 2013). In trypanosomes, SL RNA *trans*-splicing does not necessarily contribute to proteomic diversity and implies a more relaxed requirement and less accurate splicing site compared to *cis*-splicing, since the SL exon is a non-coding sequence. Most surprising was that a large number of transcripts showed differential abundance of alternative splice variants across different developmental stages. More than 198 differentially spliced transcripts were detected between amastigote 72 hpi, epimastigote and trypomastigote supporting the idea that alternative splicing may have functional consequences for the regulation of parasite survival and development, of which the corresponding regulatory factors remain elusive. Alternative splicing patterns have been observed in *T. brucei* during development as well (Nilsson, Gunasekera et al. 2010).

It has been reported in trypanosomes that alternative *trans*-splicing is one mechanism by which dual targeting of different variants of proteins has evolved.

Alternative *trans*-splicing could lead to the expression of proteins having amino-termini of different lengths that derive from the same gene. Demonstrated by the group of Ochsenreiter et al., alternative *trans*-splicing is the approach for dual localization of trypanosomal isoleucyl-tRNA synthetase (IleRS) in *T. brucei* by creating a long and a short spliced variant. The protein product of the longer spliced variant with a presequence was targeted exclusively to mitochondria. In contrast, the shorter spliced variant is translated to a cytosol-specific isoform lacking the presequence (Rettig, Wang et al. 2012). Evidence in *T. cruzi* has also shown that, the protein encoded by *LYT1* gene is involved in haemolytic activity at acid pH and the regulation of stage development. *LYT1* gene can generate two protein products differing in the presence or absence of 28 amino acids at the end (Benabdellah, Gonzalez-Rey et al. 2007). The shorter products localizes in the mitochondrial kinetoflagellar zone, whereas the longer one is found on the plasma membrane. Our experiment captured identical alternative *trans*-splicing sites for gene *LYT1* reported by Swindle's team using PCR amplification of reverse transcribed mRNA in 2002. This further confirmed the accuracy and sensitivity of our high throughput sequencing approach. Alternative *trans*-splicing events can be an efficient mechanism for the regulation of protein compartmentalization. In addition, we postulate that different splice leader addition site at polycistronically transcribed genes can result in different stability or translational efficiency for each gene thus affecting protein abundance. Of all the genes that underwent alternative splicing events, we systematically selected 122 genes that applied different primary SL-addition site at various stages, of

which the reliability and dominance of the primary site were confirmed and the distance between primary and secondary was more than 20 nt. We propose that the preference of different primary *trans*-splicing events can be applied as a mechanism leading to the regulation of protein variant targeting to different cellular compartments.

More questions can be asked with regard to the different splicing patterns: Are the splicing events are regulated? If yes, how does the splicing machinery adapt under different developmental stages? Are the resulting variants targeted to different cellular compartments or do they possess altered stabilities? Can the inclusion of some sequence motifs on the variants lead to different translational efficiency? The identification of potential sequence motifs located in the 5' or 3' UTRs, particularly those that will be selectively included through different splicing events can be very helpful in explaining the function of alternative splicing. Our work here provided a very interesting gene list from the pathogen that may play a very importance role in the survival and adaptation strategies of the parasite and is of great significance in the efforts of understanding the biology of the pathogen.

Chapter 3: Simultaneous Interrogation of the Transcriptomes of the Human Pathogen *Trypanosoma cruzi* and its Infected Host Cell

2.1 Objective of Study

In this study, we aimed to construct the transcriptome of *T. cruzi* and infected human host cells across different time points of the infection cycle and also the transcriptome of *T. cruzi* at its extracellular stages using RNA-Seq technology. We sought to identify genes with significant regulation and to profile gene expression from both species simultaneously. The overall goal of this chapter is to capture the transcriptomic signature for both pathogen and host during the infection cycle, characterize their adaptation strategy in different environments applied by the parasite and understand the biology behind it.

2.2 Materials and Methods

2.2.1 Materials

T. cruzi (Y strain) trypomastigotes were harvested from tissue culture supernatants of infected LLcMK2 cells (ATCC®CCL-7) on Days 7-10 post infection as described (Tardieux, Webster et al. 1992). To greatly minimize contamination from amastigotes present in the culture supernatants (typically 5-30%) parasites were centrifuged at 1,000*g* for 10 min and pelleted parasites were incubated at 37ºC for 4 hours to allow trypomastigotes to swim away from the pelleted amastigotes. This procedure greatly

enriches for trypomastigotes, where 95-99% purity can be achieved as assessed by morphology. Epimastigotes were grown in liver infusion tryptose (LIT) at 27ºC and harvested at mid-log phase, when metacyclics are not present.

2.2.2 Methods

RNA isolation and library construction for simultaneous transcriptome profiling of host and pathogen.

RNA was isolated using Trizol® reagent, quality determined by an Agilent 2100 bioanalyzer and quantified by qPCR using a KAPA Biosystems library quantification kit. Standard Illumina protocols were used for mRNA-Seq sample preparation. RNA-Seq libraries were constructed from polyA-enriched mRNAs of 8 *T. cruzi* developmental stages: epimastigotes, trypomastigotes, and intracellular amastigotes at 4, 6, 12, 24, 48 and 72 hrs post infection (hpi) of HFF cells. Libraries were also constructed from uninfected HFF cells at the same time points. For each condition, 2-4 independent biological replicates were sequenced on an Illumina HiSeq 1000. A total of 2.7 billion reads from 35 samples (See Table 6, 7,and 8 for the full experimental design) were generated from 101 bp paired-ends.

Raw data pre-processing and quality trimming.

The quality of the reads was evaluated using the FastQC tool [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] and one nucleotide was trimmed using the FastX toolkit (Hannon Lab, CSHL) when the mean of the quality score fell below 30 in the last position (see Table S1).

Mapping cDNA fragments to reference genomes, abundance estimation and data normalization.

Reads were aligned to the reference human genome sequence (hg19, GRCh37) using Tophat v2.0.8 (Trapnell, Pachter et al. 2009), allowing for two mismatches per read and for mapping to more than one locus). Reads were also independently aligned to the Esmeraldo haplotype (El-Sayed, Myler et al. 2005) of the *T. cruzi* CL Brener reference genome (v.4.1) obtained from the TriTrypDB database (www.tritrypdb.org) since the *T. cruzi* Y strain belongs to the *T. cruzi* Type II clade (Cortez, Martins et al. 2012) as the Esmeraldo parent strain. Two mismatches per read were allowed, however reads were allowed to map only to a single locus.The abundance of reads mapping to each coding sequence (CDS) was determined using HTSeq [http://www-huber.embl.de/users/anders/HTSeq/].

Data quality assessment by statistical sample clustering and visualization.

Very weakly expressed genes were removed. Those were defined as having less than 1 read per million in n of the samples, where n is the size of the smallest group of replicates (Anders, McCarthy et al. 2013) (here n=2 and 3 for the *T. cruzi* and human samples, respectively). Multiple approaches were used to evaluate replicates and the relationships between samples across time points and to visualize sample-sample distances. Those included Pearson correlation and standardized median correlation analyses, box plots, Principal Component Analysis (PCA) and Euclidean distances-based hierarchical clustering. All components of our

statistical pipeline, named cbcbSEQ, can be accessed on GitHub (https://github.com/kokrah/cbcbSEQ/) and the specific R code for the analyses is included in the Supplementary material section.

We removed samples that did not pass the following quality assessment procedure: for each sample we computed the median pairwise correlation (mpc) to all other samples in the dataset. We then applied a standard outlier identification method to remove samples that have low correlation to the other samples: samples were removed if their median pairwise correlation (mpc) is less than Q1 (mpc) – 1.5 IQR (mpc) where Q1 (mpc) and IQR (mpc) are first the quartile and inter-quartile range of the median pairwise correlation across all samples respectively. Two samples from a single sequencing batch were removed as a result.

Differential Expression Analysis.

A quantile normalization scheme was applied to all samples (Bolstad, Irizarry et al. 2003). Following log2 transformation of the data, we used **`Limma`** (a Bioconductor package) to conduct our differential expression analyses. **`Limma`** utilizes a standard variance moderated across all genes using a Bayesian model and produces *p*-values with greater degrees of freedom (Smyth 2004). When appropriate, the **`Voom`** module was used to transform the data based on observational level weights derived from the mean-variance relationship (Law, Chen et al. 2014). To control for expression profile changes in human cells over culture, we subtracted the 'infected' from 'uninfected' normalized-log2-transformed expression values for each gene in

the paired uninfected/infected HFF samples at each time point. Experimental batch effects were adjusted for by including experimental batch as a covariate in our statistical model (Soneson and Delorenzi 2013). To correct for Type I error rooted from multiple comparisons, we used q-value as the statistical parameter (Storey 2002). A contrast matrix was used within **`Limma`**. We identified differentially expressed genes between infected human cells and human controls at the same time point, as well as between infected human cells at different time points post invasion. For the samples of *T. cruzi*, we identified differentially expressed genes between intracellular and extracellular stages, between the two extracellular stages and between different time points of intracellular stages. Differentially expressed genes were defined as genes with q-value < 0.05. We also generated DEG lists for both human and *T. cruzi* requiring q-value < 0.05 and fold change greater than 2. DEGs were also compared to published gene list from primary siRNA screening (Caradonna, Engel et al. 2013) and intersections were outputted as tables.

K means clustering and motif analysis

K-means clustering analysis was performed to identify genes with similar expression profiles across different developmental stages for both human host and *T. cruzi* with the R function “kmeans” and using the Hartigan-Wong algorithm (Hartigan 1979). Euclidean distance was used as distance metric; 20 partitions were used to generate the clusters. Quantile-normalized and batch-effect-adjusted expression values were used for clustering. Both the 3’ and 5’UTR sequences of genes from these clusters were retrieved from Chaper 2 and used

as input for XXmotif to identify potential regulatory motifs *de novo (Luehr, Hartmann et al. 2012).* We limited the motif search to a single strand with "--revcomp NO" option and constrained the maximum motif size to 6 nt with "--max-match-positions 6". The output motifs were further examined with the 53 motifs identified from groups of genes with similar metabolically related functions reported by Frasch and colleagues.

GOSeq analysis

Significantly regulated genes in each comparison of the differential expression analysis and genes from each K-mean cluster were then classified into GO functional categories with package GOSeq, which applied Wallenius approximation to correct the bias of over-detection of differential expression for long and highly expressed transcripts(Young, Wakefield et al. 2010). FDR is controlled by Benjamini and Hochberg's procedure (Benjamini and Hochberg 1995).

Gene Set Enrichment Analysis (GSEA).

Differential expression as well as k means clustering gene lists were subjected to Gene Set Enrichment Analysis (GSEA) (Subramanian, Tamayo et al. 2005). We used the pre-ranked gene tool of this Java-based software developed by Broad Institute. For human host, we investigated enriched gene set in both KEGG pathways and Gene Ontology categories from MSigDB; while for *T. cruzi*, we focused on the enriched gene set from GO database from TritrypDB. To adjust for multiple testing, we used the FDR < 0.1 as the statistical significance cutoff. The

number of permutations in the test was set as 1000. We ranked the genes according to their fold change for the gene lists from differential expression analysis and chose to use weighted calculation mode assigning higher enrichment scores to genes on top of the list. We used the classic mode for the gene belong to the same cluster of k-mean cluster treating genes in the same list as equal

Databases.

The coding sequence coordinates and Gene Ontology terms for *T. cruzi* genes were also obtained from www.tritrypdb.org. The human reference genome was downloaded from UCSC Genome Bioinformatics database (http://hgdownload.cse.ucsc.edu/downloads.html), hg19 GRCh37. We used Molecular Signatures Databases (MSigDB) collections from Broad Institute (http://www.broadinstitute.org/gsea/msigdb/collections.jsp) for Gene Set enrichment analysis, which includes KEGG gene sets and GO gene sets for human host.

2.4 Results

Defining *T. cruzi* and human cell transcriptomes by RNA sequencing

To capture the global transcriptional response during the initiation and maintenance of intracellular infection by the Chagas disease parasite, *Trypanosoma cruzi*, the transcriptomes of the parasite and infected human host cells were simultaneously profiled by RNA-Seq. Monolayers of low passage primary human foreskin fibroblasts (HFF) were infected with tissue culture-derived trypomastigotes of the *T. cruzi* Y strain. Messenger RNA was isolated from pure trypomastigote populations, from uninfected and infected HFF monolayers at 4 hrs, 6 hrs, 12 hrs, 24 hrs, 48 hrs and 72 hrs and from axenically-grown *T. cruzi* epimastigotes that correspond to the replicating insect vector stage of the parasite (**Figure 26**). RNA-Seq libraries were constructed from poly(A) + RNA and paired-end sequence reads of 101 nucleotides were obtained on the HiSeq-1500 platform. Two to four independent biological replicates were sequenced for each condition generating 2.7 billion high-quality reads for 34 samples (**Tables 6, 7, 8, see appendix 2**). Reads that passed Illumina quality filters were further examined using FastQC software. The pipeline of our analysis is shown in **Figure 27**.

Examination of sequenced data

To ensure the read quality of our large scale dataset, we quantitatively examined the raw sequenced reads coming from high throughput sequencing pipelines. We applied the tool FastQC on Linux system as well as in-house R script to obtain an

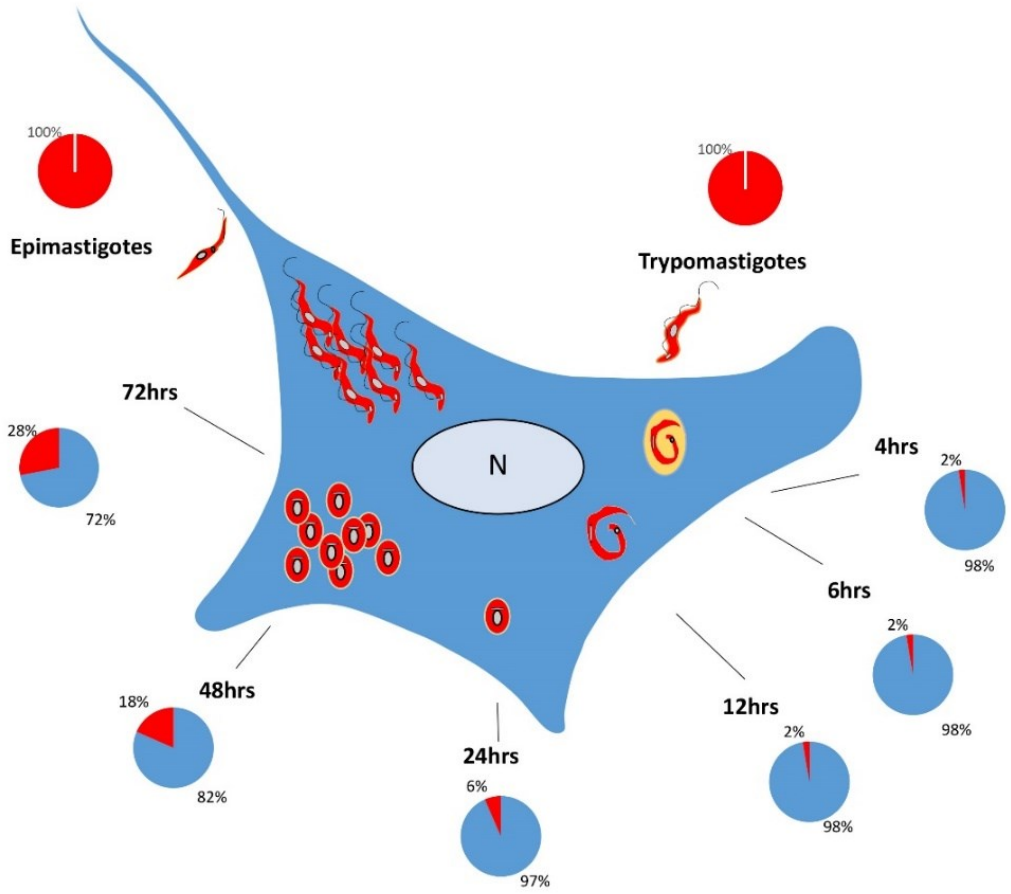

Figure 26 Simultaneous interrogation of the host and pathogen transcriptomes - Experimental design.

Human foreskin fibroblasts (HFF) were exposed to *T. cruzi* trypomastigotes (Y strain) for 2 hrs to permit invasion. Extracellular parasites were then removed by extensive washing in PBS, culture medium (DMEM+2% FBS) replaced and the infection allowed to progress for 72 hours at 37ºC. RNA samples were collected at 4, 6, 12, 24, 48 and 72 hrs post invasion (hpi). Following the attachment of infective trypomastigotes to HFFs, they are targeted to vacuoles that have fused with lysosomes. Within the vacuole, trypomastigotes differentiate into

amastigotes 2-8 hpi and gradually egress into the cytoplasm (8-16 hpi). Amastigotes begin to replicate around 20 hpi and continue to divide, eventually transform back into trypomastigotes, rupturing the cell and spreading the infection within 96-120 hpi. Biological replicates (2 to 4) were collected for each of the stages of infection. Experimental and data collection details are included in Table S1. The pie charts indicate the proportion of sequence reads assigned to the parasite (red) or human (blue).

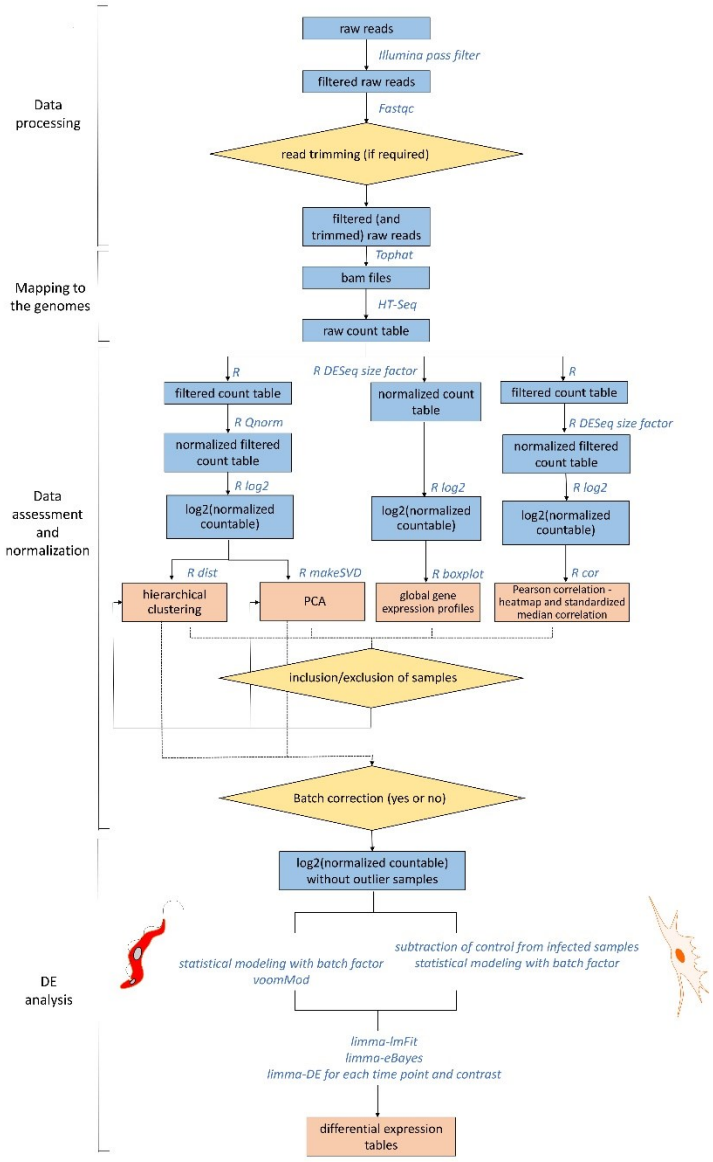

Figure 27 Overview of the pipeline for differential expression analysis.

The input/output for each step of the data processing and analysis is depicted in rectangular boxes. Software/scripts or methodological components are shown in italicized blue text. Decision-making steps are represented diamond-shaped boxes. The four general steps are shown in the left margin.

overview of the sequence data before doing any further analysis. The quality scores of each samples were examined. Out of the 34 samples, the mean quality score of both reads in 23 samples exceed 30 at all nucleotide positions, whereas the mean quality score of the last nucleotide from reads of the other 11 samples fell below 30 and was trimmed before downstream analysis. Two examples of sequencing quality score check were included in **Figure 28**, of which the one in panel B requiring trimming of the last nucleotide. The nucleotide composition of sequenced reads for each samples were also examined (**Figure 29**). In a random library, it is expected that little or no difference should be observed between different bases of a sequence run, so the lines, representing nucleotide composition, should be parallel and around 25% mark. However, we noticed a distinctive biased pattern at the first 12-14 bp of the read across all samples. There was an overrepresentation of sequence combination (CTNAAATCT) at the 5' end of the read. After the first 12-14 positions, the nucleotide frequencies became independent of position. This selection bias has been reported in other RNA-Seq experiments as well and is not relevant to organism or laboratory (Hansen, Brenner et al. 2010). This dependence of nucleotide frequency on position at the 5' end of the read was not caused by sequencing errors but originated from a biased selection of random hexamer primers during the cDNA reverse transcription step in library construction. The extension of the pattern beyond the hexamer primer, out to 12 to 14 bases can probably be explained by the length of two hexamer primers or by the dependencies between the first 6 nucleotides and the adjacent nucleotides.

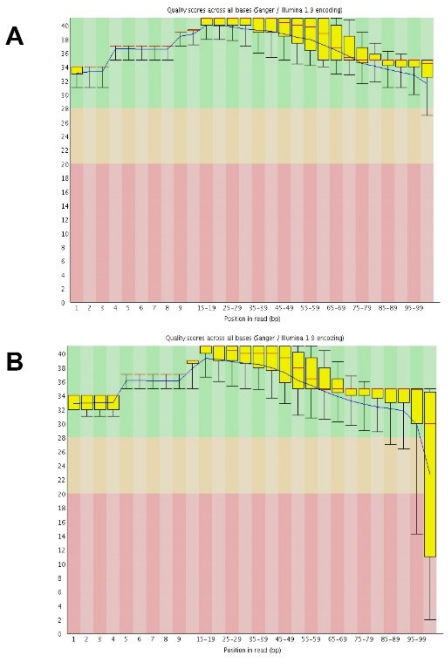

Figure 28 Examples of quality scores across all bases from sequenced samples by FastQC.

We examined the quality scores of sequenced reads for all sequenced samples and displayed two examples here. (A) HPGL0063 Read 2 and (B) HPGL0062 Read 2. For each position, a Boxplot is drawn to visualize the quality score distribution. The yellow box indicates the inter-quartile range with the central red line presenting the median value and the blue line represents the mean quality. The upper and lower whiskers denotes the 10% and 90% points. If the mean quality score of the last nucleotide was lower than 28, it would be trimmed before aligning to the genome (for example HPGL0062 in panel B).

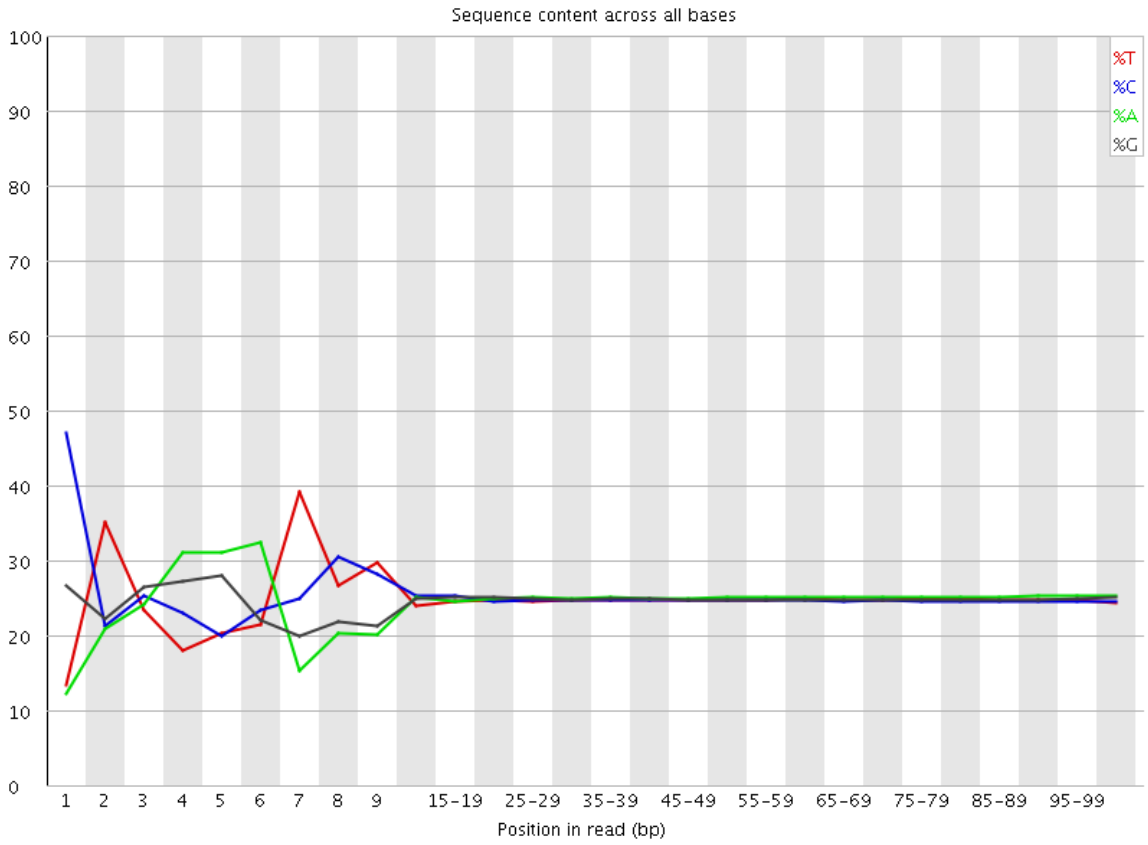

Figure 29 An example of nucleotide composition from sequenced reads.

The nucleotide composition of each base position were calculated and visualized by software FastQC. A clear bias were noted at the first 14 nucleotides at the 5' end of sequence reads: the composition of A, T, C and G were greater or less than 25%. This phenomena was caused by the effects of random hexamer priming step applied in the cDNA library construction (Hansen, Brenner et al. 2010).

In addition, to gathering a general idea about coverage saturation, we investigated the minimum sequencing coverage required for the detection of genes required for both the parasite and human (**Figure 30**). We combined mapped reads that came from biological replicates at an increasing order for each developmental stage to simulate the increase of sequencing depth. The number of genes detected were defined as the number of genes with at least one read mapped into the CDS region. For example, we examined the number of non-zero genes detected in epimastigotes with only HPGL0251, or combined samples of HPGL0251 and HPGL0252, or HPGL0251, HPGL0252 and HPGL0253. For human samples from the same time point, we observed a positive correlation between the number of non-zero genes detected and the number of sequenced reads included (**Figure 30A**). However, the comparison between the number of non-zero genes and sequencing coverage across samples from different time points did not reveal the same correlation, indicating different number of genes were expressed at various stages: with similar sequencing coverage, the number of non-zero gene detected was different across stages. On the contrary, for *T. cruzi* samples at the same developmental stage, we only noticed increased number of detected genes with deeper coverage in stages with relatively low sequencing depth (**Figure 30B**). In trypomastigote and epimastigote samples, where deeper coverage was retrieved, the number of detected genes reached saturation with even only one replicate included. This discovery was consistent with the global expression pattern observed in **Figure 34**: Majority of the *T. cruzi* genes were expressed

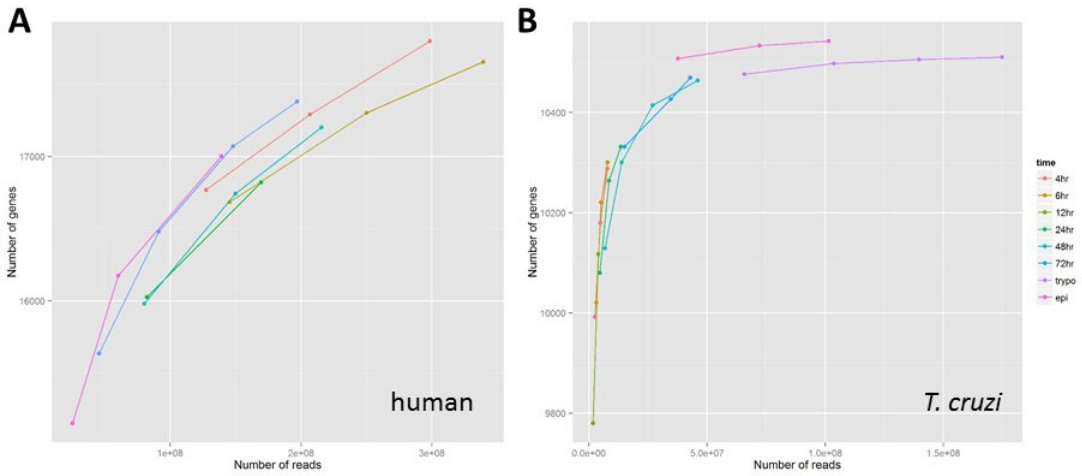

Figure 30 Number of non-zero genes detected in samples with various sequencing depth.

We plotted the number of genes with at least one read mapped to the CDS (y axis) against the number of sequenced reads from each of the developmental stages (x axis) included in this study for (A) human and (B) *T. cruzi*. We combined reads from different number of biological replicates of the same stage to investigate the effect of various sequencing coverage on the detection of genes.

across different developmental stages, in contrast with human, in which a large portion of the genes were regulated at the transcriptional level and were not expressed, so only in *T. cruzi*, the increase of sequencing depth can result in the plateau of gene detection.

Proportion of reads that can be aligned to *T. cruzi* or human genomes

We also examined the number of reads from the pool that mapped to the *T. cruzi* or human reference genomes (**Figure 31; Table 8**). Each RNA-Seq library generated from infected cells consisted of a pool of mixed reads from *T. cruzi* and human fibroblasts. To parse these out, the pre-processed RNA-Seq reads were mapped against corresponding reference genomes using the Tophat aligner program (Trapnell, Pachter et al. 2009). The genome of the *T. cruzi* CL Brener Esmeraldo haplotype (DTU II) (Berriman, Ghedin et al. 2005) was used as the reference to map *T. cruzi* Y strain sequences (also DTU II) (de Freitas, Augusto-Pinto et al. 2006). The human hg19 genome sequences provided a reference to map human host cell sequences. The fraction of reads mapping to the parasite or human genome references were used to estimate the proportion of RNA molecules from each source. In samples from infected cells, where ~30% of host cells contained intracellular parasites, a relatively small fraction (2% at 4-12 hpi) of the mapped read pool aligned to the *T. cruzi* reference genome at pre-replication time points (**Figure 26**). This fraction increased to 28% by 72 hpi, reflective of parasite doublings that occur between 24-72 hpi. Nevertheless, the fraction of parasite reads is likely to be underestimated since a significant

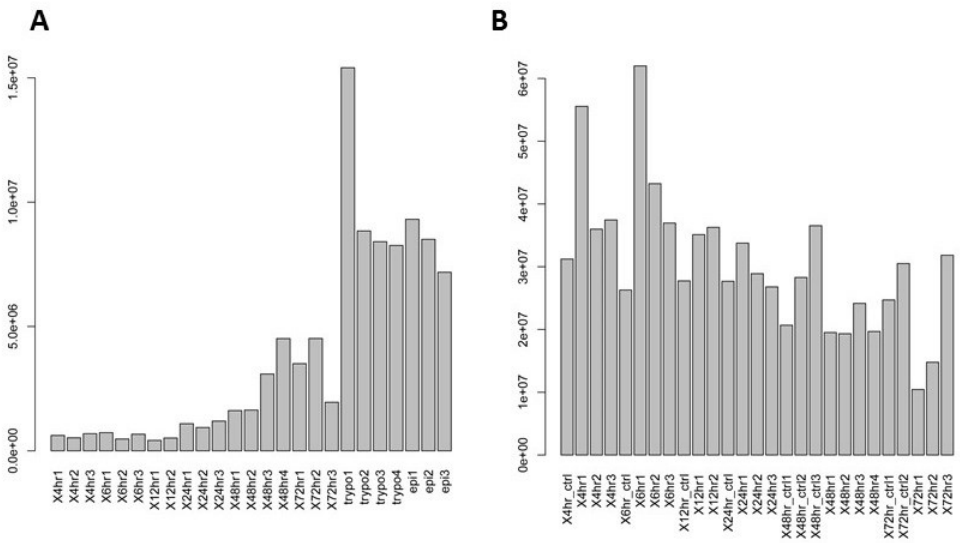

Figure 31 Number of mapped reads for *T. cruzi* and human samples.

The number of reads that were mapped to (A) *T. cruzi* and (B) human reference genome were computed and displayed for each sample. The x axis indicated the samples included in the analysis.

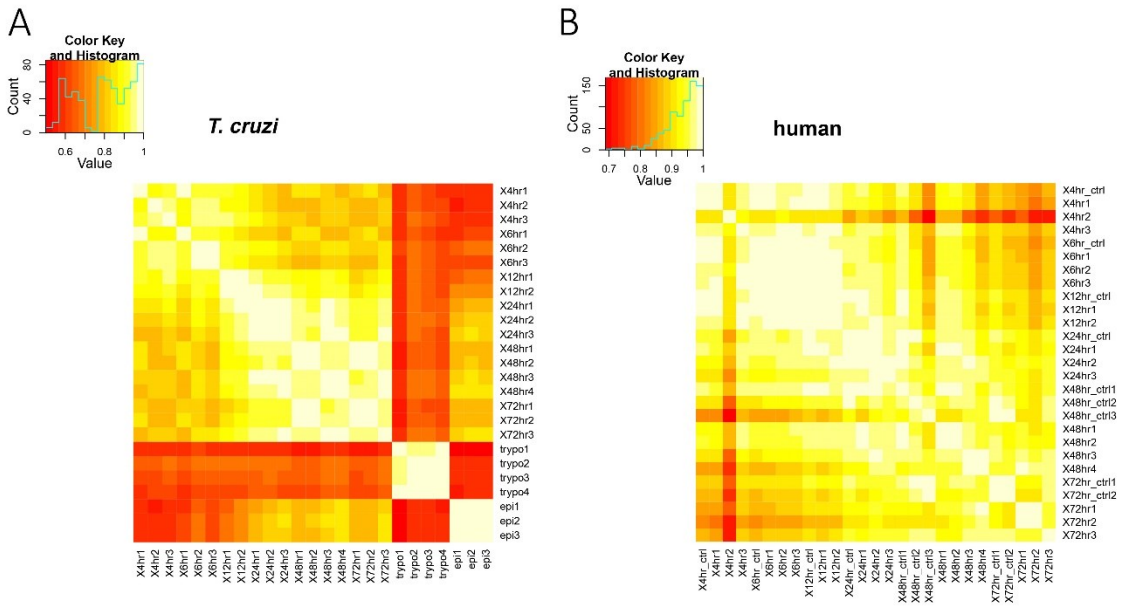

Figure 32 Heatmap of Pearson correlation between samples.

Gene counts were normalized for sequencing library size. All pairwise Pearson correlations were calculated and plotted as a heatmap to view the relatedness of samples and identify outliers for (A) *T. cruzi* and (B) human.

number of parasite reads do not align to the reference genome when constrained by the allowance of no more than two mismatches. This is evident in extracellular *T. cruzi* trypomastigote and epimastigote samples where up to 38% and 44% (respectively) of the reads could not be aligned to the reference genome using our stringent parameters. Allowing for more mismatches (5) increased the reads mapping to the *T. cruzi* genome to between ~60-70%. The remaining unmappable reads likely result from highly polymorphic multicopy gene family members, which are usually fast evolving and more divergent in *T. cruzi* (Cerqueira, Bartholomeu et al. 2008). Thus, for the differential gene expression analyses reads were mapped with two or fewer mismatches. The relative abundance of each CDS was estimated by counting the number of reads mapped within the CDS boundaries with the HT-Seq software [http://www-huber.embl.de/users/anders/HTSeq/]. For *T. cruzi* samples, mean-variance bias of read counts were corrected by package Voom, whereas for human samples, the ratio of infected/control counts were used and no mean-variance bias were detected (**Figure 33**).

Global expression profiles

Our initial analyses examined global expression profiles for consistency between similar samples and evaluated experimental variation and reproducibility within our datasets. To examine the global gene expression levels of *T. cruzi* and its human host cell, the log2-transformed and size-factor-normalized gene counts for both species were calculated and one representative sample from each group

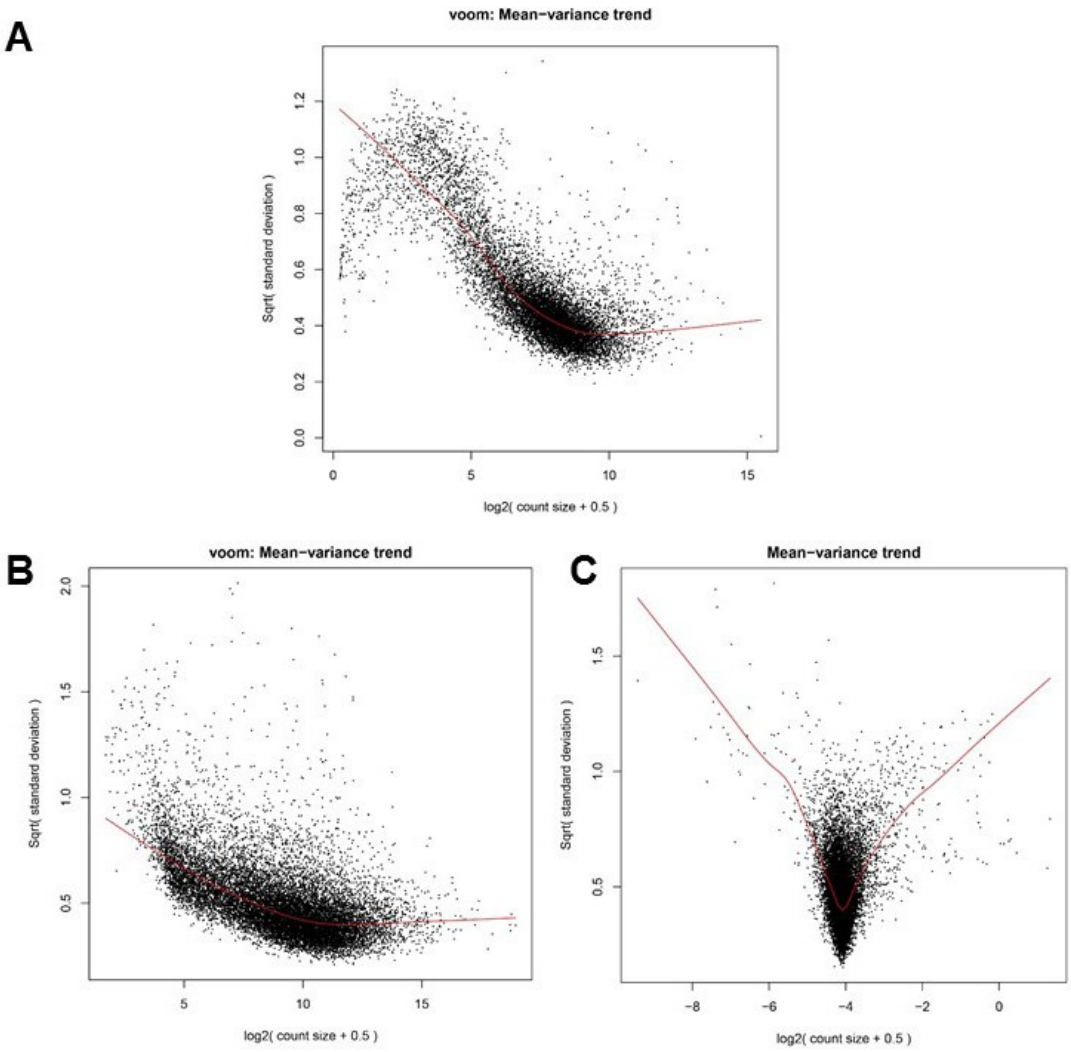

Figure 33 Mean-variance trend of *T. cruzi* and human samples.

We applied Voom package on the R platform to estimate and the mean-variance relationship of the log-counts for both *T. cruzi* (A) and human (B) samples. To remove the time progression effects from uninfected human samples, we used the ratio of infected/control at each time point in the investigation of differential expression genes (See method for details). The mean-variance relationship of the logged ratio was plotted in panel C.

is graphically depicted as box plots (**Figure 34 and Table S14**). Mapped read counts from parasite extracellular and intracellular stage samples as well as human cells displayed consistent degrees of dispersion as indicated by the nearly identical quartile distributions in similar samples. A particularly interesting observation stems from the examination of the median expression values which displayed a compact distribution for parasite genes (with an interquartile range of 8.8 to 10.8 in epimastigote samples, for example, in **Figure 34**, *T. cruzi*) and a much broader dynamic range for human genes (interquartile range of 0 to 10) (**Figure 34**, human), highlighting marked differences in transcriptional regulation between the two organisms. Gene expression in mammalian cells shows highly restricted tissue-specific patterns with complex and tight regulatory mechanisms exerted at multiple levels (Cortez, Martins et al. 2012). The resulting gene expression profile is a supple one with a wide range of expression and a large number of genes not expressed (only 61% of the genes in HFFs are expressed here). In contrast, a large majority (98.7%) of trypanosomatid genes are expressed during various developmental stages, yet with a much narrower range of steady-state transcript level variation. Both features are the result of a general lack of control at the transcriptional level with most of the fine-tuning of monocistronic mRNA levels occurring mostly at the post-transcriptional level through pre-mRNA processing and RNA degradation (Martinez-Calvillo, Vizuet-de-Rueda et al. 2010).

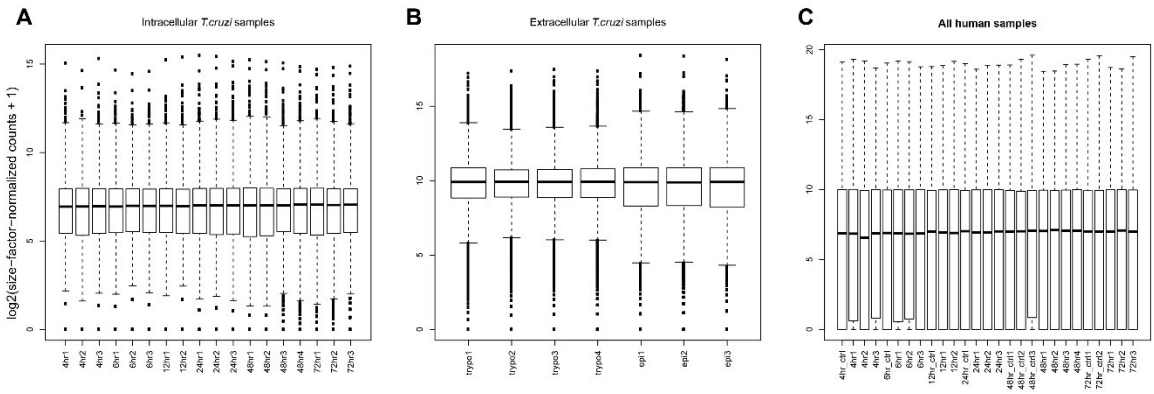

Figure 34 Distribution of global gene expression levels in *T. cruzi* and human samples.

For samples from (A) *T. cruzi* intracellular and (B) extracellular stages, and (C) human , counts were normalized for sequencing library size and a boxplot was generated to compare the distribution of per-gene counts (log2 counts per million with an offset of 1). The ends of the whiskers represent the lowest datum still within 1.5 interquartile range (IQR) of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile. Gene features with extremely high or low expression levels are shown as open circles above and below the whiskers, respectively.

Assessment of the data with various statistical analysis

We used Pearson correlation and median pairwise correlation analyses to evaluate the reproducibility and experimental variation within our samples. For parasite samples, the Pearson correlation matrix revealed high correlation levels among biological replicates for each of the developmental stages (**Figures 32A, 39**) but also highlighted similarities between the intracellular stages (4 hpi to 72 hpi) and a contrast with the extracellular stages (trypomastigotes and epimastigotes). The correlation matrix for the human transcriptome samples also displayed expected correlation levels between biological replicates with the exception of one sample ("4hr2", HPGL0111) which was removed from downstream analysis (**Figures 32B, 40**). The same sample was identified as an outlier when we carried out a more systematic median pairwise correlation analysis (**Figure 35**).

In order to investigate general trends in the data and identify and quantify batch effects, we carried out principal component analysis (PCA) as well as hierarchical clustering visualization of all samples. A high degree of similarity between biological replicates is evident in the PCA plots for both *T. cruzi* and human samples (**Figure 36**). The spreading of points along the first principal component reflects the time progression of intracellular amastigote infection. Samples from the same time points of infection clustered together for both *T. cruzi* and human samples, although the correlation between infected human samples was less rigid. Uninfected human samples clustered according to time point reflecting the

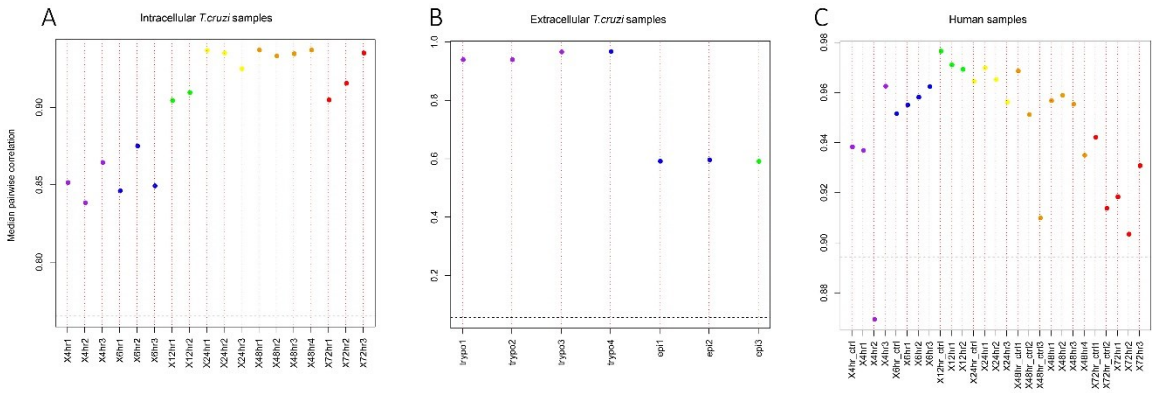

Figure 35 Standardized median Pearson correlation between *T. cruzi* and human samples.

Gene counts were normalized for sequencing library size. The standardized median Pearson correlation between each sample and all other samples was plotted to view the relatedness of samples and identify outliers for (A) intracellular *T. cruzi,* extracellular *T. cruzi* (B) and (C) human. Letters [A-E] in the sample name refer to experimental batch.

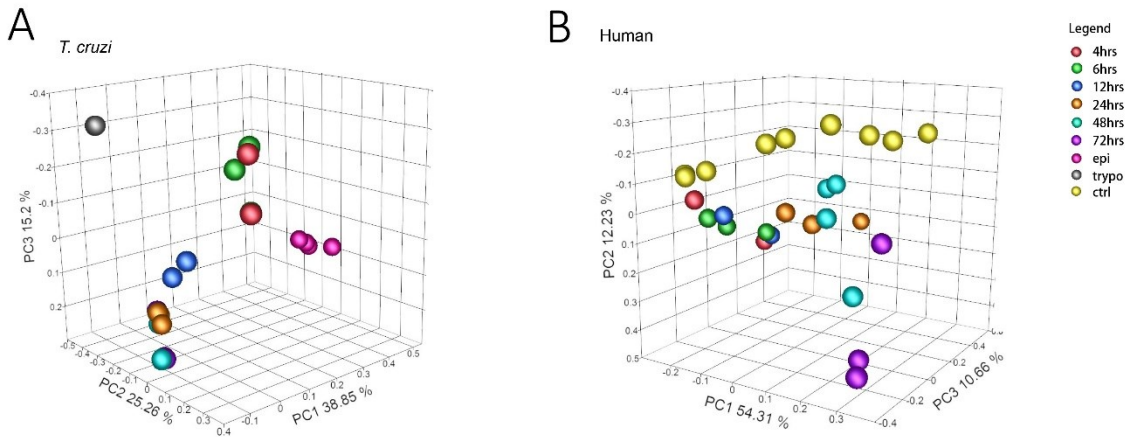

Figure 36 Principal component analysis of global transcriptome profiles in *T. cruzi* and human host cells at various stages of the infection.

To evaluate our biological replicates and the relationships between samples across time points and to visualize sample-sample distances for genes in (A) *T. cruzi* (extracellular and intracellular forms) and (B) human cells (infected and uninfected controls), we conducted a comprehensive principal component analysis (PCA) on all collected samples. Very weakly expressed genes were removed and counts were quantile-normalized and log2-transformed before PCA.

drift in the transcriptome of these cells as they grow in tissue culture. Individual samples from trypomastigotes and epimastigotes were tightly clustered (**Figure 36**). Unsupervised hierarchical clustering of *T. cruzi* samples identified four distinct groups: trypomastigotes, epimastigotes and intracellular amastigotes at early (4, 6, and 12 hpi) and late developmental stages (24, 48, and 72 hpi), with the exception of a few samples (**Figure 37**). Hierarchical clustering of human samples presented the difference between early and late developmental stages of infected samples (**Figure 38**). The same pattern was observed in uninfected human samples indicating a change in the HFF transcriptome over time. In conclusion, our analyses confirm reproducibility between biological replicates, enable us to detect temporal progression of samples, and reveal the variation in individual transcriptomes.

RNA-Seq, like many other technologies applied in biology, require a complicated set of reagents, hardware and highly trained personnel to yield accurate and replicable results. Batch effects were defined as sub-groups of measurements that have qualitatively inconsistent behavior across conditions and are irrelevant to the biological factors investigated in a study (Leek, Scharpf et al. 2010). In our study, samples were prepared by the same laboratory but the infection experiments were carried out on five different dates spanning a year period. So here we considered experiment date as a distinctive batch source. In this study, we carefully examined batch effects in samples by using both PCA and hierarchical clustering dendrograms (**Figure 36, 37, 38**). We did not observe

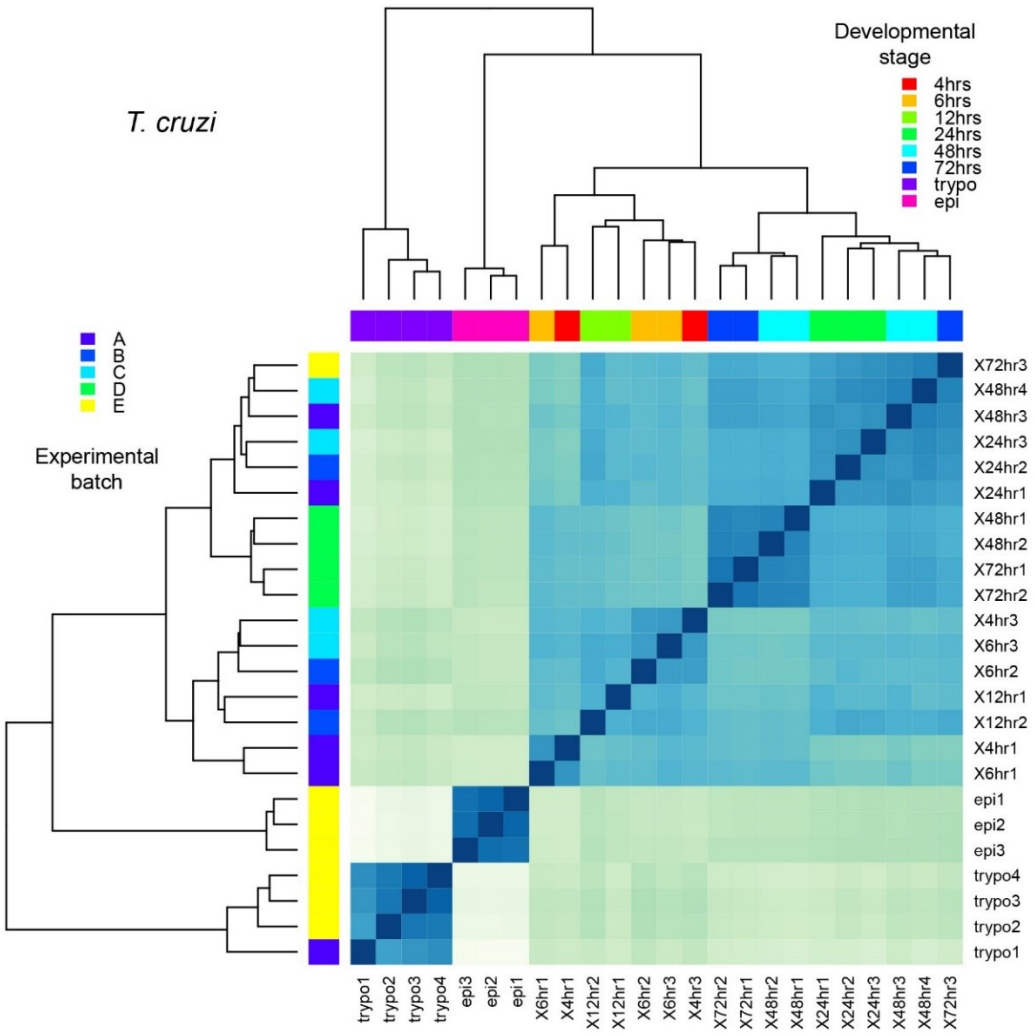

Figure 37 Hierarchical clustering of *T. cruzi* samples.

Hierarchical clustering analysis based on Euclidean distance was performed using all *T. cruzi* genes after filtering for weakly expressed genes, quantile normalization, and inclusion of the batch variable in the statistical model used by **`limma`**. Colors along the top of the heatmap indicate the developmental stage and colors along the left side of the heatmap indicate the batch/experimental date.

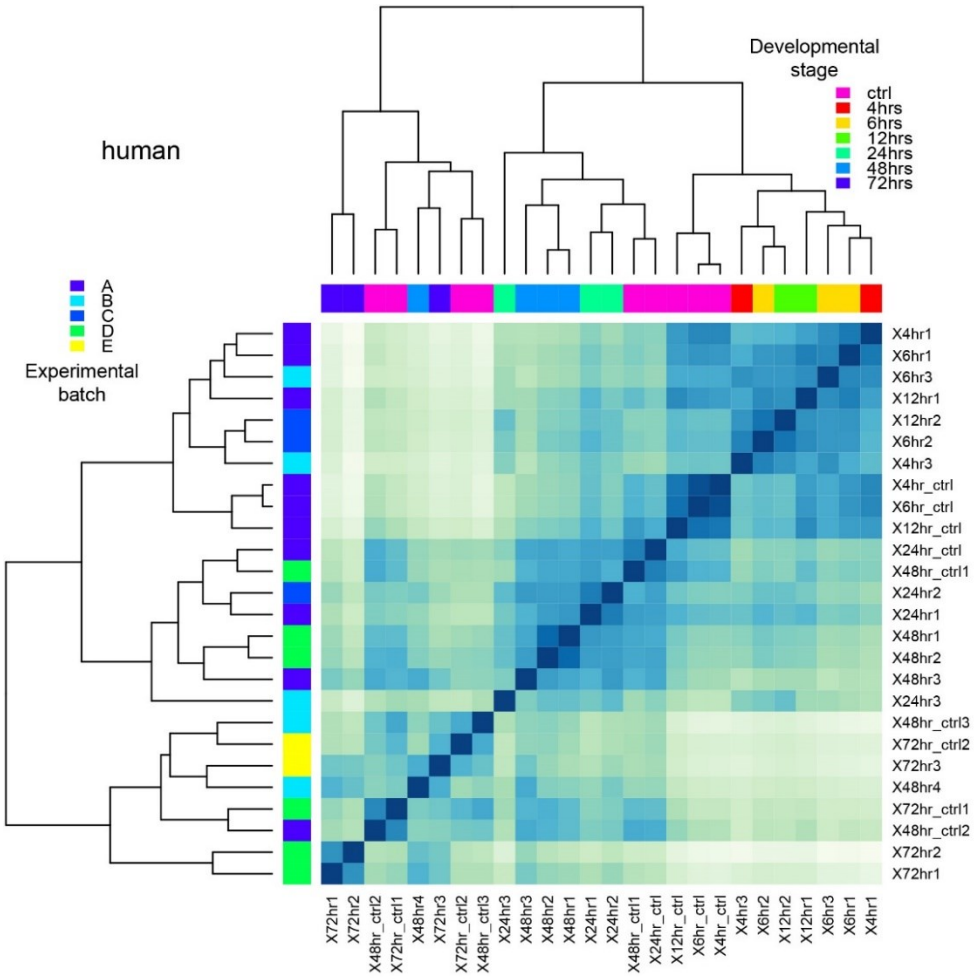

Figure 38 Hierarchical clustering of human samples.

Hierarchical clustering analysis based on Euclidean distance was performed using all human genes after filtering for weakly expressed genes, quantile normalization, and inclusion of the batch variable in the statistical model used by **`limma`**. Colors along the top of the heatmap indicate the developmental stage and colors along the left side of the heatmap indicate the batch/experimental date.

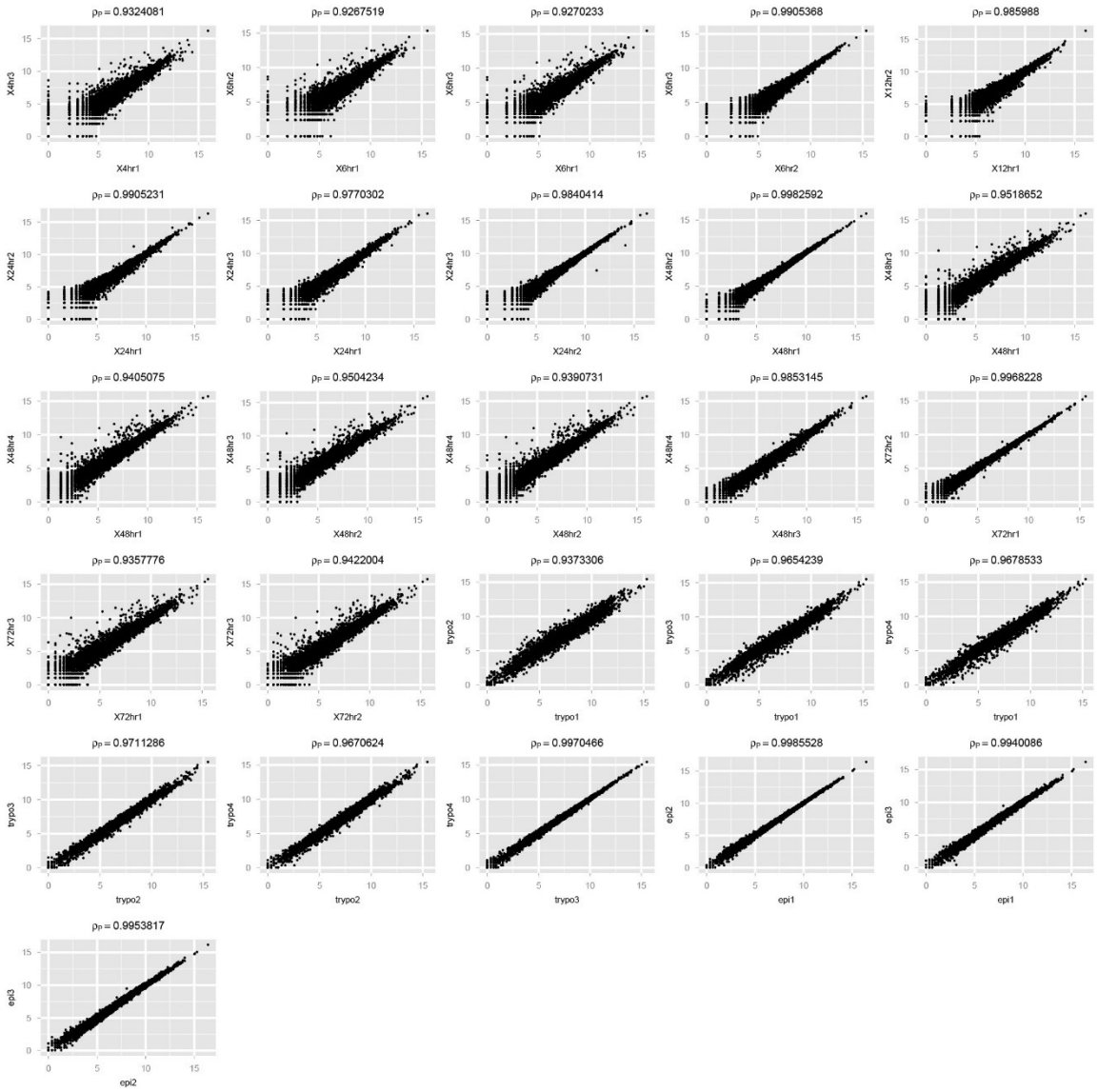

Figure 39 Pairwise Pearson correlation between *T. cruzi* samples.

Gene counts were normalized for sequencing library size. The Pearson correlation between each sample and all other samples was calculated and plotted to view the relatedness of samples and identify outliers.

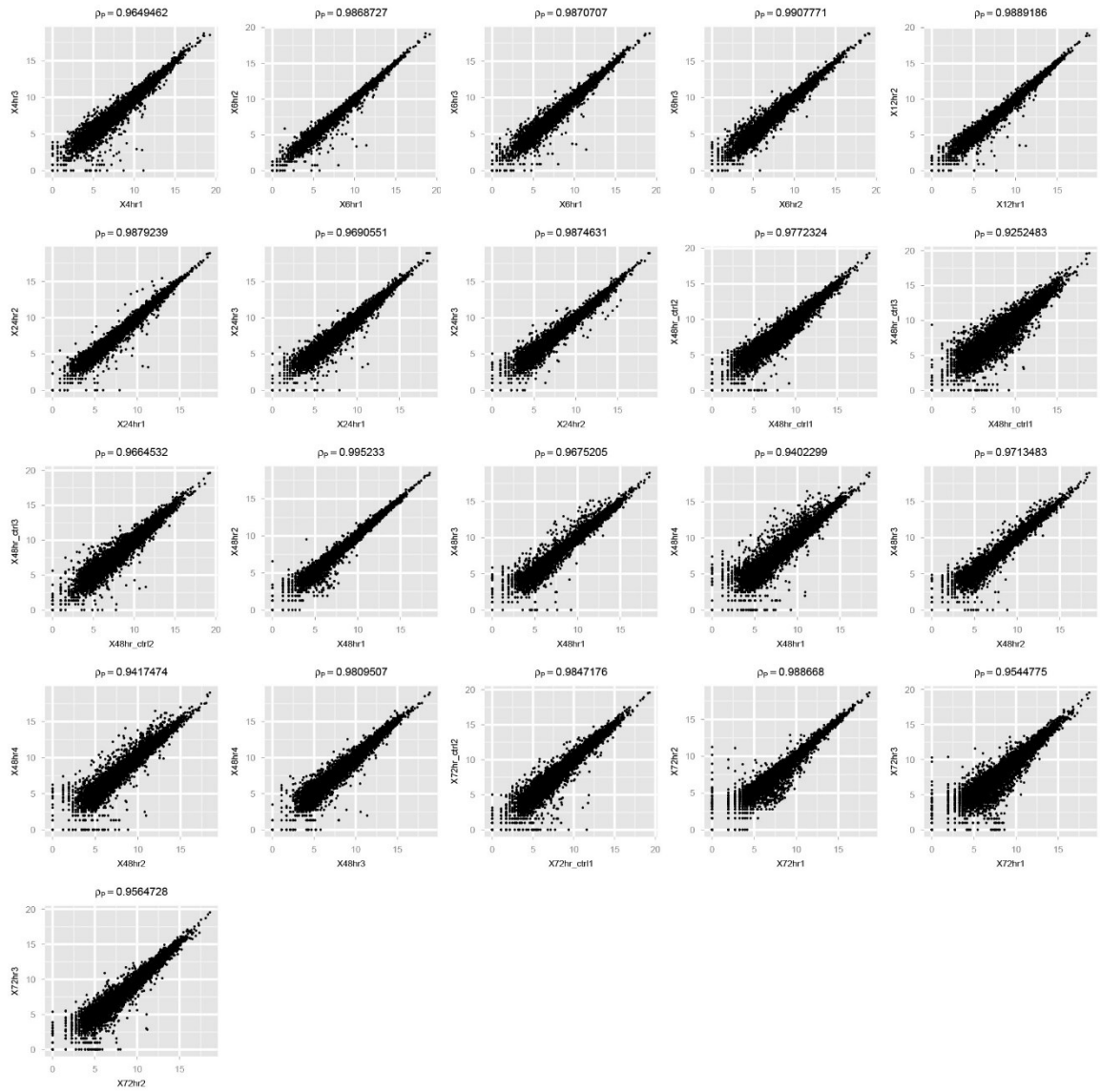

Figure 40 Pairwise Pearson correlation between human samples.

Gene counts were normalized for sequencing library size. The Pearson correlation between each sample and all other samples was calculated and plotted to view the relatedness of samples and identify outliers.

strong batch effects, but a relatively mild one: samples were clustered by developmental stages rather than batches; principal components were associated with biological conditions, rather than batch effects. However, in our samples from certain time points, such as trypomastigote or epimastigote, batch effects were highly cofounded with biological variables of interest (developmental stages and/or infection). This correlation makes it difficult to determine whether observed differences across biological groups are due to biology or artifacts. In addition, because of this unbalanced batch design of our samples, particularly the extracellular *T. cruzi* samples, we did not apply Combat package to remove batch effects to prevent the overcorrection or bias introduced by the package. Instead, we incorporated batch factor as a variable in the linear model that were used in statistical analyses. Differentially expressed genes (DEGs) in *T. cruzi* across developmental stages or in mock- or *T. cruzi*-infected human cells at different time points of infection, were identified using the Limma software package with t-statistics (**Methods**). Type I error introduced by multiple testing was corrected with q-value (Storey 2002). Differentially expressed genes were defined as genes with q-value < 0.05. In addition, to obtain gene lists with bigger biological changes, we imposed an arbitrary fold-change cut-off (2-fold) and outputted the results. Genes that were identified as differential expressed were organized into supplementary tables (**Tables S15, S16, S17, S18, S19**).

Unique signatures of gene expression in *T. cruzi* developmental stages.

Pairwise comparisons of differential gene expression were made between *T. cruzi* developmental and intracellular infection stages and significantly enriched GO functional categories were assigned. General features of intracellular *T. cruzi* developmental stages emerging in these analyses include a boost in cellular metabolic processes, DNA replication and protein translation as parasites transition from non-dividing, extracellular trypomastigotes to replicative amastigotes in the mammalian host cell (**Table S17)**. Energy generation via citric acid cycle (TCA) cycle and proton-coupled ATP synthesis are enriched functions in replicating intracellular amastigotes, as are carbohydrate and lipid metabolism including fatty acid, isoprenoid and ergosterol synthesis (**Table S17**). Glycolysis is also found to be an enriched GO function in replicating amastigotes, contrary to the belief that glucose utilization, while elevated in trypomastigotes, is dampened in amastigotes in favor of amino acid and fatty acid catabolism.

During its digenetic life cycle, *T. cruzi* switches between different developmental forms that are well-adapted to life in different hosts and niches within hosts. As a reflection of their divergent lifestyles, we note that the greatest differences in steady state transcriptomes are observed between distinct life cycle stages of the parasite (**Figure 42**). With the constraint of minimum fold change of two, between ~2800 and 3600 DEGs are observed when the transcriptome of the invasive extracellular trypomastigote stage is compared with any of the intracellular parasite stages or with epimastigotes (**Figure 42)**. In contrast, once the

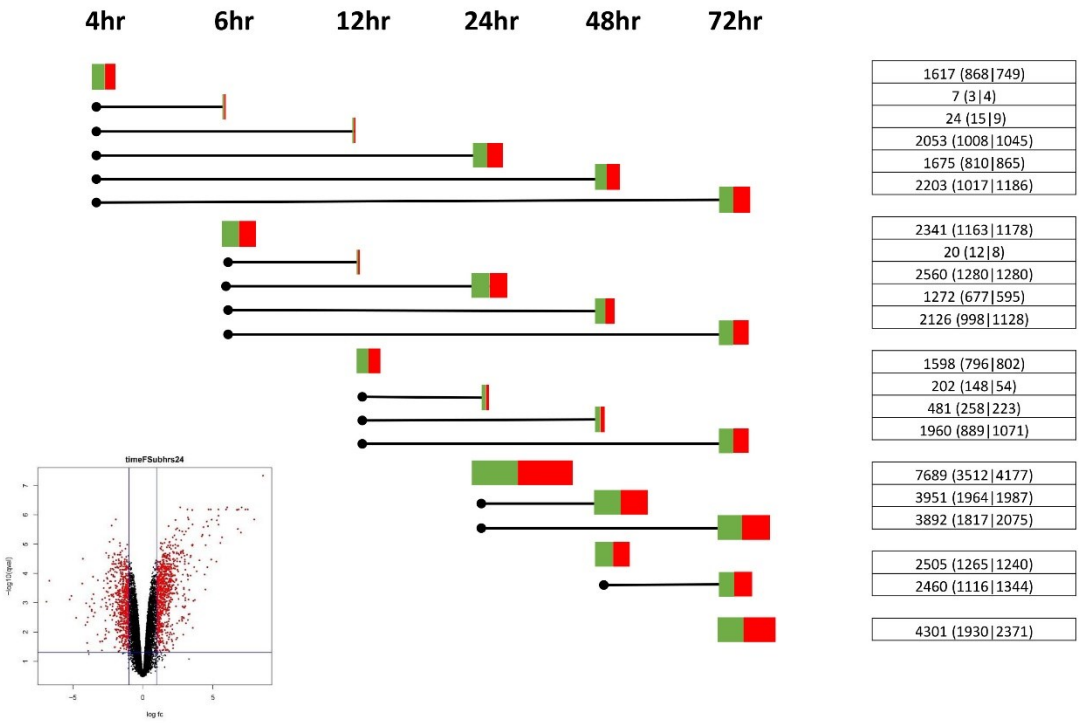

Figure 41 Differentially expressed (DE) genes from human at various stages of the infection.

Edges indicate the pairs of time points for each DE profile. The size of each box represents the number of DE genes, with the red and green portions denoting up- and down-regulated genes, respectively. A volcano plot from the comparison between control and infected human samples were included in the lower panel.

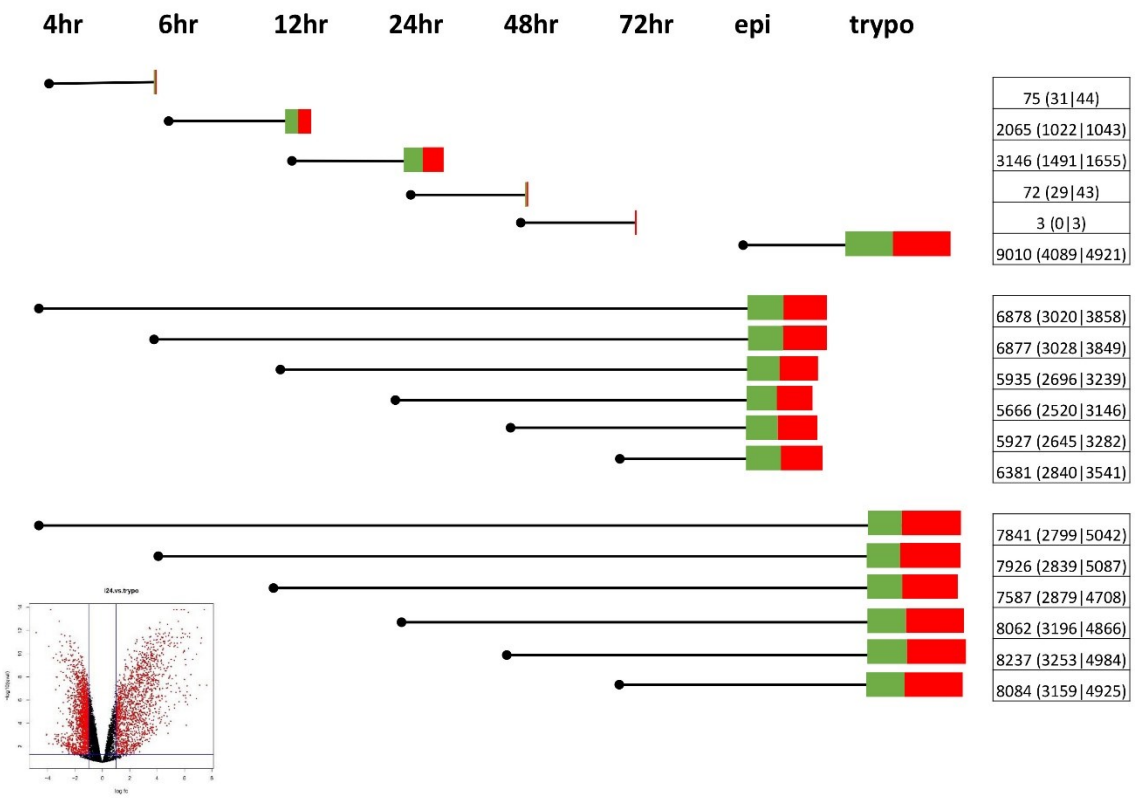

Figure 42 Differentially expressed (DE) genes from *T. cruzi* at various stages of the infection.

Edges indicate the pairs of time points for each DE profile. The size of each box represents the number of DE genes, with the red and green portions denoting up- and down-regulated genes, respectively. A volcano plot from the comparison between 24 hpi amastigote and trypomastigote samples were included in the lower panel.

transition from trypomastigotes to intracellular amastigotes is initiated by 4 hpi, additional changes in parasite gene expression are comparatively modest with ~400 additional DEGs by 72 hpi (**Figure 42, 43**; **Table S15**). Moreover, once replication competent amastigotes are formed by 24 hpi, few additional changes in the steady state transcriptome are noted (9 genes) suggesting that large shifts in the steady state transcriptome of *T. cruzi* occur only when the parasite undergoes a developmental change. Without the constraint of fold change, ~ 7000 and ~9000 genes were detected as differential expressed between amastigote and trypomastigote or epimastigote, marking the striking changes of transcriptome between intracellular and extracellular stages (**Table S15**).

Transcripts that were highly over-represented in a single life cycle stage of *T. cruzi,* and poorly represented in others, were considered to be stage-specific (**Table S20**). For example, only genes that were significantly upregulated in one specific stage in comparison with the other two were included in this analysis. The set of trypomastigote-'enriched' genes reveals that in addition to hundreds of polymorphic surface antigens (comprised of *trans*-sialidases, mucins, MASPs and gp63 surface protease), several glycosyltransferases, protein kinases and other signaling proteins such as receptor-type adenylate cyclases were enriched in this life cycle stage (**Table S20**). Epimastigotes had an abundance of 'stage-specific' transcripts encoding metabolic functions (**Table S20**). The number of unique transcripts associated with intracellular amastigotes, was relatively low with 34 genes selectively expressed in replicating amastigotes as compared to

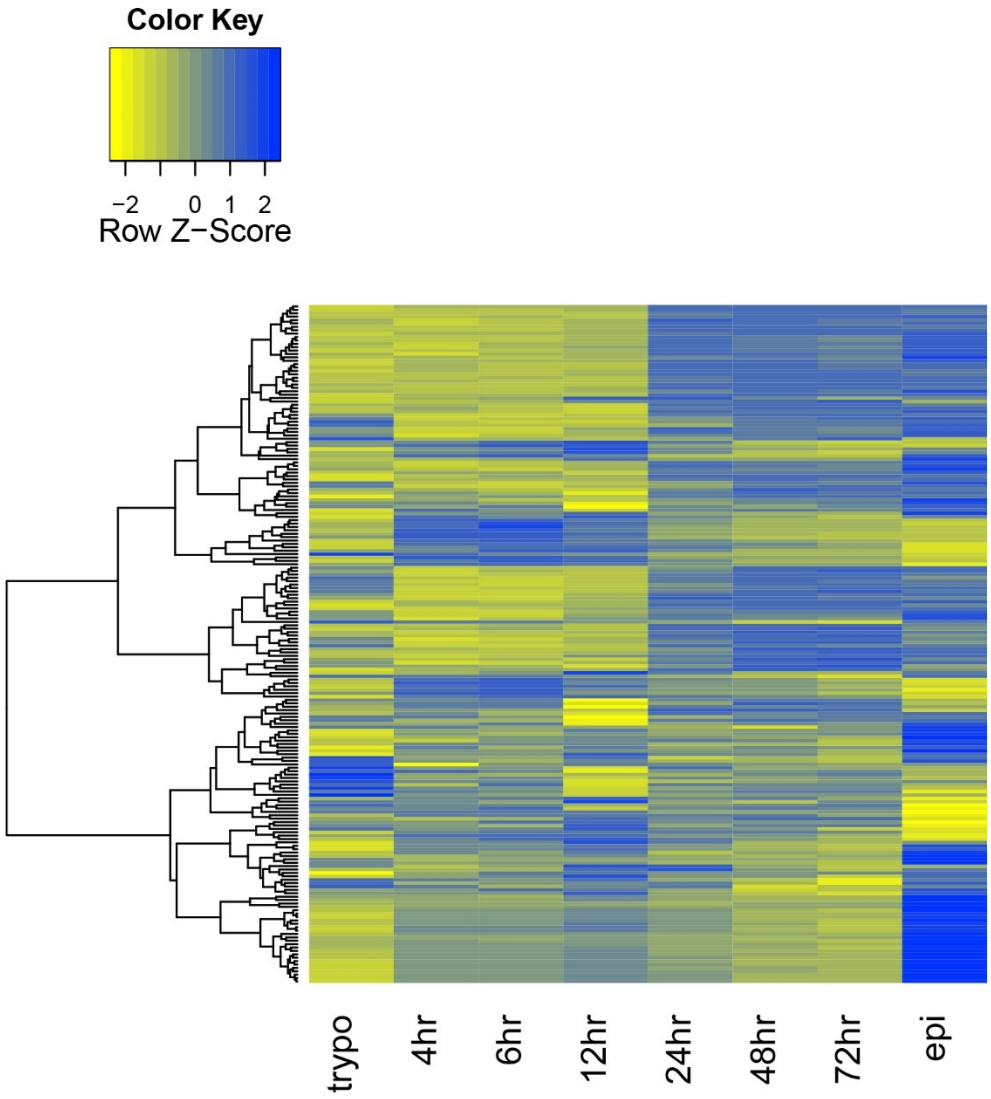

Figure 43 Heatmap of the top 200 *T. cruzi* genes significantly regulated pre- and post- replication at the intracellular stages.

Yellow bars represent downregulation and blue bars represent upregulation. The top 200 significantly regulated with the biggest fold change between 12hpi and 72 hpi were selected and hierarchically clustered.

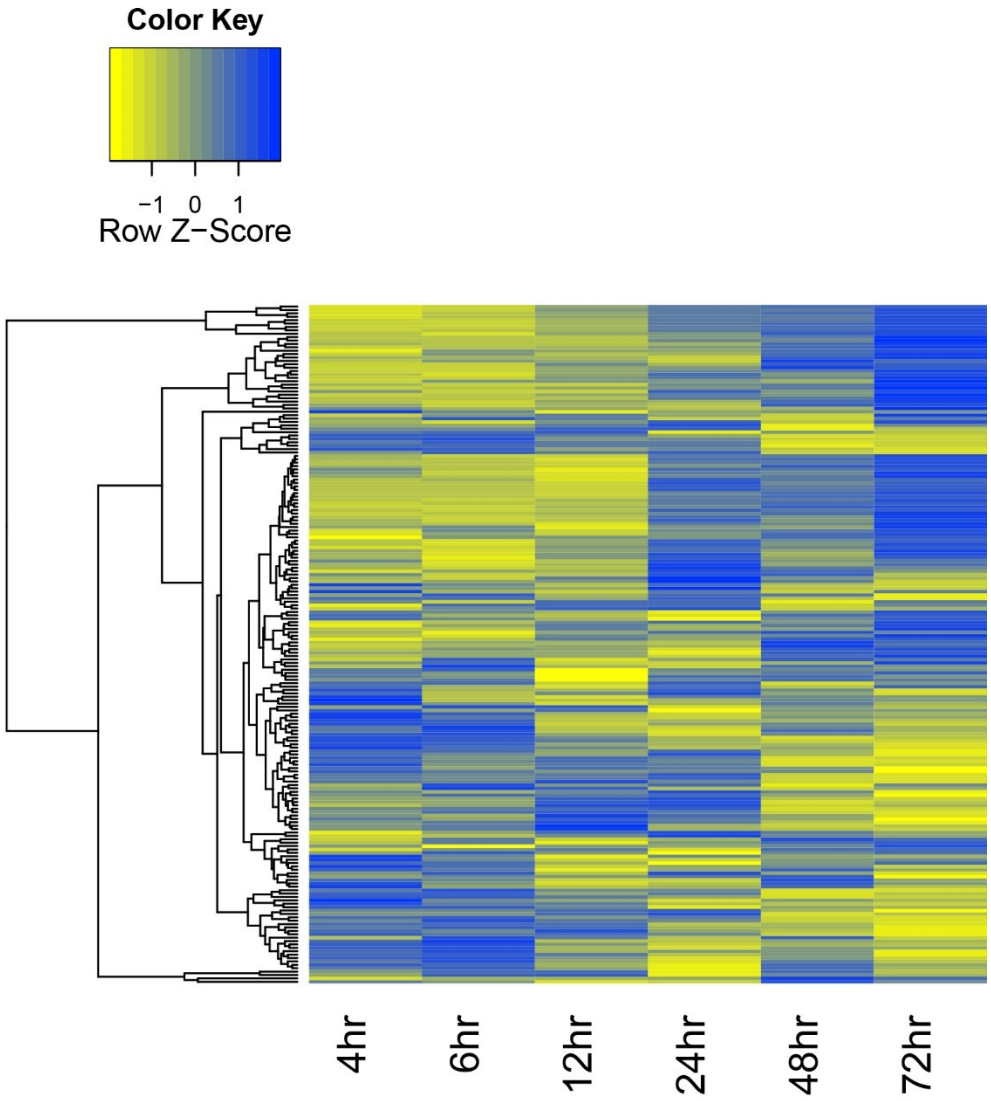

Figure 44 Heatmap of the top 200 human genes significantly regulated pre- and post- intracellular parasite replication.

Yellow bars represent downregulation and blue bars represent upregulation. The top 200 significantly regulated with the biggest fold change between 12hpi and 72 hpi were selected and hierarchically clustered.

the extracellular stages. These included genes coding for surface proteins such as amastins - which are one of the few amastigote-specific gene families to be characterized (Teixeira, Kirchhoff et al. 1995; Cruz, Souza-Melo et al. 2012). Analysis of the hypothetical protein coding genes revealed three highly related genes with unique structural features (Tc00.1047053438923.10; Tc00.1047053506495.40; Tc00.1047053506495.50) that have not been previously described. As the translated proteins are predicted to have signal peptides and a variable number of nearly identical repeating peptide sequences. Two of them also have predicted GPI anchor attachment motifs (Pierleoni, Martelli et al. 2008), these proteins are predicted to be trafficked to the surface of the intracellular amastigote where they are in a position to intersect host function. As hypothetical protein coding genes comprise ~60% of the parasite genome, parasite stage-specific expression analysis is a powerful tool for discovery of novel genes that are expressed in intracellular amastigotes in the context of mammalian cell infection.

Early amastigote differentiation in mammalian host cells.

The time points of intracellular *T. cruzi* infection analyzed in this study span the critical stages of development including trypomastigote-to-amastigote differentiation (0, 4, 6, 12 hpi) and intracellular amastigote replication (24, 48, 72 hpi) (**Figure 26**). As illustrated in **Figure 42**, a dramatic shift in the steady state *T. cruzi* transcriptome coincides with the early establishment of mammalian cell infection (**Table S19**) whereupon non-dividing, cell invasive trypomastigotes

receive developmental cues to trigger the transformation to replicative amastigotes. Most notable is the rapid downregulation of a large number of transcripts encoding the GPI-anchored surface proteins: *trans*-sialidases, mucins and MASPs within the first 4 hours of infection (**Table S19**) and a concomitant increase in the expression of amastigote-specific genes, such as members of the amastin gene family (Teixeira, Kirchhoff et al. 1995; Cruz, Souza-Melo et al. 2012). Other anticipated findings include the downregulation of parasite genes encoding flagellar and paraflagellar rod proteins as motile trypomastigotes transition to non-motile amastigotes.

Evidence of altered signal transduction capacities is reflected in the DEG data. Trypomastigotes exhibit higher expression of components of cAMP-dependent signaling processes such as receptor-type adenylate cyclase, cAMP-specific phosphodiesterases, cAMP-dependent protein kinase (PKA) catalytic and regulatory subunits than amastigotes. However, the induction of a different PKA regulatory subunit gene and a phosphodiesterase in amastigotes at 4 hpi suggests stage-specific regulation of a pathway that is recognized as a regulator of *T. cruzi* growth and differentiation (Gonzales-Perdomo, Romero et al. 1988; Bao, Weiss et al. 2009). Parasite signaling proteins upregulated in intracellular amastigotes include an activated protein kinase C receptor, several putative protein kinases, serine/threonine phosphatases as well as a putative protein tyrosine phosphatase (**Table S19, S21**). Among the most highly expressed, annotated genes in early amastigotes are a GPI-inositol deacylase (Guther,

Prescott et al. 2003; Hong, Nagamune et al. 2006) and members of a secreted lipase family known to be differentially expressed in the amastigote stage (Belaunzaran, Wilkowsky et al. 2013). These findings suggest the need for new parasite GPI anchor synthesis and present the possibility that parasite lipases might participate in vacuole remodeling and egress in addition to manipulation of host signaling and cellular processes (Martins Vde, Galizzi et al. 2010). In addition to the early induction of signaling components, strong expression of several parasite metabolic genes, including urocanate hydratase and imidazolonepropionase involved in histidine metabolism, was observed at 4 hpi. A number of amino acid permeases are also rapidly upregulated in intracellular parasites suggesting an increased reliance on amino acids for energy and/or protein synthesis. The significant upregulation of multiple genes encoding rRNA subunit proteins, elongation factors, RNA polymerases are consistent with the predicted need for increased protein synthesis. Despite the fact that these intracellular amastigotes will not undergo their first round of replication for another ~20 hours, preparation for the eventuality of cell doubling is already apparent at 4 hpi. The upregulation of nucleoside transporter as well as pyrimidine synthesis and purine salvage enzymes are also consistent with increased RNA synthesis at this stage and for DNA synthesis at later time points. As intracellular amastigotes advance through their developmental program from 4 - 24 hpi, an increasing number of *T. cruzi* DEGs is observed (from 15 to 415) in a comparison of consecutive time points (**Figure 42; Table S15**).

Host cell response after invasion of *T. cruzi*

Along with the generation of the first comprehensive *T. cruzi* transcriptome dataset, our RNA-Seq analysis enabled us to simultaneously capture changes in the steady state transcriptome of infected human host cells. To identify host cell genes that were differentially expressed upon infection at different time points, gene expression values obtained from matched mock-infected fibroblasts were subtracted from values obtained from infected cells. In the early hours of infection (4-12 hpi), 500-600 genes were identified as significantly altered in *T. cruzi*-infected fibroblasts as compared to uninfected controls (**Table S15**)**.** Consistent with previous microarray studies, a type I interferon (IFN) signature (Vaena de Avalos, Blader et al. 2002; Chessler, Unnikrishnan et al. 2009; Costales, Daily et al. 2009) is observed in our human fibroblast transcriptome. This signature is detectable as early as 4 hpi, where IFN-inducible genes are among the most highly expressed genes in infected cells. By 24 hpi, 1458 human genes were differentially expressed reflecting both the response to cytosolic amastigotes as well as secondary and tertiary responses to soluble factors including type I IFNs. A similar number of fibroblast genes (1176) were modulated at 72 hpi, whereas only 339 DEGs were detected at 48 hpi. Variation between independent samples at the 48 hpi time point, as shown in the PCA and hierarchical clustering analysis, is the likely cause of this anomalously low value. All DEGs for comparisons between each pair of time points are reported (**Figures 41, 44; Table S16, S17, S22**).

K-mean cluster and Gene Ontology enrichment analysis for clusters

To investigate the function representation of genes sharing similar expression patterns, K-mean clustering was exploited to group both *T. cruzi* and human genes by the developmental dynamics of their expression patterns (**Figures 45, 46)** where enriched functions in each cluster were revealed with GOSeq analysis. We defined k equals to 20. For human host cells, most of the GO categories showed enrichment for particular clusters of gene expression (**Table S24**). For example, genes that encode enzymes for DNA synthesis, cell cycle regulation and chromatin structure, cytoskeleton assembly and organization, protein metabolism, potential signaling proteins (phosphatidylinositol or inositol lipid-mediated signaling) were greatly enriched in clusters 10 and 14, suggesting human cells were down-regulating cell proliferation shortly after the replication of intracellular parasites started. Genes that show peak expression at 24 hpi or 72 hpi (clusters 6 and 15) included those that were involved in biological processes of production and response to different cytokines, antigen processing and presentation, regulation of apoptosis, various pathways involved in immune reaction (MDA-5, interferon-gamma-mediated, RIG-I or cytoplasmic pattern recognition receptor signaling pathways), reflecting the stressful and defensive state of host cells and suggesting that the immune response activates soon after the parasite infection and intensifies through different stages. Finally, in cluster 7 and 20, gene expression is nearly exclusively committed to metabolism and biosynthesis with predominant expression of genes that encode enzymes for nitrogen compound, nucleobase-containing compound, heterocycle, isoprenoid,

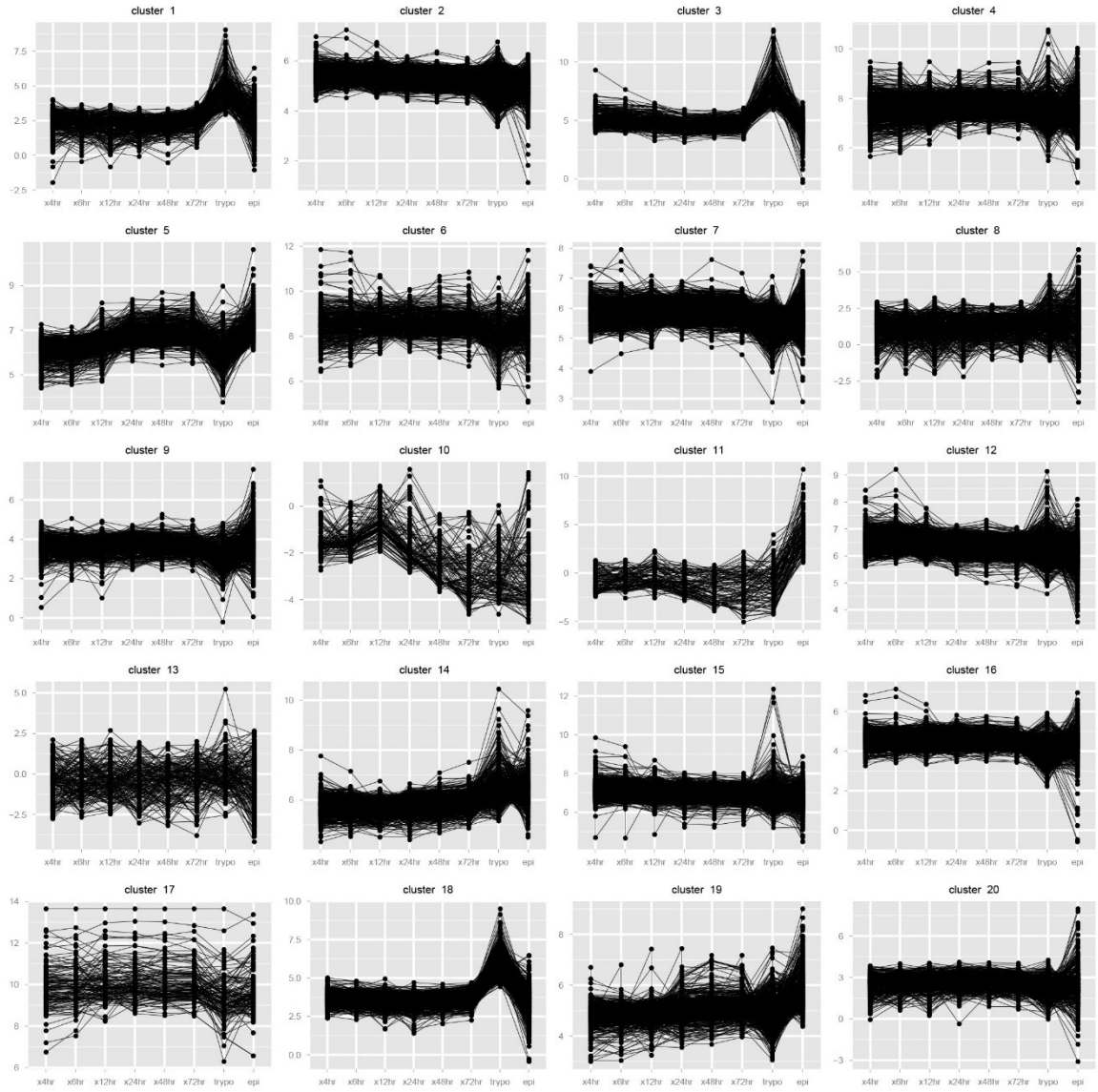

Figure 45 K-means clustering of *T. cruzi* transcriptome based on the dynamic progression of gene profiles across different developmental stages.

The x-axis shows the time points included in the analysis: intracellular amastigotes 4hpi-72hpi, trypomastigotes, and epimastigote. The y-axis represents the quantile-normalized batch-adjusted gene expression values.

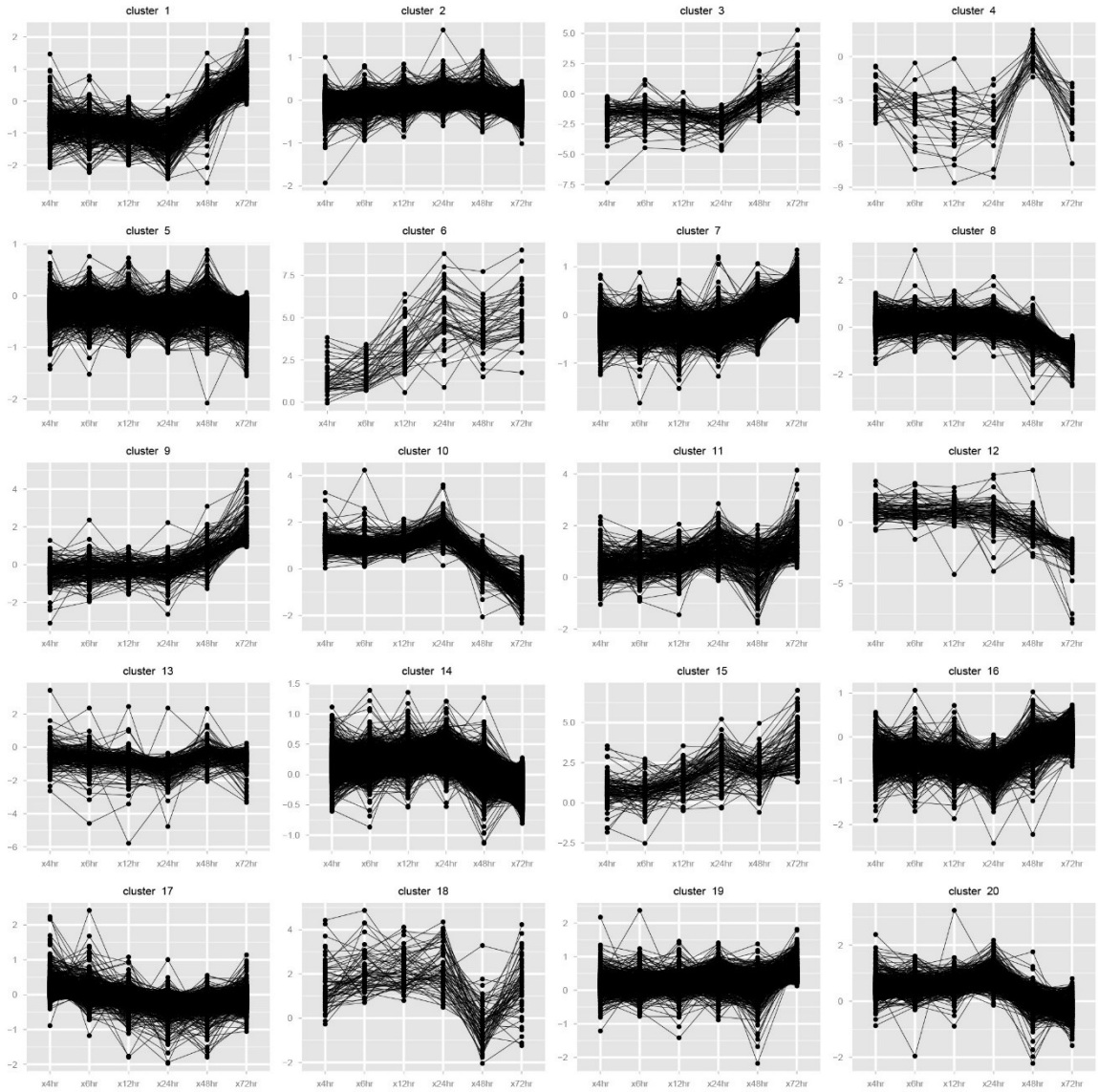

Figure 46 K-means clustering of human transcriptome based on the dynamic progression of gene profiles across different developmental stages.

The x-axis shows the time points included in the analysis: 4hpi-72hpi. The y-axis represents the quantile-normalized batch-adjusted gene expression values.

alcohol, organic hydroxyl compound and lipid biosynthesis. Genes that encode enzymes and transporters for organic cyclic compound metabolism, aromatic compound metabolism, ubiquitin-dependent protein catabolic process and muscle cell differentiation were also greatly enriched in this segment. Together, these data revealed that the major biochemical shifts along the infection process are produced in part by highly dynamic, coordinated and localized transitions in mRNA abundance.

For *T. cruzi*, most of the GO categories also showed enrichment for particular clusters of gene expression (**Figure 45**; **Table S23**). For example, in cluster 5 very abundant transcripts were detected for genes annotated as encoding DNA recombination, DNA replication, DNA repair, and telomere maintenance. Genes predicted to encode proteins involved in ATP synthesis coupled proton transport, phosphorylation and cell redox homeostasis were also disproportionally represented in the same cluster. This group of genes increased their expression levels throughout the intracellular stages with the highest level at epimastigote stage and the lowest at trypomastigote stage. Genes contributing to Cluster 6 included those annotated as regulation of translational initiation, elongation or termination, ribosome biogenesis, protein folding, tRNA aminoacylation, translational frameshifting, suggesting active translational regulation at epimastigote stage. The greatest enrichment in cluster 20 was observed for genes annotated as vesicle docking involved in exocytosis, vesicle-mediated

transport, cell adhesion, proteolysis followed by translational elongation and pathogenesis, with major changes occurred at epimastigote stage.

Gene set enrichment analysis

The Gene set enrichment analysis tool (GSEA) developed by the Broad Institute was applied to investigate functional associations of gene expression changes between different set of comparisons (Subramanian, Tamayo et al. 2005), supplementary materials). DEGs that were up- or down- regulated were analyzed by functional clustering. Kyoto Encyclopedia of Genes and Genomes (KEGG) was selected as the *priori* defined annotation category for enrichment analysis of human genes and Gene Ontology (GO) terms was selected for *T. cruzi* (Ashburner, Ball et al. 2000; Kanehisa and Goto 2000). Once the tool has ranked the DEGS according to the magnitude of their fold change, it identifies enriched KEGG/GO categories and assigns a statistical significance based on the number and the ranking of the constituent genes. We performed this analysis on all significantly regulated gene lists reported in the previous step.

For the human samples, functions associated to cytokine receptor interaction, chemokine signaling pathway, Nod-like receptor signaling pathway, Rig-like receptor signaling pathway, and Toll-like receptor signaling pathway were enriched in most of the developmental stages post invasion of *T. cruzi*. The enrichment of other functions related to human immune response were also reported, such as, intestinal immune network for IGA production at 12 hpi, B cell

receptor signaling pathway and epithelial cell signaling in helicobacter pylori infection at 24 hpi, and cytosolic DNA-sensing pathway at 72 hpi. The Jak-STAT signaling pathway were also up-regulated at 12 hpi. Interestingly, the enrichment of KEGG categories related to the rise of cytosolic Ca_2^+ concentration was also detected at late stages of infection. More than 20 genes, encoding adenylate cyclase, angiotensin II receptor, calcitonin receptor, potassium large conductance calcium-activated channel, guanylate cyclase, protein kinase phospholipase A2 were significantly upregulated at 48 and 72 hpi. Typical clinical manifestations of chronic phase of Chagas disease are grouped into three major forms: cardiac, digestive, and cardiodigestive, all of which can be linked to the attacks against the smooth muscle cells (SMCs), combined with the inflammation caused by long-term infection. However, the reason for the proneness of SMCs to the attack from *T. cruzi* and the mechanisms behind these symptoms were still unclear. The principal mechanisms involved in the regulation of smooth muscle cells rely on the modulation of cytosolic Ca_2^+ concentration. Specifically, the increase of calcium concentration triggers the Ca_2^+-CaM-MLCK pathway and stimulates MLC20 phosphorylation, leading to myosin-actin interactions and, hence, the development of contractile force. During receptor stimulation, the contractile force is greatly boosted by the inhibition of myosin phosphatase (Abdel-Latif 2001). Our data provided an interesting gene list that may contribute to the understanding of the disease and may help in identification of targets for chemotherapeutic intervention of Chagas disease.

By contrast, genes from *T. cruzi* known to be involved in sialidase activity, pathogenesis, and cell outer membrane were enriched in trypomastigotes. An overrepresentation of gene functions related to nucleosome assembly, DNA replication, DNA packaging, DNA recombination, translation, ribosome, kinetoplast, protein folding, ATP synthesis coupled proton transport and cell cycle at epimastigotes and amastigotes 24-72 hpi indicated the ongoing process of replication inside the insect host. Other enriched functions detected at the same time period, such as cell redox homeostasis, oxidation-reduction process, iron-sulfur cluster assembly, implied the actively regulated reaction in mitochondrial electron transport. We also noticed the overlap of many enriched functions between amastigotes 24-72 hpi and epimastigotes, between the early time points of 4 hpi and 6 hpi.

Motif analysis for coexpressed *T. cruzi* genes

In organisms like *Trypanosoma cruzi* and its kinetoplastid relatives, genes are transcribed as long polycistronic units that are *trans*-spliced, cleaved and processed post-transcriptionally. Thus, genome-wide differences in transcript abundance, as highlighted in the analyses of our RNA-Seq data, is primarily a consequence of post-transcriptional processes that exploit *trans*-acting regulatory factors. RNA-binding proteins are predicted to function in this capacity by recognizing *cis*-acting post-transcriptional regulatory elements located in the UTRs of mRNAs, as in other cellular systems. As such, the discovery of sequence motifs enriched in coexpressed gene sets is critical to our functional

genomic analysis. We applied XXmotif software (Luehr, Hartmann et al. 2012) to identify enriched sequence motifs in both the 3' and 5' UTRs of *T. cruzi* transcripts that have carefully defined in Chapter 2 for each of the coexpression gene clusters (**Figure 47**). XXmotif is a tool that can directly optimize the statistical significance of position weight matrix (PWMs), and score conservation and positional clustering of motifs. The motif size was restricted to 6 nucleotides in our analysis, since one of the common motifs identified in trypanosomes is the Adenylate/Uridylate (AU)-rich elements (ARE), of which the core sequence is usually consisted of less than 6 nucleotides (Chen and Shyu 1995; Peng, Chen et al. 1996). In general, more motifs were identified in 5' UTRs as compared to 3' UTRs. A total of 1566 motifs were identified in the 5' UTRs of the 20 clusters versus 544 in the 3' UTRs.

Several of the motifs we identified (**Tables S25, S26**) match known *cis*-acting elements involved in post-transcriptional regulation in trypanosomes. For example, the canonical cycling sequence (CS) CAUAGAAG, located in the 5'-UTR, is known to be recognized by Cycling Sequence Binding Protein (CSBP) II complex that is responsible for the regulation of a cohort of transcripts (Pasion, Hines et al. 1996; Mahmood, Mittra et al. 2001). Analysis of *T. cruzi* gene expression clusters revealed the enrichment an UAGAAG motif in the 5' UTRs of cluster 14, a group of genes for which expression gradually increases over the course of the intracellular infection cycle and were most highly expressed in epimastigotes and trypomastigotes. Interestingly, we also observed the

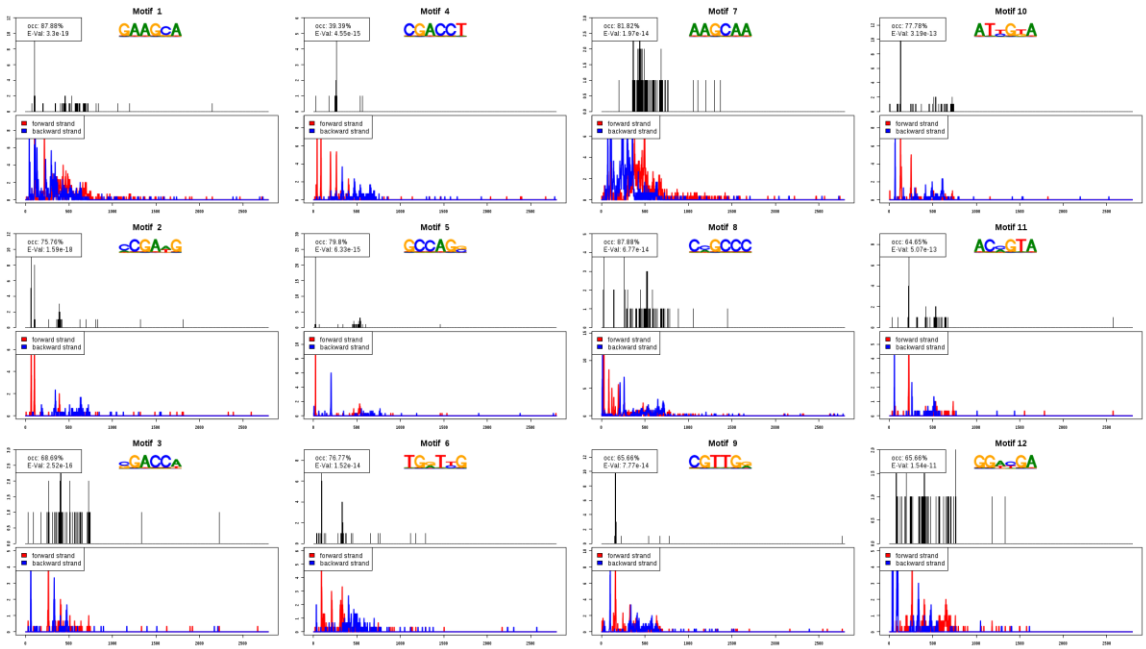

Figure 47 Overrepresented motifs detected in the 3' UTR of cluster 1 in *T. cruzi.*

Package XXmotif were applied to identify enrichment of motifs in the UTRs of gene clusters in *T. cruzi*. The top 12 of overrepresented motifs were presented here. The black bars indicate the position weighted matrix of motifs; the blue and red bars represent the occurrence of motifs on forward or backward strand.

remarkably increased expression of cell cycle sequence binding phosphoprotein (Tc00.1047053508541.190 and Tc00.1047053506777.50) at both epimastigote and trypomastigote stages in the differential expression analysis implying the regulation by the two *trans*-acting factors on this group of genes. Overrepresentation of AU-rich elements were also detected in our analysis. For example, the well conserved AUGUAU motif (Gerber, Herschlag et al. 2004; Archer, Inchaustegui et al. 2011), which contains a putative PUF-family RNA-binding domain core UGUA, was detected in the UTRs of multiple gene clusters (Cluster 2, 4, 7, 12, 14, 15, 16, and 19). As RNA-binding proteins are known to form complexes with other regulatory proteins, the presence of a putative PUF recognition motif in multiple coexpressed gene clusters may indicate combinatorial control of regulation. Recently, Frasch and colleagues investigated the enrichment of motifs in 53 gene clusters that share similar KEGG pathways (De Gaudenzi, Carmona et al. 2013). They were able to capture several sequence patterns in the 3' UTR for each of the metabolically related gene clusters. It should be noted that in their analysis, without the accurate definition of 3' UTR boundaries, they searched 350 nt downstream of the end of the CDS, whereas in our motif analysis, we were able to examine the UTRs that were precisely identified in previous chapter. A comparison of the motifs detected in our analysis to those reported by Frasch's team, revealed significant overlap of motifs between the coexpressed gene clusters and genes sharing similar KEGG pathways (**Tables S25, S26**). In particular, we were able to match the sequence patterns captured in the 5' UTRs of our analysis to those identified in the 3' UTR

in previous study indicating the RBP may function at either location, a phenomena that were previously noted in *T. brucei* as well (Archer, Inchaustegui et al. 2011). The conservation of sequence motifs in these regulated transcripts in trypanosomes further reinforce their potential functional consequences.

Relating transcriptome and function.

To determine if any of the host genes found to be differentially expressed following *T. cruzi* infection, have predicted functional roles in the parasite infection process, DEG data from our current study was compared with genome-scale functional data from the recent RNA-interference screen in HeLa cells to find the intersection (Caradonna, Engel et al. 2013). Of ~700 genes, that when silenced with siRNA pools exhibited a significant decrease (~380 genes) or increase (~320 genes) in intracellular *T. cruzi* infection, we find overlap with 50 differentially expressed human fibroblast genes in at least one of the infection time points (4-72 hpi; **Table S27**). Gating on human genes for which a > 2-fold change in expression occurred at any of the infection time points, we noted a strong correlation between the most highly upregulated genes and decreased parasite infection when these genes were silenced in the RNAi screen. This suggests that during the course of intracellular infection, *T. cruzi* induces (directly or indirectly) the expression of host genes that are beneficial to the establishment and maintenance of parasite infection. In support of this prediction, we highlight an example for which functional validation exists: GTP-cyclohydrolase 1 (GCH1) (Caradonna, Engel et al. 2013). GCH1 is the rate limiting enzyme in the

synthesis of tetrahydrobiopterin (BH_4). As pterin auxotrophs, *T. cruzi* amastigotes must scavenge these essential nutrients from the host. Our previous work showed that siRNA-mediated knockdown of host GCH1 impairs intracellular *T. cruzi* growth in a manner that could be rescued by the addition of exogenous BH_2 (Caradonna, Engel et al. 2013). Our transcriptomic data now reveals potential compensatory upregulation of this host pathway to service the increased demand for biopterin with the induction of GCH1 transcript levels beginning at 12 hpi (**Table S27**). Compensation appears to be occurring in the parasite as well, with the rapid upregulation of pterin-4-alpha-carbinolamine dehydratase in *T. cruzi* at 4 hpi. The involvement of this enzyme in BH_4 recycling is consistent with a need to increase flux through this essential pathway to fuel parasite growth.

2.5 Conclusion and discussion

Characterization of parasite and host transcriptomes during the course of intracellular infection provides fundamental insight into the dynamics of the host-pathogen interaction. Several studies have exploited microarray expression profiling to investigate the host response to *T. cruzi* infection in a variety of cells *in vitro* and *in vivo* (Shigihara, Hashimoto et al. 2008; Minning, Weatherly et al. 2009; Zhang, Kim et al. 2010; Manque, Probst et al. 2011; Tanowitz, Mukhopadhyay et al. 2011; Grynberg, Passos-Silva et al. 2012; Caradonna, Engel et al. 2013). The work presented here is the first simultaneous deep transcriptomic surveys of both *T. cruzi* and its human host cells across an intracellular infection time course. To enhance the comparison, we have obtained transcriptomic data for the invasive trypomastigote form as well as the axenic insect vector stage, the epimastigote form. The earliest infection time point analyzed, 4 hpi, reveals a substantial number of changes in steady state transcript abundance in both *T. cruzi* and its host. The parasite, having entered the host cell as highly motile, non-dividing trypomastigotes that express an abundance of several polymorphic surface antigens, quickly downregulates these surface antigens, along with motility functions as an early step in the transition to a non-motile amastigote. The surface antigens: *trans*-sialidases, mucins and MASPs are predominantly expressed in the trypomastigote form (including metacyclic trypomastigotes that arise from epimastigotes and not profiled here) and can mediate extracellular host-parasite interactions by triggering signaling

pathways in mammalian cells (Chuenkova, Furnari et al. 2001; Chuenkova and PereiraPerrin 2009). Consistent with the need for *T. cruzi* to prepare for replication in the host cell cytosol, there is evidence from the transcriptome of generalized upregulation of biosynthetic processes, occurring as early as 4 hpi. These include RNA, protein, fatty acid, isoprenoid and sterol synthesis. As the amastigote developmental program is pursued, DNA replication functions in the parasite are upregulated to prepare the amastigote for the initiation of S-phase at ~16 hpi and finally cell division at ~22 hpi. Strikingly, few additional changes in parasite gene expression occur between 24 - 72 hpi while the parasite is undergoing active replication. On the host side, parasite infection triggered an innate immune response within the first few hours of infection and maintained this feature throughout the course of infection. These findings differ markedly from those of early microarray studies, where significant upregulation of host genes in HFF was not observed in the first few hours of infection (up to 6 hpi) while the overall response at 24 hpi had a similar type I IFN signature (Vaena de Avalos, Blader et al. 2002). At the chronic phase of Chagas disease, patients often develop digestive or/and cardiac clinical manifestation, all of which can be linked to the attacks against the smooth muscle cells (SMC), combined with the inflammation caused by long-term infection (Lescure, Le Loup et al. 2010; Rassi, Rassi et al. 2010). However, the reason for the proneness of SMCs to the attack from *T. cruzi* and the mechanisms behind these symptoms were still unclear. Interestingly, our data has revealed that genes related to the modulation of cytosolic Ca_2^+ concentration were significantly upregulated in the late stages of

infection, which can further change the myosin-actin interactions and, hence, the development of contractile force of SMCs. In contrast to earlier microarray studies, deep sequencing methods used here have enabled us to detect early changes in host gene expression in response to parasite infection. These results provide a clear example of how transcriptomic profiling of intracellular *T. cruzi* amastigotes and human host can provide critical insights into the biochemical features of this pathogen and the identification of potential drug targets.

For both host and parasite, genes that shared dynamic expression patterns were clustered to reveal enriched functions and pathways. Protein-coding genes in *T. cruzi* do not have pol II promoters like the majority of eukaryotes and the regulation of gene expression levels occurs primarily at the post-transcriptional level (Teixeira, Kirchhoff et al. 1995; Teixeira and daRocha 2003; Martinez-Calvillo, Vizuet-de-Rueda et al. 2010). However, neither a genome-wide gene structure analysis nor a systematic search of regulatory elements has been done for *T. cruzi*. Through the use of ultra high-throughput sequencing, we completed the gene structure model of *T. cruzi* in Chapter 2. We also identified potential regulatory elements located in the UTR regions for each gene cluster in *T. cruzi*. The transcriptional and functional studies reported in this work provided a fundamental framework for various network analyses for both species, such as coexpression and regulatory networks, during the host-pathogen interaction.

Some strains of *T. cruzi* parasite (for example Y strain and CL-Brener strain) can cause life-threatening disease while others can be non-virulent (such as G strain). It has been proposed that differences in the pathogenic potential between disease isolates might be influenced by small genetic differences in genes from the core genome (de Freitas, Augusto-Pinto et al. 2006; Cortez, Martins et al. 2012). However, differences in the pathogenic potential of strains can be also attributed to differences at the transcriptional level that can be identified by performing comparative transcriptomic analyses of different parasite strains. We propose that the analysis of the transcriptional response of *T. cruzi* strains with/without the infection might represent a new approach to discriminate between the virulence potentials of different isolates.

In the field of *T. cruzi* pathogenesis, this first analysis of gene expression after invasion of human hosts significantly increases the knowledge of how these parasites survive and manipulate responds to human blood and causes sepsis. The findings reported in this study could also be helpful to identify the function of gene products annotated as hypothetical proteins, understand the regulation of vaccine antigens in blood and ultimately develop diagnostic and therapeutic strategies to control a neglected but devastating disease.

Chapter 4: Conclusion remarks and future perspective

Trypanosoma cruzi is a unicellular eukaryotic parasite with a digenetic life cycle alternating between the triatomine insect and a variety of mammalian hosts. It can cause Chagas disease, which is ranked as one of the seventeen neglected diseases by WHO. An estimated 10 million people suffer from this forgotten scourge (Rassi, Rassi et al. 2010). Most of them are the most vulnerable populations with limited resources and public voice. No vaccine is available and current treatments are very limited and highly toxic. This disease has significant economic impact in Latin America and has been classified by the World Bank as one of the main public health problems on South American continent (WHO 2002). Besides its importance as a human and veterinary pathogen, *T. cruzi*, together with sister species like *T. brucei* and *L. major,* has been key to the discovery and understanding of general biological principles such as *trans*-splicing, RNA editing, antigenic variation, and immune evasion (Borst and Cross 1982; Sutton and Boothroyd 1986; Simpson, Shaw et al. 1988; McCabe and Mullins 1990; Thomas, Martinez et al. 2007). In this study, we applied a high throughput RNA sequencing approach to investigate the transcriptome of the pathogen and infected human hosts simultaneously. In addition, we included the transcriptome of two extracellular stages of *T. cruzi* in order to capture its complete life cycle. In our experiments, a total of 2.7 billion reads from 34 samples across 6 intracellular stages and 2 extracellular stages were sequenced. The sequencing depth and single-nucleotide resolution allowed us to scrutinize both transcriptomes in unprecedented detail.

Unlike other eukaryotes, trypanosome genes coding for proteins with unrelated functions are organized into co-directional clusters that undergo polycistronic transcription by RNA polymerase II (Pol II) (El-Sayed, Myler et al. 2005; Siegel, Tan et al. 2005; Siegel, Gunasekera et al. 2011). *Trans*-splicing, together with polyadenylation, allows polycistronic transcripts to be processed into monocistronic units ready for translation (El-Sayed, Myler et al. 2005; Daniels, Gull et al. 2010).The lack of identifiable RNA pol II promoters in trypanosomatids suggests that these organisms lack precise transcriptional control over the majority of their genes (Siegel, Gunasekera et al. 2011). Thus, regulation of gene expression occurs mainly at the post-transcriptional level, through pre-mRNA processing, RNA degradation, or translational repression (Martinez-Calvillo, Vizuet-de-Rueda et al. 2010). To date, very little is known about the way in which trypanosomes regulate transcription or translation, but in many other eukaryotes, the 5' and 3' UTRs contain sequence elements that are responsible for regulating RNA or protein synthesis. So characterizing the RNA processing events in *T. cruzi* is of great significance in the efforts of understanding the biology of the pathogen.

Because our study design was mainly aimed at a quantitatively unbiased profiling of the transcriptome of the parasite at various developmental stages, we did not enrich for SL- or poly(T)-containing reads during library construction and relied

on deep coverage to collect relatively large numbers of reads from both ends of transcripts (6.4 million SL- and 1.1 million polyA-containing reads). This approach permitted us to identify as well as quantitate differential RNA processing events. Strikingly, our genome-wide study successfully identified *trans*-splicing and polyadenylation sites for more than 80% of the genes. A very high degree of heterogeneity of RNA processing sites were noted in *T. cruzi*. This phenomena were also observed in other trypanosomes as well. In addition, our experiment captured identical alternative *trans*-splicing sites for gene *LYT1* reported by Swindle's team using PCR amplification of reverse transcribed mRNA in 2002. This further confirmed the accuracy and sensitivity of our high throughput sequencing approach. Most surprising was that a large number of transcripts showed differential abundance of alternative splice variants across different life stages. More than 198 differentially spliced transcripts were detected between amastigote 72 hpi, epimastigote and trypomastigote supporting the idea that alternative splicing may have functional consequences for the regulation of parasite survival and development, of which the corresponding regulatory factors remain elusive. Alternative splicing patterns have been observed in *T. brucei* during development as well (Nilsson, Gunasekera et al. 2010). Four possible functional consequences can be summarized: First, alternative splicing can result in potentially untranslatable transcript with incomplete ORFs; second, it can lead to different length of transcripts with the inclusion or exclusion of signal peptide or other anchoring signals present as a potential mechanism for dual localization of proteins; third, small regulatory upstream ORFs may be included in the longer

variant of the transcripts; fourth, splicing events may allow usage of a novel ORF because of the inclusion of alternative start codon. More questions can be asked with regards to the different splicing patterns: if the splicing events are regulated; if yes, how does the splicing machinery adapt under different developmental stages; if the resulting variants are targeted to different cellular compartments or possess altered stabilities; if the inclusion of some sequence motifs on the variants can lead to different translational efficiency. The identification of potential sequence motifs located in the 5' or 3' UTRs, particularly those that will be selectively included through different splicing events can be very helpful in explaining the function of alternative splicing.

Several studies have been conducted with the oligonucleotide microarray technique (microarray) to investigate either the transcriptome of *T. cruzi* or infected human cells (Shigihara, Hashimoto et al. 2008; Minning, Weatherly et al. 2009; Zhang, Kim et al. 2010; Manque, Probst et al. 2011; Tanowitz, Mukhopadhyay et al. 2011; Grynberg, Passos-Silva et al. 2012; Caradonna, Engel et al. 2013). The research group of Tarleton have conducted genome-wide expression profile analyses, revealing differentially expressed genes at four life-cycle stages of *T. cruzi* (Minning, Weatherly et al. 2009). In another study, researchers investigated the transcriptome of *T. cruzi*-infected HeLa cells and detected 64 genes with three-fold changes (Shigihara, Hashimoto et al. 2008). However, these studies only examined gene expressions in either *T. cruzi* or infected human cells as a ratio of expression levels and none of them captured

the transcriptomes across the infection cycle which are essential for the understanding of host-pathogen interaction. The expression profiles of both *T. cruzi* and human genes in a spatially and temporally equivalent biological sample could help us to understand the molecular mechanisms of both host defense and pathogen infection strategies simultaneously. To date, our study is the first one analyzing the genome-wide expression profiles of both the human and *T. cruzi* genes simultaneously in the same infected sample with RNA-Seq technology. Compared with microarrays, RNA-Seq is known to have a wider dynamic range, higher technical reproducibility, and provide a better estimate of absolute expression levels (Marioni, Mason et al. 2008; Fu, Fu et al. 2009).

RNA-Seq, like many other technologies applied in biology, requires a complicated set of reagents, hardware and highly trained personnel to yield accurate and replicable results. Batch effects were defined as sub-groups of measurements that have qualitatively inconsistent behavior across conditions and are irrelevant to the biological factors investigated in a study (Leek, Scharpf et al. 2010). These kind of technical effects can originate from various laboratory conditions, reagent brands and lots, equipment, as well as personnel differences, If not properly handled, batch effects can be a common and powerful source of variation, making the task of combining data from different batches difficult. Several examples have been reported in published studies, which biological variables were impacted by batch effects and raised concerns about the validity of conclusions (Baggerly, Edmonson et al. 2004; Akey, Biswas et al. 2007; Leek,

Scharpf et al. 2010). Currently, several statistical solutions have been proposed to adjust for batch effect, including surrogate variable analysis (SVA) (Leek and Storey 2007) and Combat (Johnson, Li et al. 2007). In our study, samples were prepared by the same laboratory but the infection experiments were carried out on five different dates spanning a year period. So here we considered experiment date as a distinctive batch source. We quantified and visualized batch effects using both PCA and hierarchical clustering dendrograms (**Figure**s **36, 37, 38**). We did not observe strong batch effects, but a mild one: samples were clustered by developmental stages rather than batches; principal components were associated with biological conditions, rather than batch effects. However, in our samples from certain time points, such as trypomastigote or epimastigote, batch effects were highly cofounded with biological variables of interest (developmental stages and/or infection). This correlation makes it difficult to determine whether observed differences across biological groups are due to biology or artifacts. In addition, because of this unbalanced batch design of our samples, particularly the extracellular *T. cruzi* samples, we did not apply Combat to remove batch effects to prevent the overcorrection or bias introduced by the package. Instead, we incorporated batch factor as a variable in the linear model that were used in statistical analyses. In future studies with high throughput data, especially for those dealing with time series, a more balanced experiment design may minimize the batch effects or facilitate the correction of potential technical variation. Close collaborations between biologists and statisticians are also

needed so that the specific sources of batch effects can be isolated and the dependence on surrogates can be reduced.

Complementary research that can be done after this study is network analysis. The establishment of transcriptome profiles for both *T. cruzi* and human will allow researchers to study gene coexpression, the process by which genes are expressed in coordination to produce proteins. For example, weighted gene correlation network analysis (WGCNA) can be used for finding modules of highly correlated genes and for summarizing these clusters of genes by identifying their eigengene (Langfelder and Horvath 2008). Network analyses have been proposed as a solution to systems biology studies, particularly those involving transcriptomic datasets(Zhang and Horvath 2005). This approach can model the interactions of real biological networks and can intuitively visualize the relationships between modules (Aoki, Ogata et al. 2007; Horvath and Dong 2008). The clustering of co-expressed genes into "modules" mirrors regulatory associations found in biological systems and provides information on unknown genes through "guilt by association" with well-characterized ones in the same cluster (Zhang and Horvath 2005). Motif findings can also be carried out in genes that share similar expression patterns (Wernicke and Rasche 2006). In particular, the highly connected “hubs” identified in *T. cruzi* coexpression networks may play an essential role either by influencing the expression patterns of other genes or alternatively communicating changes that occur elsewhere in the network and may present as a potential drug target.

With the rapid drop of sequencing cost, comparative sequencing has entered a new era. A wide range of evolutionary and pathological questions within the *T. cruz*i lineage can be answered by sequencing analyses of additional *T. cruzi* isolates. In this study, we constructed the transcriptome of *T. cruzi* Y strain (DTU II), which is one of the most virulent strains and often proposes as a lethal challenge. In contrast, BALB/c mice injected with CL-14 trypomastigotes, a non-virulent strain, showed no or limited parasitaemia but high level of resistance (Lima, Lenzi et al. 1995). Comparisons between the transcriptomes of Y strain and non-virulent strains like CL-14 or G strain may help identify the key factors involved in the pathogenesis of Chagas and generate an interesting gene list that may restore the virulence of CL-14. In addition, we can compare the usage of different RNA processing sites between strains. The conservation of site usage indicates the functional consequences of the RNA processing sites and may help identify motifs involved in these choices; whereas, different *trans*-splicing sites, resulting in transcripts with different lengths, may propose as an explanation for the virulence of Y strain.

Ribosomal profiling (Ribo-Seq) technology can be a good complement to RNA-Seq experiments helping us understand both *trans*-splicing and polyadenylation events. Ribo-Seq produces a global picture of the positions of all the active ribosomes in a cell, thus allowing us to accurately pinpoint the translational start site and evaluate the translational efficiency. This technology, together with RNA-

Seq, can help us examine whether alternative *trans*-splicing events, or the presence of certain motifs, can lead to different translational start site, thus various protein products. In addition, it can allow us to correctly identify the start codon for annotated and novel ORFs. It will also be of particular interest to investigate the association of alternative RNA processing events with translational efficiency. If a positive correlation can be established between the two, the diverse splicing patterns detected at different developmental stages could be an efficient approach by *T. cruzi* to adjust its protein levels across time. Other genomic features, for example the spacing between *trans*-splicing sites and upstream polyadenylation site or the length of UTRs, may also play a role in the regulation of alternative translation start site or translational efficiency. Ribo-Seq may help in the elucidation of the above questions. The combination of RNA-sequencing and ribosome profiling technologies can bridge the gap between gene express level and protein synthesis (Ingolia 2014). A future direction and good complement of this study is to construct the ribosome profiling (Ribo-Seq) for *T. cruzi* across its development stages. Since trypanosomes do not follow a rigid regulation at the transcriptional level, it will be of particular interests to generate global measurements of translation that are as precise and detailed as data derived from RNA-Seq. Ribo-Seq can be used to identify translated sequences within the profiled parasite transcriptome, to monitor the process of translation and the maturation of nascent polypeptides *in vivo*, and to quantify profiles of cellular protein synthesis. The correlation and discrepancy

between transcriptional level and translational level can provide new insights into the posttranscriptional mechanism exploited by the parasite.

Acknowledgement

We thank Dr. Barbara Burleigh from the school of public health at Harvard University for extracting and providing RNA samples for our experiments; Dr. Hector Corrada-Bravo from the Computer Science for helping us improve the statistical pipelines applied.

The work was supported by the National Institutes of Health AI094773 and AI094195.

Appendices

Appendix 1

Table 6 Summary of experiment design and treatments.

Lab sample ID	Developmental stage	Infected	Batch	Trimmed
HPGL0063	Amastigote 4 hpi	N	A	
HPGL0064		Y	A	Y
HPGL0111		Y	B	
HPGL0112		Y	C	
HPGL0065	Amastigote 6hr	N	A	
HPGL0066		Y	A	Y
HPGL0113		Y	B	
HPGL0114		Y	C	Y
HPGL0067	Amastigote 12hr	N	A	
HPGL0068		Y	A	Y
HPGL0115		Y	B	Y
HPGL0069	Amastigote 24hr	N	A	
HPGL0070		Y	A	Y
HPGL0116		Y	B	Y
HPGL0117		Y	C	Y
HPGL0055	Amastigote 48hr	N	D	
HPGL0071		N	A	Y
HPGL0257		N	C	
HPGL0056		Y	D	
HPGL0060		Y	D	
HPGL0072		Y	A	Y
HPGL0258		Y	C	
HPGL0057	Amastigote 72hr	N	D	
HPGL0255		N	E	
HPGL0058		Y	D	

HPGL0061		Y	D	
HPGL0256		Y	E	
HPGL0062	trypomastigote	N	A	Y
HPGL0249		N	E	
HPGL0250		N	E	
HPGL0251		N	E	
HPGL0252	epimastigote	N	E	
HPGL0253		N	E	
HPGL0254		N	E	

Appendix 2

Table 7 Summary of total number of reads mapped to the reference genomes.

Lab sample ID	Number of reads that pass Illumina filter	Total number of reads mapped	% of total reads mapped
HPGL0063	75,455,946	72,024,476	95.45%
HPGL0064	139,558,460	130,016,017	93.16%
HPGL0111	123,249,854	81,724,984	66.31%
HPGL0112	98,909,298	94,669,006	95.71%
HPGL0065	63,327,220	60,735,117	95.91%
HPGL0066	157,401,226	148,513,778	94.35%
HPGL0113	112,249,358	106,802,654	95.15%
HPGL0114	97,411,582	91,909,631	94.35%
HPGL0067	67,337,918	64,902,003	96.38%
HPGL0068	88,441,820	83,974,766	94.95%
HPGL0115	93,933,948	89,531,503	95.31%
HPGL0069	66,477,092	64,617,122	97.20%
HPGL0070	88,843,990	84,786,038	95.43%
HPGL0116	79,347,540	73,605,848	92.76%

HPGL0117	74,297,462	70,880,898	95.40%
HPGL0055	57,701,782	48,117,142	83.39%
HPGL0071	69,490,186	68,360,413	98.37%
HPGL0257	91,887,850	90,007,520	97.95%
HPGL0056	57,701,782	52,398,585	90.81%
HPGL0060	57,692,178	52,602,786	91.18%
HPGL0072	77,919,340	70,119,342	89.99%
HPGL0258	76,483,856	67,999,152	88.91%
HPGL0057	60,211,156	58,562,661	97.26%
HPGL0255	77,417,180	76,347,021	98.62%
HPGL0058	49,143,458	40,315,329	82.04%
HPGL0061	65,873,648	54,571,365	82.84%
HPGL0256	90,450,622	87,107,335	96.30%
HPGL0062	105,944,210	65,836,443	62.14%
HPGL0249	59,465,792	37,831,725	63.62%
HPGL0250	55,394,756	36,057,541	65.09%
HPGL0251	54,424,652	35,121,149	64.53%
HPGL0252	65,967,456	37,693,521	57.14%

HPGL0253	59,156,660	34,484,732	58.29%
HPGL0254	52,314,436	29,335,454	56.08%
Total	2,710,883,714	2,361,563,057	87.11%

Appendix 3

Table 8 Summary of mapping statistics to individual genomes

Lab sample ID	Reads mapped to *T. cruzi* Esmeraldo haplotype	Reads mapped to hg19	% of reads mapped to *T. cruzi* Esmeraldo haplotype	% of reads mapped to hg19
HPGL0063		72,024,476		100.00%
HPGL0064	2,585,633	127,430,384	1.99%	98.01%
HPGL0111	2,257,327	79,467,657	2.76%	97.24%
HPGL0112	2,928,987	91,740,019	3.09%	96.91%
HPGL0065		60,735,117	0.00%	100.00%
HPGL0066	3,128,883	145,384,895	2.11%	97.89%
HPGL0113	2,014,693	104,787,961	1.89%	98.11%
HPGL0114	2,811,907	89,097,724	3.06%	96.94%
HPGL0067		64,902,003	0.00%	100.00%
HPGL0068	1,769,313	82,205,453	2.11%	97.89%
HPGL0115	2,136,351	87,395,152	2.39%	97.61%
HPGL0069		64,617,122	0.00%	100.00%
HPGL0070	4,635,117	80,150,921	5.47%	94.53%
HPGL0116	3,903,274	69,702,574	5.30%	94.70%

HPGL0117	4,982,378	65,898,520	7.03%	92.97%
HPGL0055		48,117,142	0.00%	100.00%
HPGL0071		68,360,413	0.00%	100.00%
HPGL0257		90,007,520	0.00%	100.00%
HPGL0056	6,866,656	45,531,929	13.10%	86.90%
HPGL0060	6,985,398	45,617,388	13.28%	86.72%
HPGL0072	13,138,673	56,980,669	18.74%	81.26%
HPGL0258	19,099,189	48,899,963	28.09%	71.91%
HPGL0057		58,562,661	0.00%	100.00%
HPGL0255		76,347,021	0.00%	100.00%
HPGL0058	15,056,971	25,258,358	37.35%	62.65%
HPGL0061	19,546,061	35,025,304	35.82%	64.18%
HPGL0256	8,327,776	78,779,559	9.56%	90.44%
HPGL0062	65,836,443		100.00%	
HPGL0249	37,831,725		100.00%	
HPGL0250	36,057,541		100.00%	
HPGL0251	35,121,149		100.00%	
HPGL0252	37,693,521		100.00%	
HPGL0253	34,484,732		100.00%	
HPGL0254	29,335,454		100.00%	

Total	398,535,152	1,963,027,905	14.70%	72.41%

Appendix 4: Supplementary table index

All supplementary tables were submitted together with this thesis.

Table S1. Summary of features for RNA processing events. This table contains the number of *T. cruzi* reads, SL-containing reads, poly(T)-containing reads, and the number of RNA-processing sites detected.

Table S2. *T. cruzi* 5' UTR coordinates ORF-containing 5' UTRs excluded. This table includes the coordinates of all 5' UTRs detected for annotated CDSs and the number of SL-containing reads mapped to corresponding sites at each time point. UTRs containing open reading frames were excluded from output.

Table S3. *T. cruzi* 3' UTR coordinates ORF-containing 3' UTRs excluded. This table includes the coordinates of all 3' UTRs detected for annotated CDSs and the number of poly(T)-containing reads mapped to corresponding sites at each time point. UTRs containing open reading frames were excluded from output.

Table S4. Function distribution of genomic features with extreme values. This table includes the function frequency of genes with extreme values of genomic features for *T. cruzi* and *T. brucei*. The data from *T. brucei* were extracted from published dataset.

Table S5. Coordinates of 5' UTR for all CDSs existing and novel. This table includes the coordinates of all 5' UTRs detected for both annotated and novel CDSs and the number of SL-containing reads mapped to corresponding sites at each time point.

Table S6. Coordinates of 3' UTR for all CDSs existing and novel. This table includes the coordinates of all 3' UTRs detected for both annotated and novel CDSs and the number of poly(T)-containing reads mapped to corresponding sites at each time point.

Table S7. Characterization of novel ORFs. This table contains detailed information of novel identified ORFs, including strand, boundary coordinates of ORFs and transcripts, whether or not the signal peptide, transmembrane domain, or GPI-anchor signal was present, CDD output.

Table S8. Putative alternative start codon in *T. cruzi.* This table includes the coordinates of all putative alternative start codon.

Table S9. Acceptor sequence usage for primary and alternative *trans*-splicing sites. This table includes the frequency of acceptor sequence applied at primary or minor *trans*-splicing sites.

Table S10. *T. cruzi trans*-splicing site detected in non-CDS region. This table includes the coordinates of *trans*-splicing sites detected inside annotated CDSs and the number of SL-containing reads mapped to corresponding sites at each time point.

Table S11. Genes with different primary *trans*-splicing site usage across stages. This table includes genes that applied different primary *trans*-splicing site at different developmental stages.

Table S12. Summary of alternative splicing events of *T. cruzi* at stages of trypomastigote, epimastigote and amastigote 72 hpi. This table is a quantitative presentation of Figure 4, containing the characterization of alternative *trans*-splicing events for each gene, including the number of raw and normalized SL-containing reads that mapped to the primary and secondary sites, the ratio of P/S at the stages of trypomastigote, epimastigote and amastigote 72 hpi, the status of alternative splicing, whether there are more than 3 SL-containing reads mapped to the primary sites, whether P/S >1.6, whether primary site switching was present, and whether the secondary site was at least 20 nt away from the primary site if existing.

Table S13. *T. cruzi* 5' UTR coordinates ORF-containing 5' UTRs and alternative 5UTRs located less than 20nt excluded. This table includes the coordinates of all 3' UTRs detected and the number of SL-containing reads

mapped to corresponding sites at each time point. UTRs containing open reading frames or alternative *trans*-splicing sites less than 20 nt from the primary one were excluded from output.

Table S14. Global gene expression profiles of *T. cruzi* and HFF. This table displays quantitative data that were used to plot Figure 34.

Table S15. Summary of number of DEGs. Table showing number of differentially expressed genes each comparison for human and pathogen (q-value < 0.05).

Table S16. Summary of differential expression analysis for human genes. Each sheet reports differential expression results for a specific comparison. The columns include gene id, gene function, log2 fold change, P value, q-value, adjusted P value, average expression value, and whether the q-value or adjusted P value < 0.05 (with Qvals or AdjPvals = 1) .

Table S17. Summary of differential expression analysis for human genes with indicator of both q-value and fold change. Each sheet reports differential expression results for a specific comparison. The columns include gene id, gene function, log2 fold change, P value, q-value, adjusted P value, average expression value, and whether the q-value or adjusted P value < 0.05 (with Qvals or AdjPvals = 1).

Table S18. Summary of differential expression analysis for *T. cruzi* genes. Each sheet reports differential expression results for a specific comparison. The columns include gene id, gene function, log2 fold change, P value, q-value, adjusted P value, average expression value, and whether the q-value or adjusted P value < 0.05 (with Qvals or AdjPvals = 1) .

Table S19. Summary of differential expression analysis for *T. cruzi* genes with indicator of both q-value and fold change. Each sheet reports differential expression results for a specific comparison. The columns include gene id, gene function, log2 fold change, P value, q-value, adjusted P value, average expression value, and whether the q-value or adjusted P value < 0.05 (with Qvals or AdjPvals = 1).

Table S20. Overrepresented *T. cruzi* genes in a single life stage. In this table, genes that were detected as highly expressed in a single life stage are represented. Each sheet reports signature genes for amastigote, trypomastigote and epimastigote (with or without the limitation of fold change).

Table S21. Enriched Gene Ontology terms for *T. cruzi* DEGs. This dataset contains enriched gene functions for each gene cluster, including the information of GO category, over and underrepresented P value, number of gene in the cluster that were present in the category, number of genes in the category,

adjusted P. value, description of the GO category. Each sheet reports overrepresented GO terms under biological process (BP), molecular functions (MF), and cellular components (CC) for upregulated or downregulated genes.

Table S22. Enriched Gene Ontology terms for each human DEGs. This dataset contains enriched gene functions for each gene cluster, including the information of GO category, over and underrepresented P value, number of gene in the cluster that were present in the category, number of genes in the category, adjusted P. value, description of the GO category.

Table S23. Enriched Gene Ontology terms for each *T. cruzi* K-mean cluster. This dataset contains enriched gene functions for each gene cluster, including the information of GO category, over and underrepresented P value, number of gene in the cluster that were present in the category, number of genes in the category, adjusted P. value, description of the GO category.

Table S24. Enriched Gene Ontology terms for each human K-mean cluster. This dataset contains enriched gene functions for each gene cluster, including the information of GO category, over and underrepresented P value, number of gene in the cluster that were present in the category, number of genes in the category, adjusted P. value, description of the GO category.

Table S25. Motifs identified in the 5’ UTR of K-mean cluster that overlapping with the 53 motifs identified in KEGG gene clusters. We compared motifs detected in the 5’ UTRs of gene clusters detected in our K-mean cluster analysis to those that have been identified in KEGG gene cluster. The overlapping motifs were indicated by brackets.

Table S26. Motifs identified in the 3’ UTR of K-mean cluster that overlapping with the 53 motifs identified in KEGG gene clusters. We compared motifs detected in the 3’ UTRs of gene clusters detected in our K-mean cluster analysis to those that have been identified in KEGG gene cluster. The overlapping motifs were indicated by brackets.

Table S27. Intersection between human DEGs and positive results from siRNA screening. This table contains genes that were detected differentially expressed and were positive in primary siRNA screening. Each sheet reports the overlapping DEGs from each comparison and the siRNA results.

Bibliography

(2005). "[Brazilian Consensus on Chagas disease]." Revista da Sociedade Brasileira de Medicina Tropical **38 Suppl 3**: 7-29.

Abdel-Latif, A. A. (2001). "Cross talk between cyclic nucleotides and polyphosphoinositide hydrolysis, protein kinases, and contraction in smooth muscle." Experimental biology and medicine **226**(3): 153-163.

Abrahamsohn, I. A. and R. L. Coffman (1996). "Trypanosoma cruzi: IL-10, TNF, IFN-gamma, and IL-12 regulate innate and acquired immunity to infection." Experimental parasitology **84**(2): 231-244.

Afonso, A. M., M. H. Ebell, et al. (2012). "A systematic review of high quality diagnostic tests for Chagas disease." PLoS neglected tropical diseases **6**(11): e1881.

Agabian, N. (1990). "Trans splicing of nuclear pre-mRNAs." Cell **61**(7): 1157-1160.

Akey, J. M., S. Biswas, et al. (2007). "On the design and analysis of gene expression studies in human populations." Nature genetics **39**(7): 807-808; author reply 808-809.

Aliberti, J. C., M. A. Cardoso, et al. (1996). "Interleukin-12 mediates resistance to Trypanosoma cruzi in mice and is produced by murine macrophages in response to live trypomastigotes." Infection and immunity **64**(6): 1961-1967.

Anders, S., D. J. McCarthy, et al. (2013). "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor." Nature protocols **8**(9): 1765-1786.

Andersson, B. (2011). "The Trypanosoma cruzi genome; conserved core genes and extremely variable surface molecule families." Research in microbiology **162**(6): 619-625.

Andrade, Z. A. (1983). "Mechanisms of myocardial damage in Trypanosoma cruzi infection." Ciba Foundation symposium **99**: 214-233.

Aoki, K., Y. Ogata, et al. (2007). "Approaches for extracting practical information from gene co-expression networks in plant biology." Plant & cell physiology **48**(3): 381-390.

Araujo, A. F., B. C. de Alencar, et al. (2005). "CD8+-T-cell-dependent control of Trypanosoma cruzi infection in a highly susceptible mouse strain after immunization with recombinant proteins based on amastigote surface protein 2." Infection and immunity **73**(9): 6017-6025.

Archer, S. K., D. Inchaustegui, et al. (2011). "The Cell Cycle Regulated Transcriptome of Trypanosoma brucei." PLoS One **6**(3): e18425.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics **25**(1): 25-29.

Aufderheide, A. C., W. Salo, et al. (2004). "A 9,000-year record of Chagas' disease." Proceedings of the National Academy of Sciences of the United States of America **101**(7): 2034-2039.

Baez, A. L., M. S. Lo Presti, et al. (2013). "Chronic indeterminate phase of Chagas' disease: mitochondrial involvement in infection with two strains." Parasitology **140**(3): 414-421.

Baggerly, K. A., S. R. Edmonson, et al. (2004). "High-resolution serum proteomic patterns for ovarian cancer detection." Endocrine-related cancer **11**(4): 583-584; author reply 585-587.

Bailey, J. A., R. Baertsch, et al. (2004). "Hotspots of mammalian chromosomal evolution." Genome biology **5**(4): R23.

Bao, Y., L. M. Weiss, et al. (2009). "Protein kinase A regulatory subunit interacts with P-Type ATPases in Trypanosoma cruzi." The American journal of tropical medicine and hygiene **80**(6): 941-943.

Belaunzaran, M. L., S. E. Wilkowsky, et al. (2013). "Phospholipase A1: a novel virulence factor in Trypanosoma cruzi." Molecular and biochemical parasitology **187**(2): 77-86.

Benabdellah, K., E. Gonzalez-Rey, et al. (2007). "Alternative trans-splicing of the Trypanosoma cruzi LYT1 gene transcript results in compartmental and functional switch for the encoded protein." Mol Microbiol **65**(6): 1559-1567.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.

Benz, C., D. Nilsson, et al. (2005). "Messenger RNA processing sites in Trypanosoma brucei." Mol Biochem Parasitol **143**(2): 125-134.

Benz, C., D. Nilsson, et al. (2005). "Messenger RNA processing sites in Trypanosoma brucei." Molecular and biochemical parasitology **143**(2): 125-134.

Bern, C. and S. P. Montgomery (2009). "An estimate of the burden of Chagas disease in the United States." Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **49**(5): e52-54.

Bernstein, R. E. (1984). "Darwin's illness: Chagas' disease resurgens." Journal of the Royal Society of Medicine **77**(7): 608-609.

Berriman, M., E. Ghedin, et al. (2005). "The genome of the African trypanosome Trypanosoma brucei." Science **309**(5733): 416-422.

Bloom, J. S., Z. Khan, et al. (2009). "Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays." BMC Genomics **10**: 221.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Borst, P. and G. A. Cross (1982). "Molecular basis for trypanosome antigenic variation." Cell **29**(2): 291-303.

Bradford, J. R., Y. Hey, et al. (2010). "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling." BMC Genomics **11**: 282.

Brasil, P. E., L. De Castro, et al. (2010). "ELISA versus PCR for diagnosis of chronic Chagas disease: systematic review and meta-analysis." BMC infectious diseases **10**: 337.

Brener, Z. (1971). "Life cycle of Trypanosoma cruzi." Revista do Instituto de Medicina Tropical de Sao Paulo **13**(3): 171-178.

Brener, Z. (1973). "Biology of Trypanosoma cruzi." Annual review of microbiology **27**: 347-382.

Brisse, S., J. Henriksson, et al. (2003). "Evidence for genetic exchange and hybridization in Trypanosoma cruzi based on nucleotide sequences and molecular karyotype." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **2**(3): 173-183.

Britto, C., A. Cardoso, et al. (1999). "Polymerase chain reaction (PCR) as a laboratory tool for the evaluation of the parasitological cure in Chagas disease after specific treatment." Medicina **59 Suppl 2**: 176-178.

Camargo, E. P. (2009). "Perspectives of vaccination in Chagas disease revisited." Memorias do Instituto Oswaldo Cruz **104 Suppl 1**: 275-280.

Campbell, D. A., S. J. Westenberger, et al. (2004). "The determinants of Chagas disease: connecting parasite and host genetics." Current molecular medicine **4**(6): 549-562.

Caradonna, K. L., J. C. Engel, et al. (2013). "Host metabolism regulates intracellular growth of Trypanosoma cruzi." Cell host & microbe **13**(1): 108-117.

Caradonna, K. L., J. C. Engel, et al. (2013). "Host metabolism regulates intracellular growth of Trypanosoma cruzi." Cell Host Microbe **13**(1): 108-117.

Castro, J. A., M. M. de Mecca, et al. (2006). "Toxic side effects of drugs used to treat Chagas' disease (American trypanosomiasis)." Human & experimental toxicology **25**(8): 471-479.

Cerqueira, G. C., D. C. Bartholomeu, et al. (2008). "Sequence diversity and evolution of multigene families in Trypanosoma cruzi." Molecular and biochemical parasitology **157**(1): 65-72.

Chen, C. Y. and A. B. Shyu (1995). "AU-rich elements: characterization and importance in mRNA degradation." Trends in biochemical sciences **20**(11): 465-470.

Chessler, A. D., M. Unnikrishnan, et al. (2009). "Trypanosoma cruzi triggers an early type I IFN response in vivo at the site of intradermal infection." Journal of immunology **182**(4): 2288-2296.

Chuenkova, M. V., F. B. Furnari, et al. (2001). "Trypanosoma cruzi trans-sialidase: a potent and specific survival factor for human Schwann cells by means of phosphatidylinositol 3-kinase/Akt signaling." Proceedings of the National Academy of Sciences of the United States of America **98**(17): 9936-9941.

Chuenkova, M. V. and M. PereiraPerrin (2009). "Trypanosoma cruzi targets Akt in host cells as an intracellular antiapoptotic strategy." Science signaling **2**(97): ra74.

Cloonan, N. and S. M. Grimmond (2008). "Transcriptome content and dynamics at single-nucleotide resolution." Genome Biol **9**(9): 234.

Cortez, C., R. M. Martins, et al. (2012). "Differential infectivity by the oral route of Trypanosoma cruzi lineages derived from Y strain." PLoS Negl Trop Dis **6**(10): e1804.

Costales, J. A., J. P. Daily, et al. (2009). "Cytokine-dependent and-independent gene expression changes and cell cycle block revealed in Trypanosoma cruzi-infected host cells by comparative mRNA profiling." BMC Genomics **10**: 252.

Crampton, A. and T. Vanniasinkam (2007). "Parasite vaccines: the new generation." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **7**(5): 664-673.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome research **14**(6): 1188-1190.

Cruz, M. C., N. Souza-Melo, et al. (2012). "Trypanosoma cruzi: role of delta-amastin on extracellular amastigote cell invasion and differentiation." PLoS One **7**(12): e51804.

Daniels, J. P., K. Gull, et al. (2010). "Cell biology of the trypanosome genome." Microbiol Mol Biol Rev **74**(4): 552-569.

de Freitas, J. M., L. Augusto-Pinto, et al. (2006). "Ancestral genomes, sex, and the population structure of Trypanosoma cruzi." PLoS pathogens **2**(3): e24.

de Freitas, J. M., L. Augusto-Pinto, et al. (2006). "Ancestral genomes, sex, and the population structure of Trypanosoma cruzi." PLoS Pathog **2**(3): e24.

De Gaudenzi, J. G., S. J. Carmona, et al. (2013). "Genome-wide analysis of 3'-untranslated regions supports the existence of post-transcriptional regulons controlling gene expression in trypanosomes." PeerJ **1**: e118.

Dias, E., F. S. Laranja, et al. (1956). "Chagas' disease; a clinical, epidemiologic, and pathologic study." Circulation **14**(6): 1035-1060.

Dubner, S., E. Schapachnik, et al. (2008). "Chagas disease: state-of-the-art of diagnosis and management." Cardiology journal **15**(6): 493-504.

El-Sayed, N. M., P. J. Myler, et al. (2005). "The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease." Science **309**(5733): 409-415.

El-Sayed, N. M., P. J. Myler, et al. (2005). "Comparative genomics of trypanosomatid parasitic protozoa." Science **309**(5733): 404-409.

Elkon, R., A. P. Ugalde, et al. (2013). "Alternative cleavage and polyadenylation: extent, regulation and function." Nature reviews. Genetics **14**(7): 496-506.

Euskirchen, G. M., J. S. Rozowsky, et al. (2007). "Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies." Genome Res **17**(6): 898-909.

Flores-Lopez, C. A. and C. A. Machado (2011). "Analyses of 32 loci clarify phylogenetic relationships among Trypanosoma cruzi lineages and support a single hybridization prior to human contact." PLoS Negl Trop Dis **5**(8): e1272.

Franzen, O., S. Ochaya, et al. (2011). "Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener." PLoS neglected tropical diseases **5**(3): e984.

Fu, X., N. Fu, et al. (2009). "Estimating accuracy of RNA-Seq and microarrays with proteomics." BMC Genomics **10**: 161.

Garcia, S. B., A. L. Aranha, et al. (2003). "A retrospective study of histopathological findings in 894 cases of megacolon: what is the relationship between megacolon and colonic cancer?" Revista do Instituto de Medicina Tropical de Sao Paulo **45**(2): 91-93.

Gaunt, M. W., M. Yeo, et al. (2003). "Mechanism of genetic exchange in American trypanosomes." Nature **421**(6926): 936-939.

Gerber, A. P., D. Herschlag, et al. (2004). "Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast." PLoS Biol **2**(3): E79.

Gibson, W., M. Crow, et al. (1997). "Kinetoplast DNA minicircles are inherited from both parents in genetic crosses of Trypanosoma brucei." Parasitology research **83**(5): 483-488.

Gimelli, G., M. A. Pujana, et al. (2003). "Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions." Human molecular genetics **12**(8): 849-858.

Gish, W. and D. J. States (1993). "Identification of protein coding regions by database similarity search." Nature genetics **3**(3): 266-272.

Gomes, Y. M., V. M. Lorena, et al. (2009). "Diagnosis of Chagas disease: what has been achieved? What remains to be done with regard to diagnosis and follow up studies?" Memorias do Instituto Oswaldo Cruz **104 Suppl 1**: 115-121.

Gonzales-Perdomo, M., P. Romero, et al. (1988). "Cyclic AMP and adenylate cyclase activators stimulate Trypanosoma cruzi differentiation." Experimental parasitology **66**(2): 205-212.

Graber, J. H., J. Salisbury, et al. (2007). "C. elegans sequences that control trans-splicing and operon pre-mRNA processing." RNA **13**(9): 1409-1426.

Greif, G., M. P. de Leon, et al. (2013). "Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax." BMC Genomics **14**(1): 149.

Griffith, M., O. L. Griffith, et al. (2010). "Alternative expression analysis by RNA sequencing." Nat Methods **7**(10): 843-847.

Grynberg, P., D. G. Passos-Silva, et al. (2012). "Trypanosoma cruzi gene expression in response to gamma radiation." PLoS One **7**(1): e29596.

Gunzl, A. (2010). "The pre-mRNA splicing machinery of trypanosomes: complex or simplified?" Eukaryot Cell **9**(8): 1159-1170.

Gurtler, R. E., U. Kitron, et al. (2007). "Sustainable vector control and management of Chagas disease in the Gran Chaco, Argentina." Proceedings of the National Academy of Sciences of the United States of America **104**(41): 16194-16199.

Guther, M. L., A. R. Prescott, et al. (2003). "Deletion of the GPIdeAc gene alters the location and fate of glycosylphosphatidylinositol precursors in Trypanosoma brucei." Biochemistry **42**(49): 14532-14540.

Hajduk, S. L., M. E. Harris, et al. (1993). "RNA editing in kinetoplastid mitochondria." FASEB journal : official publication of the Federation of American Societies for Experimental Biology **7**(1): 54-63.

Hansen, K. D., S. E. Brenner, et al. (2010). "Biases in Illumina transcriptome sequencing caused by random hexamer priming." Nucleic Acids Res **38**(12): e131.

Hartigan, J. A. a. W., M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1): 100-108.

Hirsh, D. and X. Y. Huang (1990). "Trans-splicing and SL RNAs in C. elegans." Molecular biology reports **14**(2-3): 115.

Hong, Y., K. Nagamune, et al. (2006). "Removal or maintenance of inositol-linked acyl chain in glycosylphosphatidylinositol is critical in trypanosome life cycle." The Journal of biological chemistry **281**(17): 11595-11602.

Horiuchi, T. and T. Aigaki (2006). "Alternative trans-splicing: a novel mode of pre-mRNA processing." Biology of the cell / under the auspices of the European Cell Biology Organization **98**(2): 135-140.

Horvath, S. and J. Dong (2008). "Geometric interpretation of gene coexpression network analysis." PLoS computational biology **4**(8): e1000117.

Hotez, P. J., E. Dumonteil, et al. (2013). "Innovation for the 'bottom 100 million': eliminating neglected tropical diseases in the Americas." Advances in experimental medicine and biology **764**: 1-12.

Huang, J. and L. H. Van der Ploeg (1991). "Requirement of a polypyrimidine tract for trans-splicing in trypanosomes: discriminating the PARP promoter from the immediately adjacent 3' splice acceptor site." The EMBO journal **10**(12): 3877-3885.

Ingolia, N. T. (2014). "Ribosome profiling: new views of translation, from single codons to genome scale." Nature reviews. Genetics **15**(3): 205-213.

Ivens, A. C., C. S. Peacock, et al. (2005). "The genome of the kinetoplastid parasite, Leishmania major." Science **309**(5733): 436-442.

Ji, G., J. Guan, et al. (2014). "Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes." Briefings in bioinformatics.

Johnson, W. E., C. Li, et al. (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research **28**(1): 27-30.

Kierszenbaum, F. (1999). "Chagas' disease and the autoimmunity hypothesis." Clinical microbiology reviews **12**(2): 210-223.

Kolev, N. G., J. B. Franklin, et al. (2010). "The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution." PLoS pathogens **6**(9): e1001090.

Kolev, N. G., J. B. Franklin, et al. (2010). "The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution." PLoS Pathog **6**(9).

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

Law, C. W., Y. Chen, et al. (2014). "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts." Genome biology **15**(2): R29.

LeBowitz, J. H., H. Q. Smith, et al. (1993). "Coupling of poly(A) site selection and trans-splicing in Leishmania." Genes Dev **7**(6): 996-1007.

Leek, J. T., R. B. Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." Nature reviews. Genetics **11**(10): 733-739.

Leek, J. T., R. B. Scharpf, et al. (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." Nat Rev Genet **11**(10): 733-739.

Leek, J. T. and J. D. Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." PLoS genetics **3**(9): 1724-1735.

Lescure, F. X., G. Le Loup, et al. (2010). "Chagas disease: changes in knowledge and management." Lancet Infect Dis **10**(8): 556-570.

Liang, X. H., A. Haritan, et al. (2003). "trans and cis splicing in trypanosomatids: mechanism, factors, and regulation." Eukaryot Cell **2**(5): 830-840.

Lima, M. T., H. L. Lenzi, et al. (1995). "Negative tissue parasitism in mice injected with a noninfective clone of Trypanosoma cruzi." Parasitology research **81**(1): 6-12.

Lorena, V. M., I. M. Lorena, et al. (2010). "Cytokine levels in serious cardiopathy of Chagas disease after in vitro stimulation with recombinant antigens from Trypanosoma cruzi." Scandinavian journal of immunology **72**(6): 529-539.

Luehr, S., H. Hartmann, et al. (2012). "The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences." Nucleic acids research **40**(Web Server issue): W104-109.

Macedo, V. (1999). "Indeterminate form of Chagas disease." Memorias do Instituto Oswaldo Cruz **94 Suppl 1**: 311-316.

Machado, C. A. and F. J. Ayala (2001). "Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi." Proceedings of the National Academy of Sciences of the United States of America **98**(13): 7396-7401.

Mahmood, R., B. Mittra, et al. (2001). "Characterization of the Crithidia fasciculata mRNA cycling sequence binding proteins." Molecular and cellular biology **21**(14): 4453-4459.

Manning-Cela, R., A. Gonzalez, et al. (2002). "Alternative splicing of LYT1 transcripts in Trypanosoma cruzi." Infect Immun **70**(8): 4726-4728.

Manque, P. A., C. M. Probst, et al. (2011). "Trypanosoma cruzi infection induces a global host cell response in cardiomyocytes." Infection and immunity **79**(5): 1855-1862.

Marchler-Bauer, A., C. Zheng, et al. (2013). "CDD: conserved domains and protein three-dimensional structure." Nucleic acids research **41**(Database issue): D348-352.

Marin-Neto, J. A., E. Cunha-Neto, et al. (2007). "Pathogenesis of chronic Chagas heart disease." Circulation **115**(9): 1109-1123.

Marin Neto, J. A., M. V. Simoes, et al. (1999). "Chagas' heart disease." Arquivos brasileiros de cardiologia **72**(3): 247-280.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-1517.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome research **18**(9): 1509-1517.

Martinez-Calvillo, S., J. C. Vizuet-de-Rueda, et al. (2010). "Gene expression in trypanosomatid parasites." J Biomed Biotechnol **2010**: 525241.

Martinez-Perez, A., F. F. Norman, et al. (2014). "An approach to the management of Trypanosoma cruzi infection (Chagas' disease) in immunocompromised patients." Expert review of anti-infective therapy **12**(3): 357-373.

Martins, G. A., L. Q. Vieira, et al. (1999). "Gamma interferon modulates CD95 (Fas) and CD95 ligand (Fas-L) expression and nitric oxide-induced apoptosis during the acute phase of Trypanosoma cruzi infection: a possible role in immune response control." Infection and immunity **67**(8): 3864-3871.

Martins Vde, P., M. Galizzi, et al. (2010). "Developmental expression of a Trypanosoma cruzi phosphoinositide-specific phospholipase C in amastigotes and stimulation of host phosphoinositide hydrolysis." Infection and immunity **78**(10): 4206-4212.

Masana, M., E. G. de Toranzo, et al. (1983). "Reductive metabolism and activation of benznidazole, a drug against Chagas' disease." Developments in toxicology and environmental science **11**: 383-386.

Matthews, K. R., C. Tschudi, et al. (1994). "A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes." Genes & development **8**(4): 491-501.

Matthews, K. R., C. Tschudi, et al. (1994). "A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes." Genes Dev **8**(4): 491-501.

McCabe, R. E. and B. T. Mullins (1990). "Failure of Trypanosoma cruzi to trigger the respiratory burst of activated macrophages. Mechanism for immune evasion and importance of oxygen-independent killing." Journal of immunology **144**(6): 2384-2388.

Michaeli, S. (2011). "Trans-splicing in trypanosomes: machinery and its impact on the parasite transcriptome." Future microbiology **6**(4): 459-474.

Miles, M. A., M. D. Feliciangeli, et al. (2003). "American trypanosomiasis (Chagas' disease) and the role of molecular epidemiology in guiding control strategies." BMJ **326**(7404): 1444-1448.

Minning, T. A., D. B. Weatherly, et al. (2009). "The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi." BMC Genomics **10**: 370.

Moller, S., M. D. Croning, et al. (2001). "Evaluation of methods for the prediction of membrane spanning regions." Bioinformatics **17**(7): 646-653.

Moncayo, A. (2003). "Chagas disease: current epidemiological trends after the interruption of vectorial and transfusional transmission in the Southern Cone countries." Memorias do Instituto Oswaldo Cruz **98**(5): 577-591.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Mount, S. M. (1983). "RNA processing. Sequences that signal where to splice." Nature **304**(5924): 309-310.

Muhich, M. L. and J. C. Boothroyd (1988). "Polycistronic transcripts in trypanosomes and their accumulation during heat shock: evidence for a precursor role in mRNA synthesis." Molecular and cellular biology **8**(9): 3837-3846.

Mulindwa, J., A. Fadda, et al. (2014). "Methods to determine the transcriptomes of trypanosomes in mixtures with mammalian cells: the effects of parasite

purification and selective cDNA amplification." PLoS neglected tropical diseases **8**(4): e2806.

Munoz-Saravia, S. G., A. Haberland, et al. (2012). "Chronic Chagas' heart disease: a disease on its way to becoming a worldwide health problem: epidemiology, etiopathology, treatment, pathogenesis and laboratory medicine." Heart failure reviews **17**(1): 45-64.

Ndao, M., T. W. Spithill, et al. (2010). "Identification of novel diagnostic serum biomarkers for Chagas' disease in asymptomatic subjects by mass spectrometric profiling." Journal of clinical microbiology **48**(4): 1139-1149.

Nilsson, D., K. Gunasekera, et al. (2010). "Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei." PLoS pathogens **6**(8): e1001037.

Oliveira Filho, A. M. (1999). "Differences of susceptibility of five triatomine species to pyrethroid insecticides - implications for Chagas disease vector control." Memorias do Instituto Oswaldo Cruz **94 Suppl 1**: 425-428.

Oshlack, A., M. D. Robinson, et al. (2010). "From RNA-seq reads to differential expression results." Genome Biol **11**(12): 220.

Oshlack, A., M. D. Robinson, et al. (2010). "From RNA-seq reads to differential expression results." Genome biology **11**(12): 220.

Ouellette, M. and B. Papadopoulou (2009). "Coordinated gene expression by post-transcriptional regulons in African trypanosomes." J Biol **8**(11): 100.

Pasion, S. G., J. C. Hines, et al. (1996). "Sequences within the 5' untranslated region regulate the levels of a kinetoplast DNA topoisomerase mRNA during the cell cycle." Molecular and cellular biology **16**(12): 6724-6735.

Peng, S. S., C. Y. Chen, et al. (1996). "Functional characterization of a non-AUUUA AU-rich element from the c-jun proto-oncogene mRNA: evidence for a novel class of AU-rich elements." Molecular and cellular biology **16**(4): 1490-1499.

Pepke, S., B. Wold, et al. (2009). "Computation for ChIP-seq and RNA-seq studies." Nat Methods **6**(11 Suppl): S22-32.

Perez-Fuentes, R., E. Torres-Rasgado, et al. (2008). "The anti-oxidant defence response in individuals with the indeterminate form of Chagas disease (American trypanosomiasis)." Annals of tropical medicine and parasitology **102**(3): 189-197.

Perez-Molina, J. A., A. Perez-Ayala, et al. (2009). "Use of benznidazole to treat chronic Chagas' disease: a systematic review with a meta-analysis." The Journal of antimicrobial chemotherapy **64**(6): 1139-1147.

Petersen, T. N., S. Brunak, et al. (2011). "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nature methods **8**(10): 785-786.

Pierleoni, A., P. L. Martelli, et al. (2008). "PredGPI: a GPI-anchor predictor." BMC Bioinformatics **9**: 392.

Pinto, A. Y., A. G. Ferreira, Jr., et al. (2009). "Urban outbreak of acute Chagas disease in Amazon region of Brazil: four-year follow-up after treatment

with benznidazole." Revista panamericana de salud publica = Pan American journal of public health **25**(1): 77-83.

Poisson, G., C. Chauve, et al. (2007). "FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring." Genomics Proteomics Bioinformatics **5**(2): 121-130.

Quijano-Hernandez, I. and E. Dumonteil (2011). "Advances and challenges towards a vaccine against Chagas disease." Human vaccines **7**(11): 1184-1191.

Ramani, A. K., J. A. Calarco, et al. (2011). "Genome-wide analysis of alternative splicing in Caenorhabditis elegans." Genome research **21**(2): 342-348.

Rassi, A., Jr., A. Rassi, et al. (2000). "Chagas' heart disease." Clinical cardiology **23**(12): 883-889.

Rassi, A., Jr., A. Rassi, et al. (2010). "Chagas disease." Lancet **375**(9723): 1388-1402.

Rassi, A., Jr., S. G. Rassi, et al. (2001). "Sudden death in Chagas' disease." Arquivos brasileiros de cardiologia **76**(1): 75-96.

Rettig, J., Y. Wang, et al. (2012). "Dual targeting of isoleucyl-tRNA synthetase in Trypanosoma brucei is mediated through alternative trans-splicing." Nucleic acids research **40**(3): 1299-1306.

Ribeiro, A. L., M. P. Nunes, et al. (2012). "Diagnosis and management of Chagas disease and cardiomyopathy." Nature reviews. Cardiology **9**(10): 576-589.

Riou, G. F. and P. Yot (1977). "Heterogeneity of the kinetoplast DNA molecules of Trypanosoma cruzi." Biochemistry **16**(11): 2390-2396.

Rojas de Arias, A., E. A. Ferro, et al. (1999). "Chagas disease vector control through different intervention modalities in endemic localities of Paraguay." Bulletin of the World Health Organization **77**(4): 331-339.

Schmunis, G. A. (1991). "Trypanosoma cruzi, the etiologic agent of Chagas' disease: status in the blood supply in endemic and nonendemic countries." Transfusion **31**(6): 547-557.

Schofield, C. J. and J. P. Dujardin (1997). "Chagas disease vector control in Central America." Parasitology today **13**(4): 141-144.

Shigihara, T., M. Hashimoto, et al. (2008). "Transcriptome profile of Trypanosoma cruzi-infected cells: simultaneous up- and down-regulation of proliferation inhibitors and promoters." Parasitology research **102**(4): 715-722.

Siegel, T. N., K. Gunasekera, et al. (2011). "Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing." Trends in parasitology **27**(10): 434-441.

Siegel, T. N., D. R. Hekstra, et al. (2009). "Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei." Genes & development **23**(9): 1063-1076.

Siegel, T. N., D. R. Hekstra, et al. (2010). "Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites." Nucleic Acids Res **38**(15): 4946-4957.

Siegel, T. N., K. S. Tan, et al. (2005). "Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei." Molecular and cellular biology **25**(21): 9586-9594.

Silva, J. S., J. C. Aliberti, et al. (1998). "The role of IL-12 in experimental Trypanosoma cruzi infection." Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica ... [et al.] **31**(1): 111-115.

Silva, L. H. P. N., V (1953). "Sôbre uma cepa de Trypanosomacruzi altamente virulenta para o camundongo branco." Folia Clin & Biol (S. Paulo) **20**: 191-208.

Simpson, L. (1973). "Structure and function of kinetoplast DNA." The Journal of protozoology **20**(1): 2-8.

Simpson, L., J. Shaw, et al. (1988). "RNA editing--a novel RNA processing phenomenon in trypanosome mitochondria." Mem Inst Oswaldo Cruz **83 Suppl 1**: 243.

Smith, C. W., E. B. Porro, et al. (1989). "Scanning from an independently specified branch point defines the 3' splice site of mammalian introns." Nature **342**(6247): 243-247.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Statistical applications in genetics and molecular biology **3**: Article3.

Soares, M. B., L. Pontes-De-Carvalho, et al. (2001). "The pathogenesis of Chagas' disease: when autoimmune and parasite-specific immune

responses meet." Anais da Academia Brasileira de Ciencias **73**(4): 547-559.

Soneson, C. and M. Delorenzi (2013). "A comparison of methods for differential expression analysis of RNA-seq data." BMC Bioinformatics **14**: 91.

Storey, J. D. (2002). "A direct approach to false discovery rates." Journal of the Royal Statistical Society Series B-Statistical Methodology **64**: 479-498.

Stuart, K. and A. K. Panigrahi (2002). "RNA editing: complexity and complications." Molecular microbiology **45**(3): 591-596.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America **102**(43): 15545-15550.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.

Sutton, R. E. and J. C. Boothroyd (1986). "Evidence for trans splicing in trypanosomes." Cell **47**(4): 527-535.

Tanowitz, H. B., A. Mukhopadhyay, et al. (2011). "Microarray analysis of the mammalian thromboxane receptor-Trypanosoma cruzi interaction." Cell Cycle **10**(7): 1132-1143.

Tardieux, I., P. Webster, et al. (1992). "Lysosome recruitment and fusion are early events required for trypanosome invasion of mammalian cells." Cell **71**(7): 1117-1130.

Tarleton, R. L. (2003). Trypanosoma cruzi and Chagas disease: cause and effect. Boston, MA, Kluwer Academic Publishers.

Tarleton, R. L. and L. Zhang (1999). "Chagas disease etiology: autoimmunity or parasite persistence?" Parasitology today **15**(3): 94-99.

Team, R. D. C. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Teixeira, S. M. and W. D. daRocha (2003). "Control of gene expression and genetic manipulation in the Trypanosomatidae." Genet Mol Res **2**(1): 148-158.

Teixeira, S. M., R. M. de Paiva, et al. (2012). "Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases." Genet Mol Biol **35**(1): 1-17.

Teixeira, S. M., L. V. Kirchhoff, et al. (1995). "Post-transcriptional elements regulating expression of mRNAs from the amastin/tuzin gene cluster of Trypanosoma cruzi." J Biol Chem **270**(38): 22586-22594.

Thomas, S., L. L. Martinez, et al. (2007). "A population study of the minicircles in Trypanosoma cruzi: predicting guide RNAs in the absence of empirical RNA editing." BMC Genomics **8**: 133.

Tostes, S., Jr., D. Bertulucci Rocha-Rodrigues, et al. (2005). "Myocardiocyte apoptosis in heart failure in chronic Chagas' disease." International journal of cardiology **99**(2): 233-237.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Vaena de Avalos, S., I. J. Blader, et al. (2002). "Immediate/early response to Trypanosoma cruzi infection involves minimal modulation of host cell transcription." The Journal of biological chemistry **277**(1): 639-644.

Vasquez, J. J., C. C. Hon, et al. (2014). "Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages." Nucleic acids research **42**(6): 3623-3637.

Vazquez-Prokopec, G. M., C. Spillmann, et al. (2009). "Cost-effectiveness of chagas disease vector control strategies in Northwestern Argentina." PLoS neglected tropical diseases **3**(1): e363.

Voigt, W. H., M. Bock, et al. (1972). "Ultrastructural observations on the activity of nifurtimox on the causative organism of Chagas' disease. I. Trypanosoma cruzi in tissue cultures." Arzneimittel-Forschung **22**(9): 1586-1589.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Weatherly, D. B., C. Boehlke, et al. (2009). "Chromosome level assembly of the hybrid Trypanosoma cruzi genome." BMC Genomics **10**: 255.

Wegner, D. H. and R. W. Rohwedder (1972). "The effect of nifurtimox in acute Chagas' infection." Arzneimittel-Forschung **22**(9): 1624-1635.

Wernicke, S. and F. Rasche (2006). "FANMOD: a tool for fast network motif detection." Bioinformatics **22**(9): 1152-1153.

Westenberger, S. J., C. Barnabe, et al. (2005). "Two hybridization events define the population structure of Trypanosoma cruzi." Genetics **171**(2): 527-543.

Westenberger, S. J., G. C. Cerqueira, et al. (2006). "Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region." BMC Genomics **7**: 60.

WHO (2002). Control of Chagas disease. Second report of the WHO Expert Committee. Technical report series no 905., Geneva: World Health Organization,.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis, Springer New York.

Wilhelm, B. T., S. Marguerat, et al. (2010). "Defining transcribed regions using RNA-seq." Nat Protoc **5**(2): 255-266.

Young, M. D., M. J. Wakefield, et al. (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol **11**(2): R14.

Zapata-Estrella, H., C. Hummel-Newell, et al. (2006). "Control of Trypanosoma cruzi infection and changes in T-cell populations induced by a therapeutic DNA vaccine in mice." Immunology letters **103**(2): 186-191.

Zhang, B. and S. Horvath (2005). "A general framework for weighted gene co-expression network analysis." Statistical applications in genetics and molecular biology **4**: Article17.

Zhang, L., S. Kasif, et al. (2004). "GC/AT-content spikes as genomic punctuation marks." Proceedings of the National Academy of Sciences of the United States of America **101**(48): 16855-16860.

Zhang, S., C. C. Kim, et al. (2010). "Delineation of diverse macrophage activation programs in response to intracellular parasites and cytokines." PLoS neglected tropical diseases **4**(3): e648.

Zingales, B., M. E. Pereira, et al. (1997). "Biological parameters and molecular markers of clone CL Brener--the reference organism of the Trypanosoma cruzi genome project." Memorias do Instituto Oswaldo Cruz **92**(6): 811-814.