ABSTRACT

Title of Dissertation: COMPARATIVE TRANSCRIPTOME PROFILING OF HUMAN FORESKIN FIBROBLASTS INFECTED WITH THE SYLVIO AND Y STRAINS OF *TRYPANOSOMA CRUZI*

Genevieve Houston-Ludlam, Doctor of Philosophy, 2016

Dissertation Directed by: Professor Najib M. El-Sayed

Department of Cell Biology and Molecular Genetics

Trypanosoma cruzi, the causative agent of Chagas Disease, is phylogenetically distributed into nearly identical genetic strains which show divergent clinical presentations including differences in rates of cardiomyopathy in humans, different vector species and transmission cycles, and differential congenital transmission in a mouse model. The population structure of these strains divides into two groups, which are geographically and clinically distinct. The aim of this study was to compare the transcriptome of two strains of *T. cruzi*, Sylvio vs. Y to identify differences in expression that could account for clinical and biochemical differences. We collected and sequenced RNA from *T. cruzi*-infected and control Human Foreskin Fibroblasts at three timepoints. Differential expression analysis identified gene expression profiles at different timepoints in Sylvio infections, and between Sylvio and Y infections in both parasite and host. The Sylvio strain parasite and the host response to Sylvio infection largely mirrored the host-pathogen interaction seen in our previous Y strain work. IL-8 was more highly expressed in Sylvio-infected HFFs than in Y-infected HFFs.

Comparative transcriptome profiling of human foreskin fibroblasts infected with the Sylvio and Y strains of *Trypanosoma cruzi*

By

Genevieve Houston-Ludlam

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
2016

Advisory Committee:
Professor Najib M. El-Sayed, Chair
Assistant Professor Hector Corrada Bravo
Professor David M. Mosser
Brian Oliver, Ph.D., National Institutes of Health
Associate Professor Michael P. Cummings, Dean's Representative

Dedication

I dedicate my degree to my husband, Mark, and our three amazing children, Alexandra, Douglas and Carmen, who bring me immeasurable joy and continually remind me to focus on what is most important in life.

Acknowledgements

I would like to thank my mentor, Dr. Najib El-Sayed, for guiding, challenging, and supporting me in my transition from student to scientist. I have learned a great deal from working with you, and I thank you for the time and effort you have invested in me on this journey.

I would like to thank my committee members – Dr. Dave Mosser, Dr. Hector Corrada Bravo, Dr. Brian Oliver, and Dr. Michael Cummings, for your time, insights, and encouragement. You have sharpened my thinking with your questions, taught me to critically evaluate things with your constructive criticisms, and improved this work with your suggestions. Thank you for investing your time an effort to work with me.

I would like to thank Trey Belew from the El-Sayed lab for his friendship and support during my time in the lab. You were always willing to listen, guide and encourage me when I was frustrated, and to answer the questions I asked, and then answer the questions I should have asked. And for making everybody in the lab laugh with your wacky sense of humor!

I would also like to thank the other members of the El-Sayed lab, past and present, for being sounding boards and friends – April Hussey, Laura Dillon, Cecilia Fernandes Dupecher, Keith Hughitt, Yuan Li, Saloe Bispo, Jungmin Choi, Ramzi Temanni, MariaSanta Mangione, Eva Morgun, Rachel Evans, Rondon Neto, Santuza Teixeira and Wanderson Da Rocha. I especially thank Jungmin for teaching me the bench skills I needed to do these experiments, Cecilia for sharing her immense knowledge of all

things *T. cruzi*, and Keith and Trey for providing, and often tweaking for my benefit, software tools to help analyze my data.

I would like to thank my father Dr. William Ludlam, may he rest in peace, for infecting me with a love of science. The discussions at the dinner table, trips to the museum and the planetarium, and the occasional dissection on the kitchen counter till mom caught us (hey, we used a box lined with Saran Wrap, right?) have helped make me the person and the mom I am today. I hope you are pleased that I have come back to Biology, the "family business". I wish you could be here to celebrate with me.

I would like to thank my mother, Dr. Natalie Paine, for her love and encouragement over the years. You served as an example of coming back to school for a doctorate later in life. I am honored to follow in your footsteps.

And last, but certainly not least, I would like to thank my husband and children, who have frequently had to put up with zapping leftovers instead of having a nice dinner, and all manner of other undone things, because I was at the lab feeding my bugs or taking a timepoint, or hunched over my computer cussing at my data. You have all been so supportive, and I could not have done this without you. And for some reason, Ali, you now want to go and do this too? Hang in there- I'm really proud of you, soon-to-be Dr. Dr. Ali.

This work was supported by National Institute of Health grant AI094773 to Dr. Najib M. El-Sayed.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ATCC	American Type Culture Collection
bp	base pair
DMEM	Dubelcco's Minimal Eagle Medium
DTU	Discrete typing unit
FBS	Fetal bovine serum
FDR	False discovery rate
GO	Gene ontology
HFF	Human foreskin fibroblast
HGNC	Hugo Gene Nomenclature Committee
hpi	hours post-infection
IQR	Inter-quartile range
KEGG	Kyoto encyclopedia of genes and genomes
LIT	Liver infusion tryptose
MOI	Multiplicity of infection
MPC	Median pairwise correlation
PBS	Phosphate buffered saline
PCA	Principal component analysis
TCT	Tissue culture trypomastigote
WHO	World health organization

Chapter 1. Introduction

1.1 Chagas disease

Trypanosoma cruzi, a protozoan parasite, is the causative agent of Chagas Disease, a disease of poverty which affects 6-7 million people primarily in Latin America. (1) It is an ancient disease, having been discovered in human mummies dated over 9000 years old. (2) It is considered a neglected tropical disease, however it is now starting to become an issue in non-endemic countries such as the US due to migration of people from endemic areas of Latin America. (3)

Chagas disease is endemic to 21 countries in Latin America. It is estimated that 6-7 million people are currently infected worldwide, including 300,000 people in the United States. (2) 70 million people are at risk for infection throughout Latin America. Every year, about 30,000 new infections occur and 10,000 people die from Chagas disease mainly due to heart complications during the late stage of the disease. (1) The economic burden of Chagas disease is over $7 billion annually. (4)

Chagas Disease was first identified by Dr. Carlos Chagas, a Brazilian physician, in 1909. (5) He named the causative agent, *Trypanosoma cruzi*, after his mentor Dr. Oswaldo Cruz. Dr. Chagas characterized the transmission mode and the entire life cycle in vectors and other hosts, a hallmark in the history of parasitology since he determined almost every aspect of a new neglected tropical disease on his own. (2)

The transmission and notable symptoms of Chagas disease are highlighted in Figure 1-1.

Chagas Disease begins when a person is bitten by an infected triatomine bug (Figure 1-1A). The insect vector, members of the Reduviidae family Triatominae sub-family, are popularly referred to as kissing bugs because they often bite the soft facial skin. The three main domestic vectors are *Rhodnius prolixus, Triatoma dimidiata*, and *Triatoma infestans.* (2)

While the insect takes a bloodmeal, it simultaneously defecates. The insect feces contain an infective form of *T. cruzi* (Figure 1-1B), which can reach underlying cells by entering the broken skin when the person rubs or scratches near the bite site. The parasite is able to actively invade nearly any type of nucleated cell in the host. (2)

Once the infection is initiated in the human host, the disease enters the acute phase. This normally lasts 4-6 weeks, and is characterized by flu-like symptoms (fever, swelling of lymph nodes and tissues, and general malaise), and high blood parasitemia. There is often marked inflammation of the area immediately surrounding the bite site, referred to as Romaña's sign (Figure 1-1C). (2)

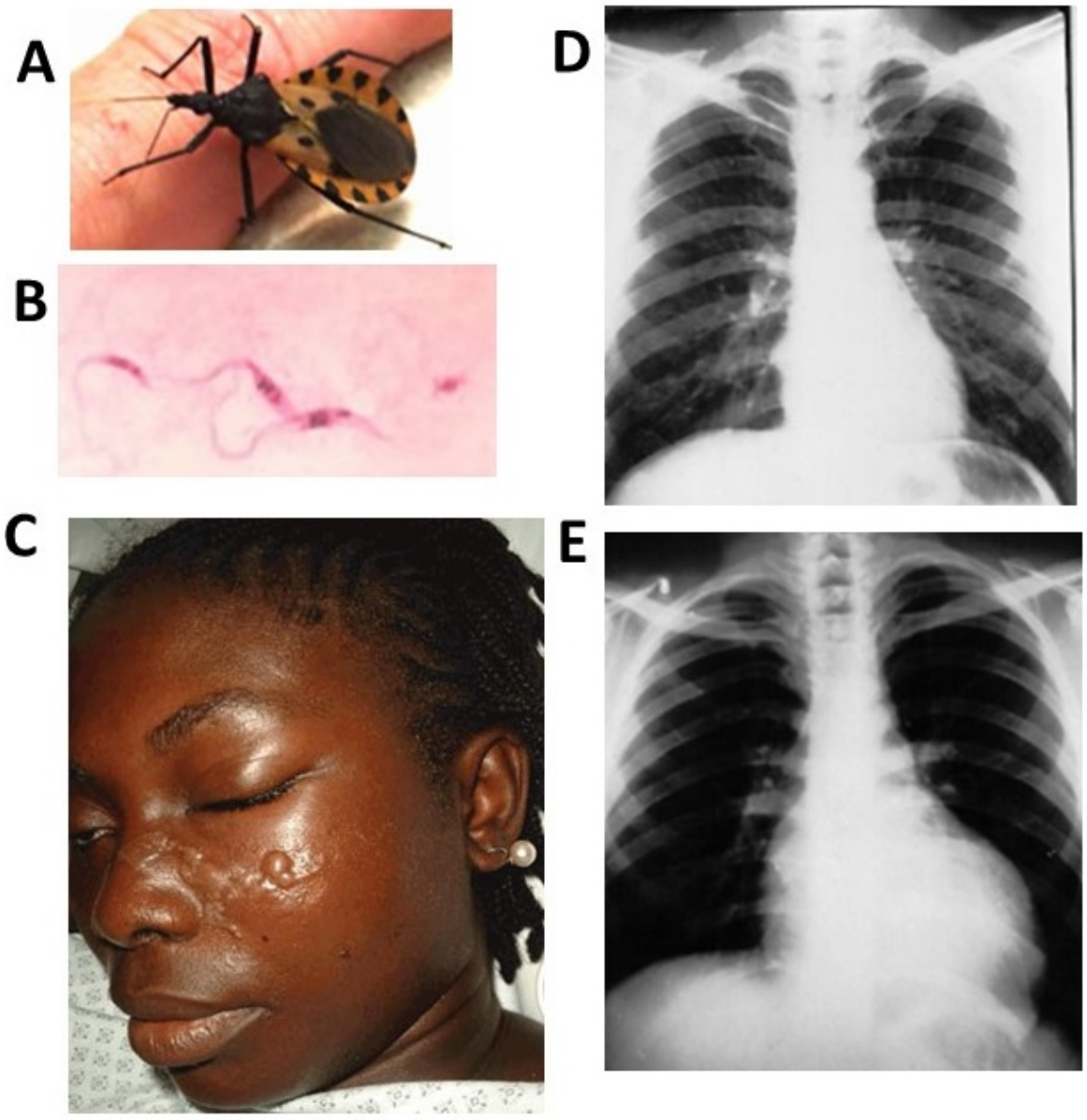

Figure 1-1. Overview of Chagas Disease. A. Insect vector for Chagas Disease, T. infestans. (6) B. The causative agent of Chagas Disease, T. cruzi (6) C. Early clinical signs of acute Chagas disease, including Romaña's sign. (6) D. Chest x-ray of normal heart. (7) E. Chest x-ray of someone with cardiac Chagas Disease, showing enlarged heart. (7)

For various reasons, including lack of access to affordable healthcare in the countries where Chagas disease is endemic, most cases are not specifically diagnosed, so most infected patients are not even aware of their infection. Unfortunately, this early acute phase is the only stage where effective treatments are available using the trypanocidal drugs Benznidazole and Nifurtimox (8, 9). Most patients resolve the acute infection spontaneously even without the use of these trypanocidal drugs, and the number of parasites in the blood and tissues significantly declines. Complications during the acute phase are usually seen in children where 2-8% of symptomatic child patients may die due to cardiac failure (10).

For most people at this point, Chagas Disease enters the indeterminate phase. Only 30% then go on to the chronic phase years (or even decades) later. Often this is the first time the patient is diagnosed with Chagas disease even though *T. cruzi* parasites may have persisted for decades (11). Host genetic susceptibility and parasite strain, amongst many other possible factors, may play a role in why such a small percentage of individuals in the indeterminate phase go on to develop chronic Chagas disease (12).

Chronic Chagas disease usually manifests as an inflammatory cardiomyopathy, which over years leads to enlargement (hypertrophy) of the heart. Figure 1-1D shows an x-ray of a normal sized heart, compared to the enlarged Chagasic heart shown in Figure 1-1E. Chagasic cardiomyopathy is the most common cause of death by heart failure in South America. Clinical symptoms include abnormal heart rhythm, heart failure, and sudden cardiac arrest (2). Both fibrosis and cell apoptosis as well as amastigotic cyst formation lead to disruption of heart electronic conduction and eventually cardiomyocyte loss (13).

Another possible but less common manifestation of chronic Chagas Disease is gastrointestinal megasyndromes, including megacolon, megaesophagus or both, leading to digestive malfunction. Amastigotic cysts in the colon or esophagus damage nerves thereby decreasing the ability of smooth muscle to dilate and contract. This loss of peristalsis leads to the inability to pass stool, leading to pain and malnutrition. A small number of people have cardiodigestive Chagas disease, which includes both cardiac as well as gastrointestinal symptoms.

Current prevention efforts have focused on vector control and public awareness campaigns (14) The Southern Cone Initiative, whose goal was to eradicate the insect vectors in many of the South American countries and thereby reduce the incidence of Chagas Disease (15) have dramatically reduced the incidence of Chagas disease in that region. There have also been extensive efforts at vaccine development, however *T. cruzi* has been refractory to vaccine treatment to date (7). Treatment of Chagas Disease is primarily accomplished with the drug Benznidazole, which is mostly effective in the acute phase of the disease (16). There are current drug trials of extended treatment with Benznidazole in chronic Chagas, and results to date have been somewhat promising (17). However, Benznidazole has very severe and unpleasant side effects, and administration is very complicated, so development of alternative treatments is very important. (7, 18)

Epidemiological data on Chagas Disease is very inconsistent. Recently, the World Health organization collected health statistics on Chagas Disease from all Latin American countries. (1) In many of these countries, health care is financially out of reach for many people, so many of the people most at risk for being infected rarely go to doctors for diagnosis and treatment, and therefore data cannot be consistently collected. The data for

these countries was estimated based on sampling, although the methods for this varied widely in quality and depth between countries.

1.2 *T. cruzi* biology/lifecycle

T. cruzi has a complex lifecycle, as do many protozoan parasites, which alternates between mammalian hosts and insect vector as shown in Figure 1-2A. There are three distinctive forms of *T. cruzi* during the lifecycle: epimastigote, trypomastigote and amastigote. The parasite takes the epimastigote form in the midgut of the insect vector where it replicates. The parasites then transit to the hindgut where they differentiate into the infective metacyclic trypomastigotes. When the insect vector takes a bloodmeal, it simultaneously defecates. The insect feces contain the metacyclic trypomastigotes which can reach underlying cells by entering the broken skin when the person rubs or scratches near the bite site. (19)

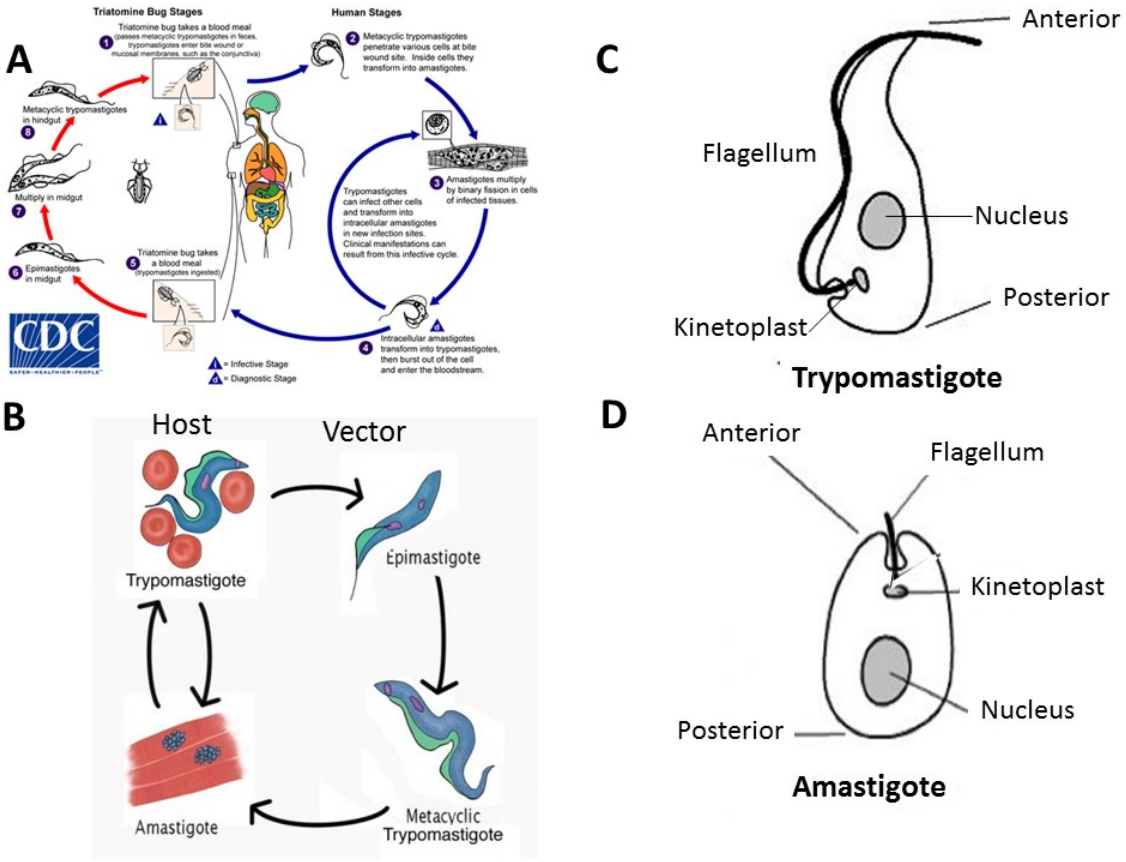

Figure 1-2. The *T. cruzi* parasite across the lifecycle. A. T. cruzi has a complex lifecycle including two distinct forms in the insect vector and two in the host. (20) B. The epimastigote is the replicative form of the parasite in the insect vector. The trypomastigote is the infective form of the parasite. The amastigote is the replicative, intracellular form of the parasite in host cells. (21) C. Trypomastigote (infective) form of the parasite. (22) D. Amastigote (host intracellular replicative) form of the parasite. (22)

Once the parasite is inside the mammalian host, the trypomastigotes can invade virtually any nucleated cell. Once inside the cell, the parasite transforms into the amastigote form within 24 hours, where it replicates for approximately four days until it transforms back into the trypomastigote form and bursts out of the cell to infect neighboring cells. If the infected host is again bitten by an insect vector, circulating trypomastigotes are ingested with the bloodmeal, which transform into epimastigotes in the midgut, completing the cycle (19).

The three forms of *T. cruzi* (epimastigote, trypomastigote and amastigote, Figure 1-2B) are most easily distinguished by the location of nucleus relative to the kinetoplast. The kinetoplast is a mitochondria-like organelle located anterior to the nucleus in the epimastigote. In the trypomastigote the nucleus is centrally positioned and kinetoplast is located towards the posterior end of the parasite. In the amastigote, the kinetoplast is located anterior to the nucleus, and the flagellum is very short. Cartoons of the two forms of *T. cruzi* used in this study are shown in Figure 1-2 C and D.

The kinetoplast, in addition to producing the energy needed for the operation of the flagellum in the same manner as a mitochondrion, contains “mini-circles” of DNA used for splicing the polycistronic transcripts produced by the parasite (23,24), and also contains guide RNAs used to edit transcripts post-transcriptionally. (25)

T. cruzi is unusual among parasites in that it can invade virtually any nucleated cell. (26) It has a wide variety of surface molecules that it uses for attachment and invasion, allowing it to be successful at invading this wide variety of cells. Figure 1-3 shows some of the molecules on the parasite and the host that are known to interact.

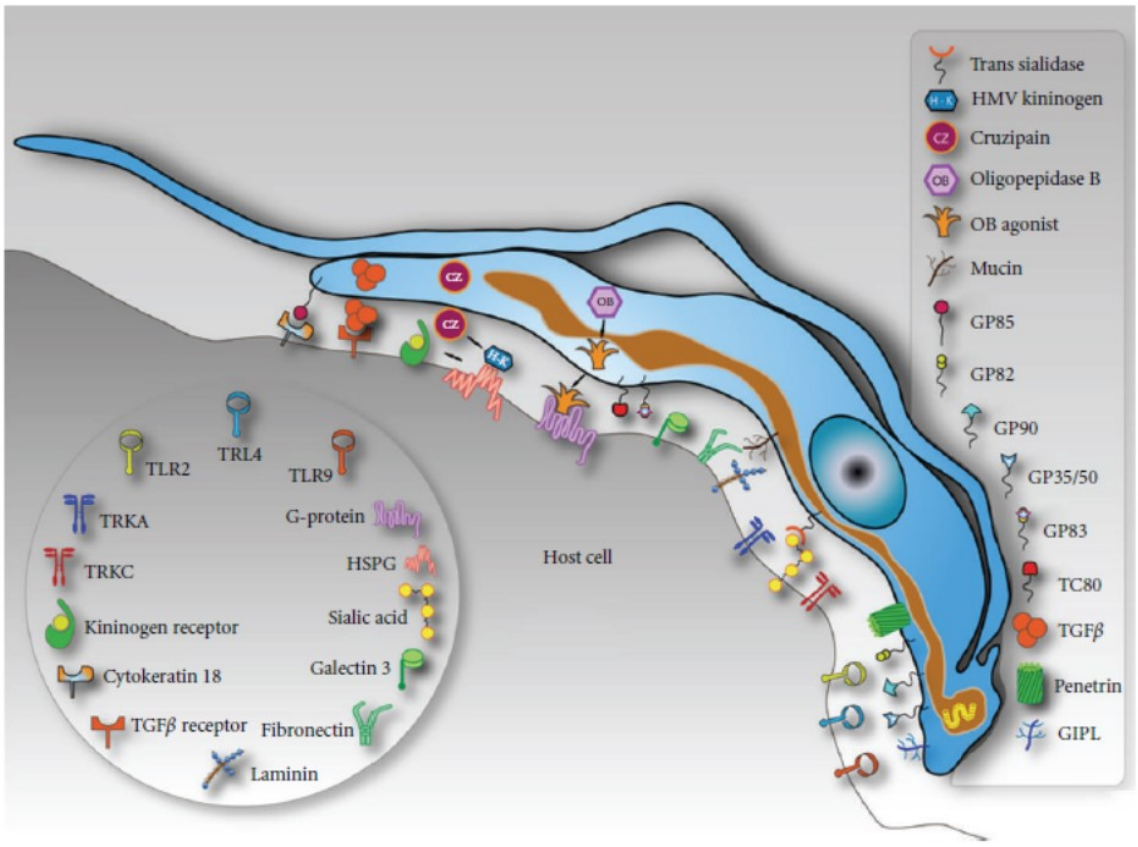

Figure 1-3. Molecules known to interact between *T. cruzi* and a host cell. (27)

Once the parasite is near a host cell, it can use one or more of these molecules, either secreting them, or “wounding” the cell membrane and bringing all of the related host cell mechanisms in to play to be exploited by the parasite (28, 19).

Figure 1-4 shows our current understanding of the infection timeline. Using one or more of the molecules identified in Figure 1-3, the parasite invades the cell within 15 minutes of attachment, and is enclosed by a portion of the host cell membrane, becoming the parasitophorous vacuole (29). This vacuole is often (but not exclusively) made out of recruited lysosomes from inside the host cell, and in fact retain the markers associated with lysosomes, such as LAMP1 (30). The lysosomes cover the parasite during its initial entry into the host, and stay in place for approximately 8 hours while the parasite begins its transformation into the intracellular, replicative amastigote form. The parasite eventually produces an acidic environment inside the parasitophorous vacuole, as well as a pore-forming protein TcTox, which helps to disrupt the membrane, and allows the parasite to escape into the cytosol. (26, 31)

The parasite completes its transformation by 24 hrs and starts dividing. By around 96 hpi, the parasite begins the transformation back to the trypomastigote form. The trypomastigotes then escape the cell and the infection process begins anew.

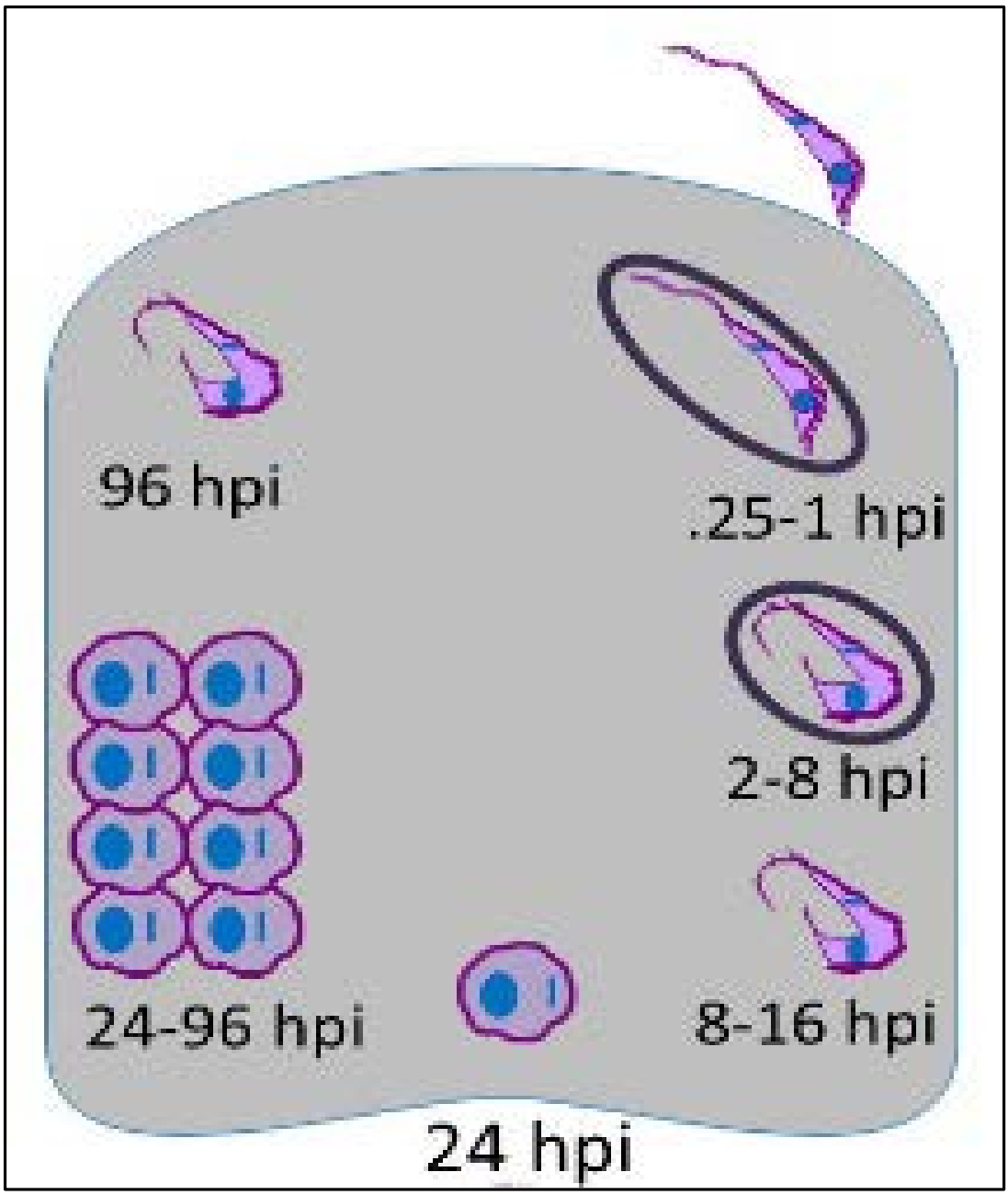

Figure 1-4. Infection dynamics of *T. cruzi*. Schematic representation of samples collected from the infection of human foreskin fibroblasts (HFF) with T. cruzi Sylvio strain. The expected progression of the T. cruzi infection is shown. Briefly, the parasite invades the cell within 15 minutes of attachment, and is enclosed by a portion of the host cell membrane (parasitophorous vacuole). By 4 hours post-infection (hpi), the parasite begins to transform from infective trypomastigote to the intracellular amastigote; it exits from the parasitophorous vacuole at around 8 hpi. The parasite completes its transformation by 24 hrs and starts dividing. By around 96 hpi, the parasite begins the transformation back to the trypomastigote form. The trypomastigotes then escape the cell and the infection process continues.

In addition to the infective trypomastigote form of the parasite, amastigotes which enter the host tissues when an infected cell bursts are also capable of invading other host cells. In some cases, the amastigotes transform back into the invasive trypomastigote form, but some strains, particularly those in DTU TcI (32, 33), have extracellular amastigotes that are capable of invading both professional and non-professional phagocytic cells by hijacking host cell actin pathways. (34, 35).

1.3 *T. cruzi* population structure

T. cruzi reproduces clonally, but there appears to be at least one historical hybridization event leading to strains with different haplotypes. The population structure is described by assigning the various strains of *T. cruzi* into six discrete typing units (DTUs): TcI to TcVI (36, 37, 38) based on molecular markers. While there is some debate as to the number of hybridization events that have occurred to produce the heterozygous strains in DTUs V and VI, there is broad agreement that DTUs I and II are pure ancient lineages. (38)

A phylogenetic analysis of 32 unlinked genetic loci in representative samples from all six DTUs have provided strong evidence for only a single hybridization event, yielding the population structure shown in Figure 1-5. (39).

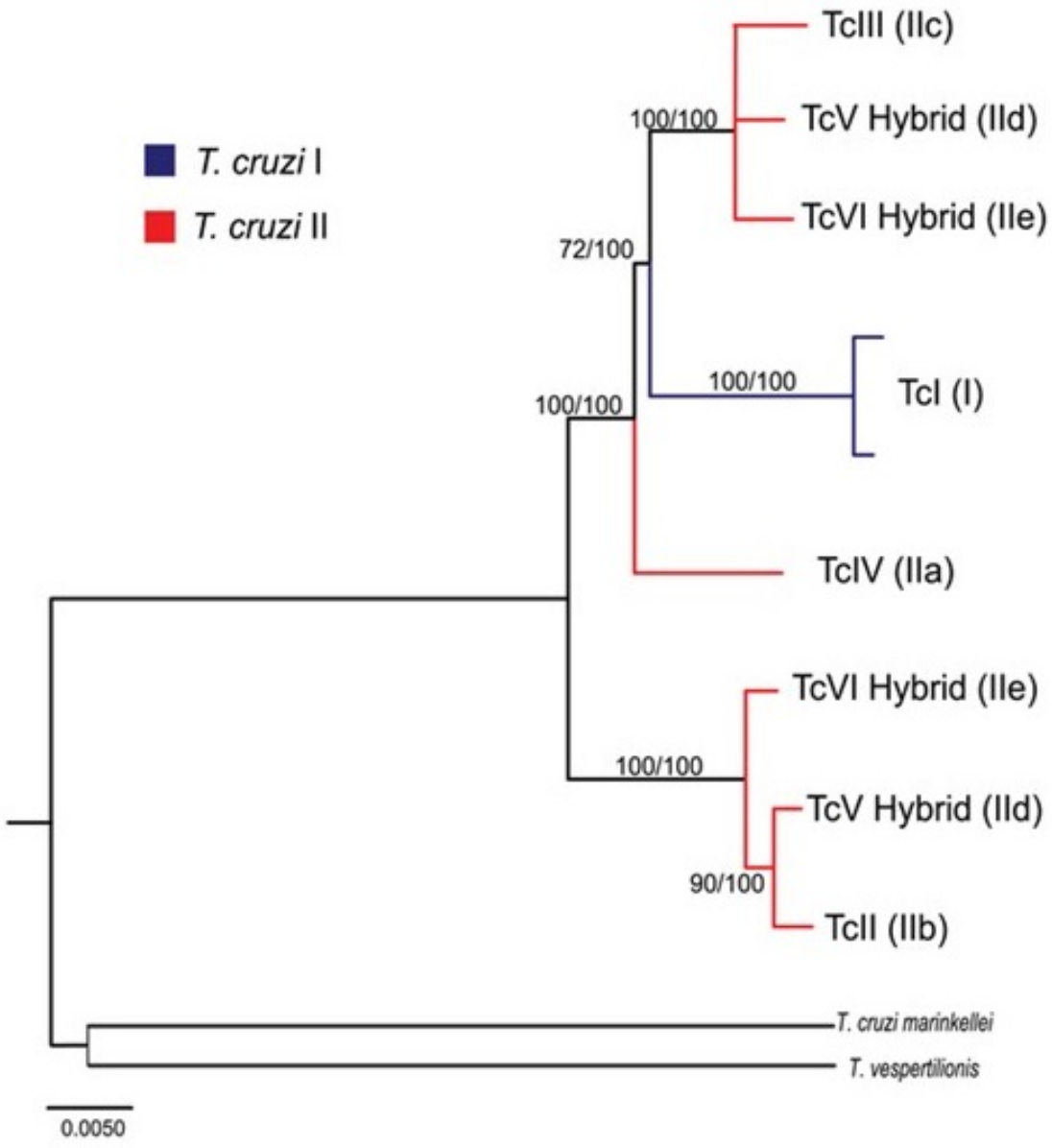

Figure 1-5. Population structure of *T. cruzi*. 32 loci in 48 strains were sequenced. Phylogenetic analysis with bootstrap support produced this tree 75% of the time, with minor changes for the rest of the trees. (39)

It appears that the historical hybridization event occurred between DTU II and DTU III, with one haplotype in each of the hybrid strains clustering with the hybrid DTUs TcV and TcVI.

There appear to be biochemical and clinical differences between the various strains of *T. cruzi* parasites. A study in rats looked at heart tissue after infection with four different strains of *T. cruzi* including a DTU TcI strain (Col1.7G2) and a DTU TcII strain (Y). The DTU TcI strain showed considerably lower cardiomyopathy, inflammation (heart and digestive organs), and cardiac denervation in these rats (40). In a study involving dogs infected with either DTU TcI strain (Col) or DTU TcII strain(Y), the TcII strain showed considerably more cardiac fibrosis than TcI (41). In a study in mice using *T. cruzi* isolates circulating in Argentina, the isolates typed as DTU TcI showed more damage to skeletal muscle compared to the DTU TcII organisms, which showed more inflammation in heart tissue. (42)

In humans, there seems to be a higher rate in gastrointestinal symptoms associated with DTU TcII parasites (43) and altered ECG (44) compared to DTU TcI.

In addition to biological differences between parasite strains, there is different geographical distribution of the strains as shown in Figure 1-6. In the Northern portion of South America, extending through Central America, strains in DTU TcI predominate. In the Southern cone area of South America, strains in DTU TcII as well as hybrid strains in DTU TcVI predominate. There is a section in the middle, particularly in Peru and the Amazon region of Brazil, where there is a mixture of strains.

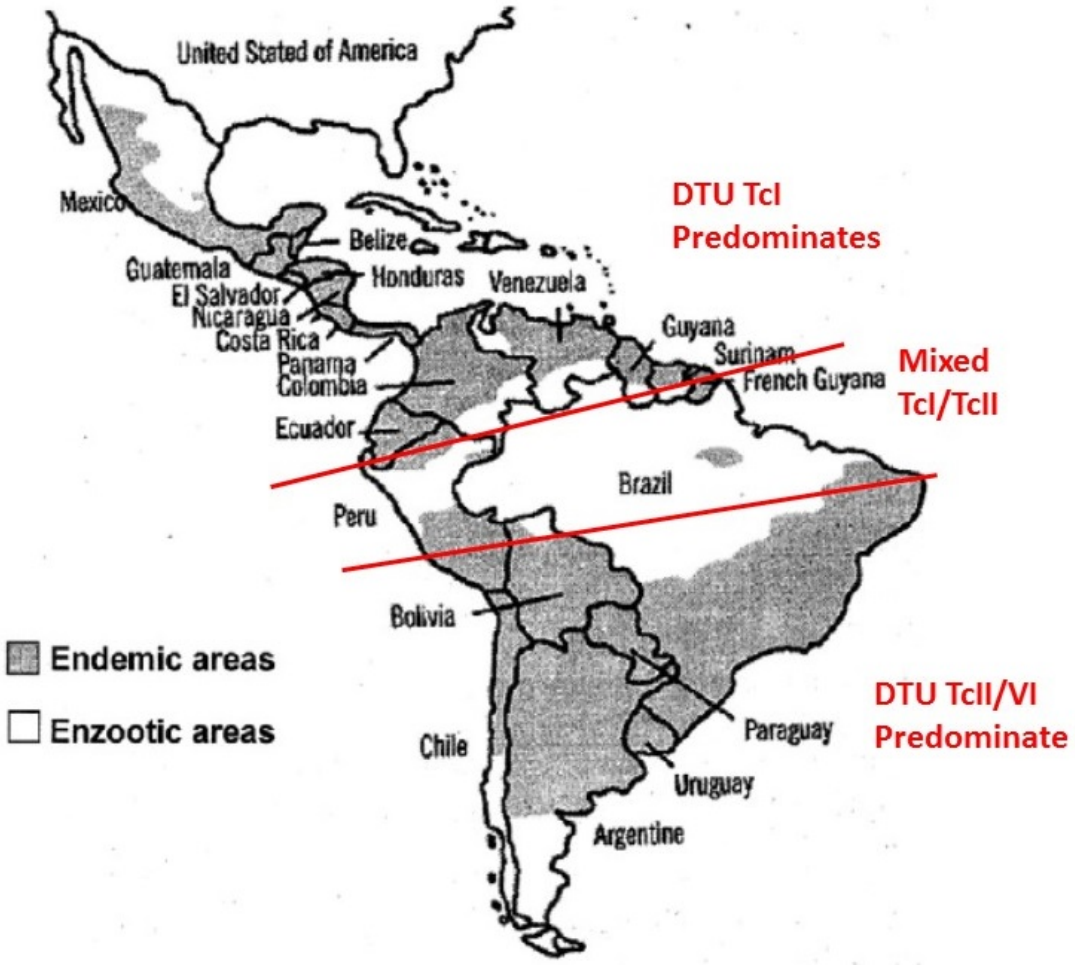

Figure 1-6. Phylogeography of *T. cruzi* strains. Original figure from (5) with annotations.

Robust epidemiological studies tying disease symptoms to parasite strain are lacking, although a recent study suggested that the rate of Chagas cardiomyopathy may be lower in countries with DTU I as the dominant strain compared to DTU II dominant countries (1). In addition, different vector species and transmission cycles have been documented for DTU I and DTU II (45).

1.4 Reference genomes

There are several reference genomes published for *T. cruzi*. The first *T. cruzi* genome was published in 2005 for the CL Brener strain (46), which was selected as the reference strain for the *T. cruzi* Genome Project (47). CL Brener, a hybrid strain assigned to DTU VI, has two heterozygous haplotypes, with approximately 50% of the genome representing repeat sequences. There are numerous multigene families, with 46 having more than 20 copies, and the largest two having over 1300 copies. These multigene families are thought to play a role in virulence (48).

The CL Brener genome was further finished and assigned to 41 pairs of chromosomes which range from approximately 78kb to 2.4Mb (49). 21133 genes were identified in the final chromosome assembly.

The second sequenced *T. cruzi* genome was for the Sylvio strain (50), the strain used in this study. The Sylvio sequencing study reported that the protein-coding regions of the two sequenced strains Sylvio strain (DTU TcI) and CL Brener (DTU TcVI) are 97-98% identical even though their geographical distribution is quite different.

Recently two additional genomes were published for the DTU TcI strain Dm28c (51) and the related bat-specific Marinkellei strain (DTU TcBat) (52). The Dm28c genome was found to be highly similar to the Sylvio genome (also a DTU TcI strain)

with a reported 98% nucleotide sequence identity and 6,094 proteins with best hits with Sylvio. The matches for CL Brener are 90% sequence identity and 5,267 protein best hits. The Marinkellei genome was compared to the Sylvio and CL Brener genomes. There was a difference in copy number in multigene families between the strains, with the Sylvio genome generally having more families than Marinkellei with the exception of the dispersed gene family. This family in Sylvio was also contracted compared to CL Brener. Nucleotide identity in orthologous gene pairs was 93% with all compared genomes (6283 orthologous genes in Sylvio, 5441 genes in CL Brener Esmeraldo-like haplotype and 5617 genes in CL Brener Non-Esmeraldo-like haplotype)

1.5 Review of current state of research and gaps

The majority of existing studies that look at infected host cells report either host or parasite individually, primarily because most studies to date have been done using microarrays.

A microarray study was done looking at the transcriptome of HeLa cells infected with the Tulahuen (TcVI) (53). This study looked at a single timepoint, 3 days post-infection. The effects of control cells were removed, and differential expression was reported for significantly up- and down-regulated genes. Their primary finding was that cell proliferation genes were being regulated to suppress host cell proliferation.

Another study looked at three different host cell types- fibroblasts, endothelial cells and smooth muscle cells, using microarrays (54). The cells were infected with Y strain and the data was collected 24 hours post-infection. The study detected a prominent type I interferon response as well as cellular repair pathways, and a reduction in mitotic cell cycle and cell division.

Rat myoblasts were interrogated using microarrays and infecting with four different strains of *T. cruzi*- Brazil (TcI), Y (TcII), CL (TcVI) and Tulahuen (TcVI) at 72 hours post-infection (55). They found that Tulahuen caused the most differential expression of these strains (17% of genes significantly DE), and Y strain the least (6% of genes). While the affected genes were not identical across strains, they found that similar pathways were altered by all four strains.

Mouse cardiomyocytes were infected with Brazil (TcI) strain and interrogated with microarrays 48 hours post-infection (56). Differentially expressed genes were primarily in the immune response, extracellular matrix and cell adhesion pathways, consistent with the kinds of pathogenesis seen in cardiac Chagas.

A study using Tulahuen infected human cardiomyocytes was performed to interrogate the early (0-120 minutes post-infection) timepoints using a microarray (57). A fibrogenic response was observed in these cells, consistent with the cardiac pathology seen in cardiac Chagas Disease.

The earliest full transcriptome using RNASeq was performed on *Trypanosoma brucei* (58). This work was used to further confirm the genome assembly, and provided evidence for bi-directional transcription from initiation sites, and also transcription from internal sites, not just Pol II initiation sites.

A microarray study of gene transcription in all four stages of the *T. cruzi* Brazil strain parasite (59) showed that transcript abundance for different genes was regulated across the parasite lifecycle. They also found that different members of multi-gene families seemed to predominate at different lifecycle stages.

Several proteomic studies of parasites have been performed, including one looking at all four stages of the parasite as the genome was published. In this study (21), they confirmed the proteins associated with many hypothetical proteins, and also proposed that the insect stages of the parasite use distinctly different energy sources, with histidine predominating in the insect forms and fatty acids used by intracellular amastigotes.

In another proteomic study, the trypomastigote versus amastigote forms of DTU TcI patient isolates were compared (60), showing that proteins associated with oxidative stress, especially tryparedoxin peroxidase, were increased, which could be associated with intracellular survival and a defense against immune response.

Finally, a meta-analysis of *T. cruzi* infection experiments was done (61) finding that there is little commonality between infection assays across the literature making it difficult to compare results between studies, and that the majority of experiments used either Tulahuen or Y strain, and very few using DTU TcI, as we do in our study.

The first RNASeq study of *T. cruzi* was our previous work with Y strain (DTU TcII) (62). This work extends this initial study by comparing the results to a DTU TcI strain, Sylvio.

1.6 Summary of dissertation work

The aim of this study was to conduct gene expression profiling of DTU TcI and DTU TcII *T. cruzi* infecting human host cells (Human foreskin fibroblasts) using RNA-Seq. We investigated the global differences at the transcription level as a proxy for translation to identify biological pathways and cellular functions being performed by the

host and parasite as the infection progressed. By following both host and parasite transcriptomes over time, we hoped to better understand the host-pathogen interaction. An improved understanding of the molecular basis for this disease will assist in developing preventive vaccines and/or targeted drugs with less toxicity. Further, the differential expression profile for the two strains of parasite will potentially give us insight into the different clinical presentation given by TcI and TcII parasite strains.

Chapter 2 provides a description of the materials and methods used to produce the data for this study.

Chapter 3 discusses the results of the infection experiments including the RNA isolation, sequencing, mapping to the genome and counting each transcript.

Chapter 4 discusses the differential expression and GOSeq analysis for the human data.

Chapter 5 discusses the differential expression and GOSeq analysis for the parasite data.

Chapter 6 contains concluding remarks and proposes future directions for this work.

This work has been submitted for publication.

Contributions to this work: Genevieve Houston-Ludlam (GHL) and Najib El-Sayed (NES) conceived the project and designed experiments; GHL performed experiments; GHL, Ashton Trey Belew (ATB), Hector Corrada Bravo (HCB) and NES analyzed data; GHL and NES wrote the paper submitted based on this work. All the sequencing was

performed at the University of Maryland Institute for Bioscience and Biotechnology Research (IBBR) sequencing core.

Chapter 2. Materials and Methods

2.1 Infection experiments

Low-passage primary human foreskin fibroblasts (HFF) BJ strain (ATCC CRL-2522) were propagated in Dulbecco's Modified Eagle Medium (DMEM, Gibco) with 10% fetal heat-inactivated bovine serum (FBS, Gemini Bioproducts), 1x MEM Non-essential Amino Acids (Cellgro) and 100 units Penicillin/100 µg Streptomycin/0.25 µg Amphotericin per mL (DMEM-10% FBS) and incubated at 37°C in 5% CO_2. Following expansion, $2.5x10^5$ cells were seeded into 6-well plates in DMEM-10% FBS for 48 hours. Identical wells were prepared for microscopy with glass microscope coverslips for parasite burden computation. Two biological replicates, each originating from separate freeze dates after expansion from the original ATCC culture, were performed.

Parasites were maintained as pure cultures of epimastigotes in LIT media supplemented with 10% HI-FBS at 27ºC. Epimastigotes were transformed to metacyclic trypomastigotes by leaving the parasites in the spent media for 14 days. The combined epimastigote/metacyclic trypomastigotes were placed on Vero Cells for 24 hours, then the media was replaced with 10% Horse Serum for 24 hours. The cells were then incubated in DMEM-2% FBS. When TCTs erupted, they were placed on a new Vero cell culture, and DMEM supplemented with 10% horse serum was used as growth media until there were no more visible epimastigotes in the flask or on the hemocytometer during cell counts. Further passages of TCTs were then grown in DMEM supplemented with 2% HI-FBS. Stabilates were made of the pure TCT cultures in 50% HI-FBS, 40% DMEM and 10% DMSO and then 1 mL vials were cryopreserved in liquid nitrogen.as the starting point for all infections.

Stabilates of *T. cruzi* strains Sylvio and Y were propagated in monolayers of Vero cells in DMEM (Cellgro) with 2% heat-inactivated FBS (Gemini Bioproducts), 1x MEM Non-Essential Amino Acids (Cellgro) and 100 units Penicillin/100 µg Streptomycin per mL (DMEM-2% FBS). Tissue culture trypomastigotes (TCTs) were harvested from the supernatant, centrifuged and resuspended in DMEM-2% FBS. Parasites (3.0×10^7 parasites per well, 100:1 infection MOI) were incubated with HFF cells for 2 hours at 37°C at 5% CO_2 to allow invasion. The remaining extracellular parasites were aspirated, each well was washed 5 times with PBS, fresh DMEM-2% FBS medium was added, and cultures were incubated for the indicated times. Timepoints were taken at 4, 12, 20, 24, 30, 48, 72 and 96 hours post-infection (hpi). Each biological replicate originated from a different stabilate of Sylvio or Y strain after expansion from the original culture.

The details of the materials used in these experiments is shown in Table 2-1.

Table 2-1. List of cells used in this study. The source for each cell line is provided

Materials	Catalog numbers	Sources
Vero Cells	CCL-81	ATCC
Human Foreskin Fibroblasts- BJ line	CRL-2522	ATCC
T. cruzi Sylvio (DTU I)	50800	ATCC
T. cruzi Y (DTU II)		Dr. Barbara Burleigh (62)

2.2 RNA sequencing

Media was aspirated from wells and 1 mL Trizol® reagent (Life Technologies) was applied to the cell culture. The Trizol® mixture was then collected and immediately placed in a -80 °C freezer until total RNA could be isolated according to the Trizol® protocol. Residual DNA was degraded using DNase and the remaining RNA was purified with a Qiagen RNeasy mini kit. RNA quantification and quality were assessed with an Agilent 2100 Bioanalyzer. PolyA+ cDNA libraries were prepared with the Illumina TruSeq Sample Preparation kit (San Diego, CA, USA) using the Illumina indexes for multiple samples per lane. Libraries were checked for quality and quantity using the Agilent Bioanalyzer 2100 and qPCR (KAPA Biosystems), then sequenced on an Illumina Hi-Seq 1500.

Wells with coverslips were fixed with 4% paraformaldehyde for ten minutes. The coverslips were washed five times with PBS, mounted onto microscope slides with ProLong Gold antifade reagent along with DAPI stain (Life Technologies) and allowed to set overnight at room temperature in the dark. Parasites were counted at 100x oil immersion on a Zeiss AxioObserver Microscope. One hundred cells on each coverslip were counted, with three coverslips per timepoint. Percent infected HFF cells and parasite burden per infected cell were computed.

2.3 RNA-Seq data processing and mapping

Paired-end reads (100 bp) were demultiplexed then trimmed for residual adapter sequences using Trimmomatic (63). Trimmed reads were checked for per-base quality

using FastQC (64) and sequences with poor quality at the ends were further trimmed with Trimmomatic. Reads were aligned with Tophat (version 2.0.10) (Kim et al., 2013) to the appropriate reference genome (hg19 for human, TriTrypDB version 4.1 for CL Brener Esmeraldo-like haplotype (proxy genome for Y-strain parasite) (46) or TriTrypDB version 6.0 for Sylvio parasite (50). Two mismatches per read were allowed and reads were allowed to map only to a single locus. The number of reads mapping to each gene feature in the corresponding annotation files was determined using HTSeq (66).

2.4 Transcriptome analysis

Data for human and parasite samples collected in this study and data from our previous work were checked for consistency between replicates using a variety of global inter-sample correlation analyses including Pearson correlation, median pairwise correlation (MPC), box plots, principal component analysis (PCA) and Euclidean distance-based hierarchical clustering. A standard method for identifying outliers was applied to the data samples (67). The median pairwise correlation (MPC) for the first quartile across samples was computed (Q1), as was the inter-quartile range (IQR). Briefly, any sample whose MPC was less than Q1-1.5(IQR) was removed as an outlier. Two samples were removed from the human data based on this criterion. No parasite samples were removed.

Weakly expressed genes, 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 (68) (here n=10 and 6 for the *T. cruzi* and human samples, respectively) were removed from subsequent analyses. Data were quantile normalized and $\log_2$ transformed (69, 70). The list of differentially expressed (DE) genes was determined using voom (71) to account for the

mean-variance relationship, then the data were fitted to a linear model, which included batch to account for experimental batch effects in the data (72), using limma (73,74). Benjamini-Hochberg multiple testing correction was applied at an FDR cutoff <0.05 to determine significance (75).

Version 8.0 genome fasta files for *T. cruzi* Sylvio strain and *T. cruzi* CL Brener strain (Esmeraldo-like haplotype) were evaluated to identify uniquely mapping orthologs. The FASTA program ggsearch36 (76), which does a global alignment, was used with default parameters (E-value <= 0.001, score > 0). Reciprocal best hits were then further refined by removing genes in the four largest multi-gene families (Mucin associated surface protein, trans-sialidase, retrotransposon hot-spot and dispersed gene family protein), thereby avoiding the problems introduced by assembly of highly similar and repetitive genes. Altogether there were 4,659 unique orthologs, of which 3,279 (70%) were annotated as hypothetical proteins. This set of genes is referred to as the "core" set of parasite genes.

The GO Term Finder (77) was applied to the human and parasite data using the server at go.princeton.edu. For the human data, Ensembl IDs were converted to HUGO Gene Nomenclature Committee Database (HGNC) IDs with the BioMart ID Conversion tool, then GO Term Finder was applied with the "GOA + HGNC Xrefs- H. sapiens (human)" annotation. The human data was also interrogated with g:Profiler (78) to simultaneously identify GO, KEGG and Reactome term enrichment. For *T. cruzi*, the GO terms were extracted from the Version 9.0 TriTrypDB text file into a standard gene association file (GAF format) and provided in the advanced input options.

Chapter 3. Infection Experiments

In this study, we investigated whether we could detect differences in the global transcriptomes of two *T. cruzi* strains Sylvio (DTU I) and Y (DTU II) and their host cell given the distinct clinical and biochemical profiles suggested by previous reports (75,36,76,50,77,37). Using RNA-Seq, we simultaneously interrogated the transcriptomes of both the host and the parasite over the course of the intracellular infection. Three independent Sylvio infections were carried out in HFF cells. We performed a single limited experiment with Y strain-infected HFF cells during replicate 1 of the Sylvio experiments to check for consistency across experimental dates and to allow us to evaluate batch effects.

Three replicate experiments were performed as shown in Table 3-1. Each biological replicate originated from a different stabilate of Sylvio or Y strain. To avoid the impact of infective amastigotes in the media (35), the number of amastigotes was counted by hemocytometer along with trypomastigotes prior to the infection, with a cutoff of 5%.

Table 3-1. Description of three replicate experiments. Dates of the respective experiments, stabilates used and percentage of amastigotes in the media used to initiate infection are reported.

Replicate	Date of Experiment	HFF Stabilate	Sylvio Stabilate	Y Stabilate	% Amas
1	6/16/2013	3/31/2103	2/26/2013	5/24/2012	3.20%
2	7/7/2013	3/15/2013	2/26/2013	NA	4.09%
3	8/11/2013	3/31/2013	2/26/2013	NA	3.03%

T. cruzi strains Sylvio and Y were propagated in monolayers of Vero cells in DMEM (Cellgro) with 2% heat-inactivated Fetal Bovine Serum (Gemini Bioproducts), 1x MEM Non-Essential Amino Acids (Cellgro) and 100 units Penicillin/100 µg Streptomycin per mL (DMEM-2% FBS). Tissue culture trypomastigotes (TCTs) were harvested from the supernatant, centrifuged and resuspended in DMEM-2% FBS. Parasites (3.0×10^7 parasites per well, 100:1 infection MOI) were incubated with HFF cells for 2 hours at 37°C at 5% CO_2 to allow invasion. The remaining extracellular parasites were aspirated, each well was washed 5 times with PBS, fresh DMEM-2% FBS medium was added, and cultures were incubated for the indicated times. For each replicate, timepoints were taken at 4, 12, 20, 24, 30, 48, 72 and 96 hours post-infection. An aliquot of the trypomastigotes used to begin the infection was also collected. RNA from each timepoint was harvested from infected cells and matched uninfected control cells using Trizol® reagent, and the parallel wells with microscope cover slips were fixed, stained with DAPI and mounted on microscope slides.

3.1 Microscopy

The stained coverslips from each timepoint were counted using a Zeiss Observer.Z1 Fluorescent Microscope. Pictures were captured using the attached Zeiss Axiocam MRm Rev. 3 camera, and analyzed and stored using AxioVision Software V4.8.2.0.

Each timepoint had three replicate coverslips for infected cells. One coverslip from control cells was also observed to ensure that there was no readily observable difference in the growth characteristics between control and infected host cells. 100 cells per coverslip were counted to compute percent infected cells as well as parasites per infected

cell. Figure 3-1 shows an example of a snapshot of cells at each timepoint at 400x magnification.

Counting was done at 1000x with oil to clearly visualize the nucleus and kinetoplast of each parasite. Figure 3-2 shows the results of the counting performed at 1000x magnification. Figure 3-2A shows an example of an intracellular amastigote compared to a host cell nucleus at 1000x with oil. Both nucleus and kinetoplast were visualized for each parasite while counting to avoid counting cell debris. This count is therefore considered a conservative estimate. Figure 3-2B shows the percentage of infected cells across replicates. This was computed as the number of cells with at least one parasite divided by the total number of cells for that timepoint. Figure 3-2C shows the number of parasites per infected cell ("parasite burden") across replicates. This was computed as the number of parasites divided by the number of infected cells for that timepoint.

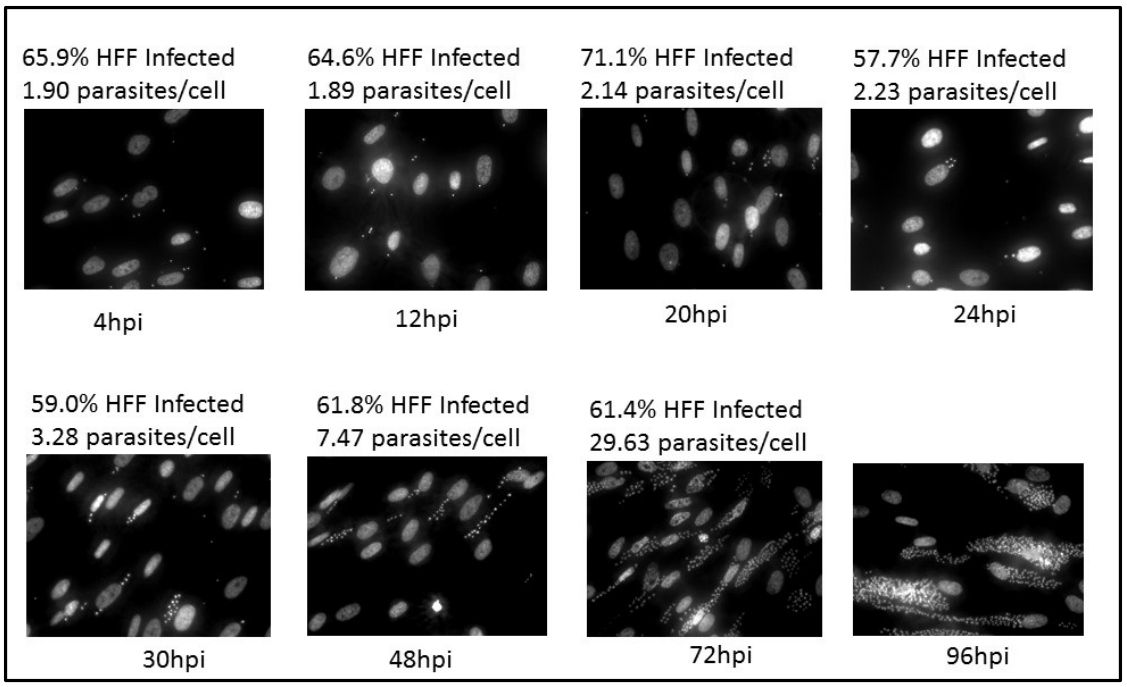

Figure 3-1. Microscope snapshots of infected HFFs across time. HFF cells infected with T. cruzi Sylvio strain were stained with DAPI and observed at 400x magnification on a Zeiss Observer.Z1 Fluorescent Microscope using UV light. The increase in the number of parasites per cell over time, especially starting at 24 hours is clearly visible.

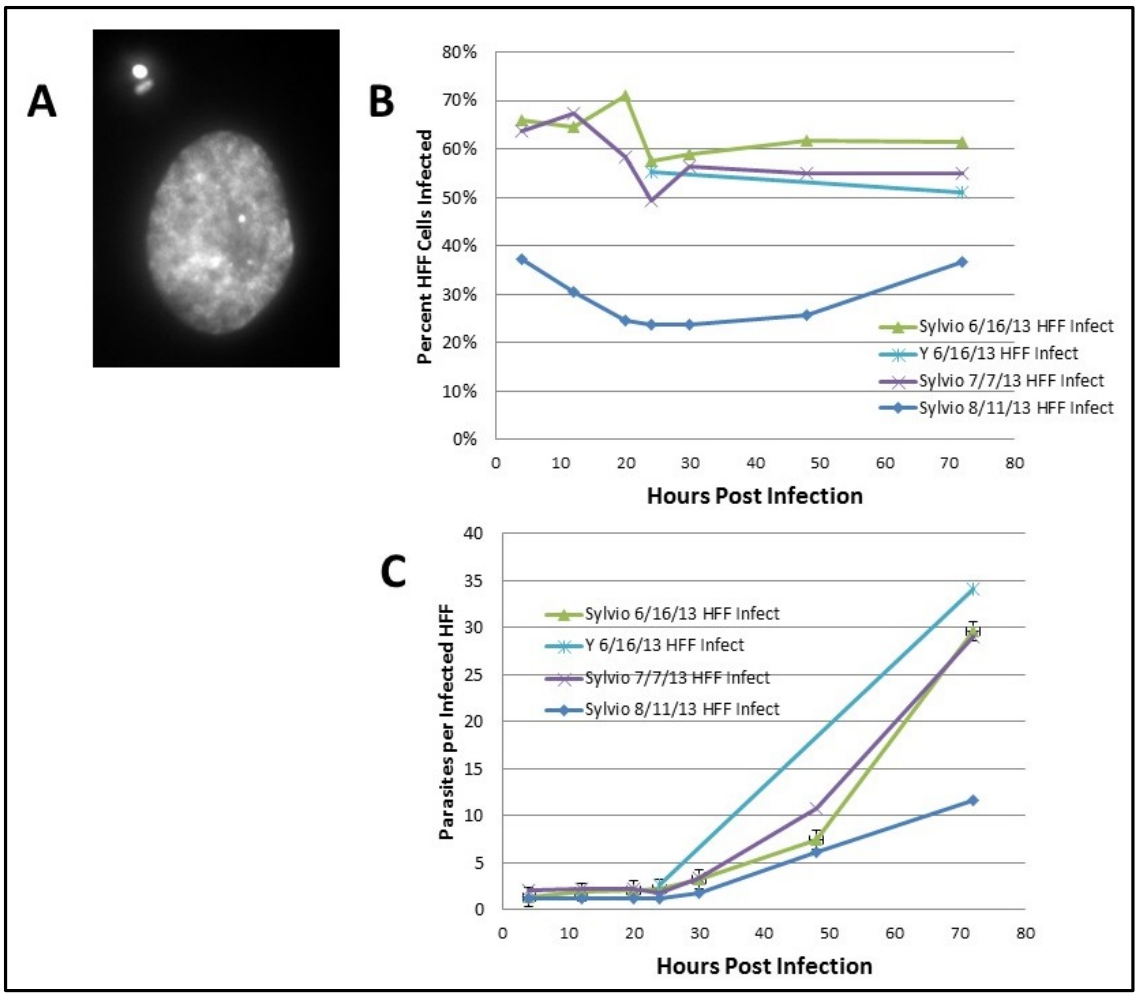

Figure 3-2. Results of microscopy for Sylvio and Y experiments. Coverslips with infected cells were counted at 1000x magnification with oil on a Zeiss Observer.Z1 Fluorescent Microscope using UV light. One hundred cells on each of three independent cover slips per experimental timepoint were counted. A. A host cell nucleus with a parasite. Both the parasite nucleus (round) as well as the kinetoplast (linear) are clearly visible. B. Percent infected cells. This was computed as the number of cells with at least one parasite divided by the total number of cells for that timepoint. C. Number of parasites per infected cell. This was computed as the number of parasites divided by the number of infected cells for that timepoint. Both the nucleus and the kinetoplast had to be seen for that parasite to be counted to ensure that cellular debris was not inadvertently counted. This count is therefore a conservative number.

As can be seen readily in Figure 3-2B, the third replicate, performed 8/11/13, had remarkably different infection dynamics compared to the other two replicates. This, combined with the smaller parasite burden as seen in the later timepoints in Figure 3-2C and the lower amount of RNA seen after RNA isolation caused us to drop this third replicate at this point and not continue with it into sequencing.

The replicate infections with *T. cruzi* Sylvio strain behaved consistently with an approximate parasite burden of two parasites per infected cell at 4 hpi and about 65% of the host cells infected. Parasite division was observed beginning at 24 hpi. Both Sylvio and Y strains displayed similar infection dynamics and behaved consistently with the existing Y strain dataset (62). Most importantly, both Sylvio and Y strains followed the expected time course parameters of an experimental *T. cruzi* infection as previously described (81).

3.2 RNA isolation and quality evaluation

Total RNA from infected and uninfected human cells and parasite TCTs was isolated for each of the biological replicates using the Trizol® protocol and purified with a Qiagen RNeasy mini kit. After isolation, the isolated total RNA was submitted for analysis by an Agilent Bioanalyzer 2100. Figure 3-3 is a sample Bioanalyzer report for total RNA from various types of samples. Figure 3-3A is parasite only, Figure 3-3B is human only (control), Figure 3-3C is an infected host sample (both signals). The spikes between 1500 and 4000 nucleotides in length represent ribosomal RNA subunits as

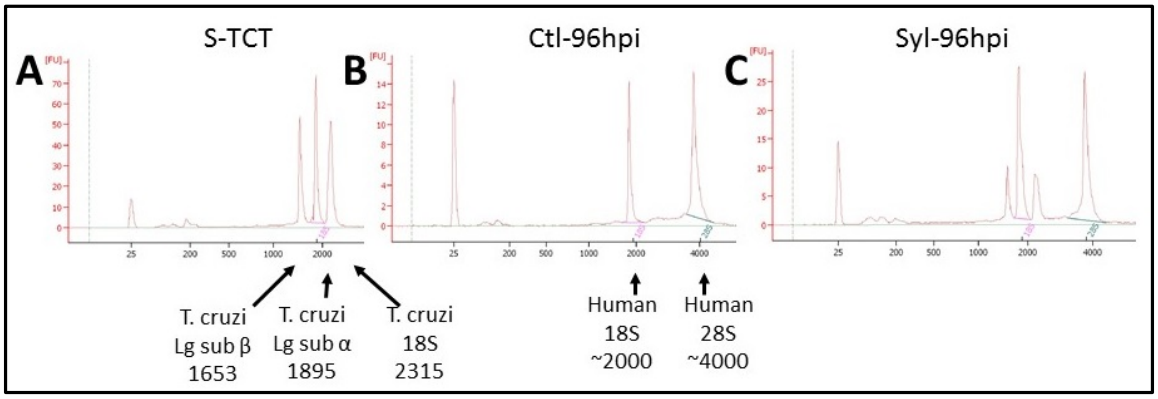

Figure 3-3. Analysis of total RNA quality and quantity. A. Example Bioanalyzer run for trypomastigote stage T. cruzi parasites. The three major rRNA peaks associated with the T. cruzi parasite are seen. B. Example Bioanalyzer run for uninfected HFF cells. The two major human rRNA peaks are seen. C. Example Bioanalyzer run for T. cruzi Sylvio-infected HFF cells. The peaks for both human and parasite are seen, with the large subunit α of the parasite combined with the 18S human peak.

labeled. All of the spikes are clearly present for both human and parasite in Figure 3-3C, the infected human cells.

3.3 Library preparation and sequencing

Poly(A)+ enriched cDNA libraries were generated using the Illumina HiSeq Library Preparation Protocol version 2. cDNA quality was assessed using the Agilent Bioanalyzer 2100. Figure 3-4 is a sample Bioanalyzer report for cDNA. The majority of cDNAs fall around 300bp, the ideal size for the Illumina sequencer.

Quantity was determined by KAPA Biosystems qPCR. The libraries were sequenced using the Illumina HiSeq 1500 platform to produce paired-end, 100bp reads. A total of ~1.1 billion reads were produced across 16 samples, with overall 94% mapping to either the host or pathogen reference genomes as shown in Table 3-3.

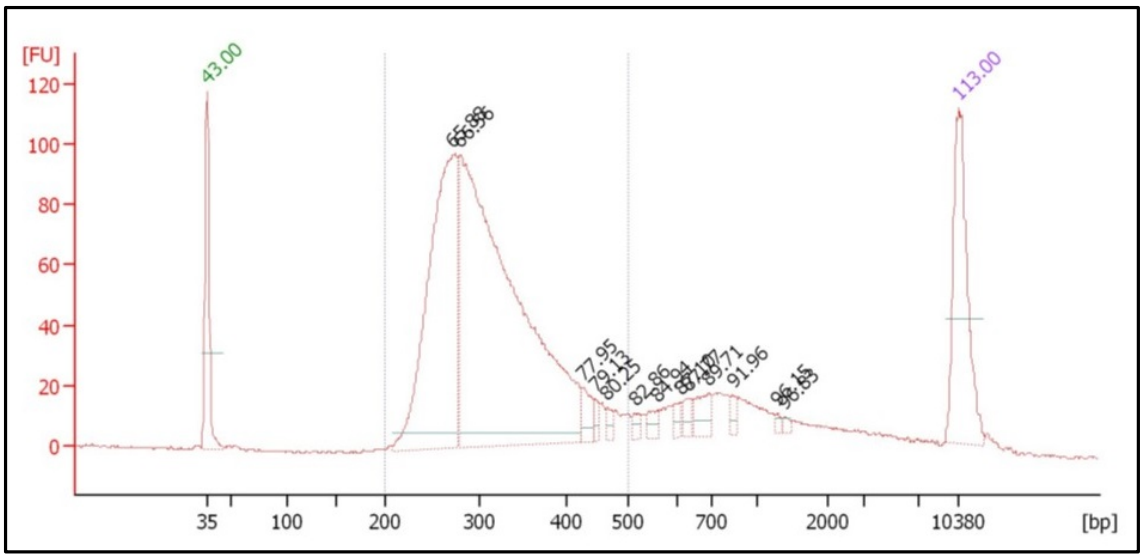

Figure 3-4. Analysis of quality of cDNA library. The majority of cDNAs fall around 300bp, the ideal size for the Illumina HiSeq 1000 sequencer.

Table 3-2. Sample information. Metadata, total sequenced reads, and mapped reads; these are defined by number and percentage aligned to human, Sylvio, and the Esmeraldo-like (Y) reference genomes for each sample**.**

Lab sample ID	Accession Number	Alias	Developmental stage	Infected	Batch	Number of reads that pass Illumina filter	Total number of reads mapped	% of total reads mapped	Reads mapped to *T. cruzi* Esmeraldo haplotype	Reads mapped to *T. cruzi* Sylvio	Reads mapped to hg19	% reads mapped to *T. cruzi* Esmeraldo	% reads mapped to *T. cruzi Sylvio*	% reads mapped to hg19
HPGL0281	SRR2177733	C4-G	Amastigote 4 hpi	N	G	79,067,424	75,287,116	95.22%			75,287,116	0.00%	0.00%	100.00%
HPGL0285	SRR2177814	S4-G		N	G	89,217,276	85,411,047	95.73%		1,518,168	83,892,879	0.00%	1.78%	98.22%
HPGL0356	SRR2180333	C4-H		Y	H	67,265,482	64,181,204	95.41%			64,181,204	0.00%	0.00%	100.00%
HPGL0360	SRR2180338	S4-H		Y	H	68,534,006	65,084,433	94.97%		1,096,879	63,987,554	0.00%	1.69%	98.31%
HPGL0282	SRR2177741	C24-G	Amastigote 24hr	N	G	74,943,922	72,117,136	96.23%			72,117,136	0.00%	0.00%	100.00%
HPGL0286	SRR2177820	S24-G		Y	G	74,121,048	71,231,671	96.10%		2,868,893	68,362,778	0.00%	4.03%	95.97%
HPGL0288	SRR2180325	Y24-G		Y	G	62,389,102	59,085,297	94.70%	2,663,709		56,421,588	4.51%	0.00%	95.49%
HPGL0357	SRR2180334	C24-H		N	H	55,024,322	52,609,972	95.61%			52,609,972	0.00%	0.00%	100.00%
HPGL0361	SRR2180340	S24-H		Y	H	59,678,804	56,335,758	94.40%		2,433,537	53,902,221	0.00%	4.32%	95.68%
HPGL0283	SRR2177785	C72-G	Amastigote 72hr	N	G	82,639,838	79,368,408	96.04%			79,368,408	0.00%	0.00%	100.00%
HPGL0287	SRR2177823	S72-G		Y	G	82,587,640	78,034,464	94.49%		25,127,565	52,906,899	0.00%	32.20%	67.80%
HPGL0289	dropped	Y72-G		Y	G	66,278,984	56,204,739	84.80%	23,908,270		32,296,469	42.54%	0.00%	57.46%
HPGL0358	SRR2180336	C72-H		N	H	59,074,892	56,452,620	95.56%			56,452,620	0.00%	0.00%	100.00%
HPGL0362	dropped	S72-H		Y	H	53,914,928	50,429,401	93.54%		21,234,180	29,195,221	0.00%	42.11%	57.89%
HPGL0284	SRR2177698	STCT-G	Trypomastigote	N	G	72,374,186	65,770,837	90.88%		65,770,837		0.00%	100.00%	0.00%
HPGL0359	SRR2177699	STCT-H		N	H	65,886,362	59,302,606	90.01%		59,302,606		0.00%	100.00%	0.00%
Total						1,112,998,216	1,046,906,709	94.06%	26,571,979	179,352,665	840,982,065	2.54%	17.13%	80.33%

After sequencing, the reads were demultiplexed into individual samples using the standard Illumina sequencing index that was incorporated into the samples during library preparation. The sequences were then trimmed for residual adapter sequences using Trimmomatic (63). Trimmed reads were checked for per-base quality using FastQC (64) and sequences with poor quality at the ends (quality score error bar dropping below 20) were further trimmed with Trimmomatic. An example per-base quality report is shown in Figure 3-5.

Reads were aligned with Tophat (version 2.0.10) (65) to the appropriate reference genome (hg19 for human, TriTrypDB version 4.1 for CL Brener Esmeraldo-like haplotype (proxy genome for Y-strain parasite) (46) or TriTrypDB version 6.0 for Sylvio parasite (50). Two mismatches per read were allowed and reads were allowed to map only to a single locus.

Because RNA-seq reads could be mapped unambiguously to the host and parasite reference genomes, we were able to track the respective fractions of RNA for each species in each sample. Our analysis revealed a proportion of 1.7% of the total mRNA population attributable to the parasite early in the infection (4 hpi). The parasite's contribution increased to 4.2% at 24 hpi, consistent with the expected doubling beginning at around that point in the infection. By 72 hpi, the parasite RNA proportion increased to 36.1%.

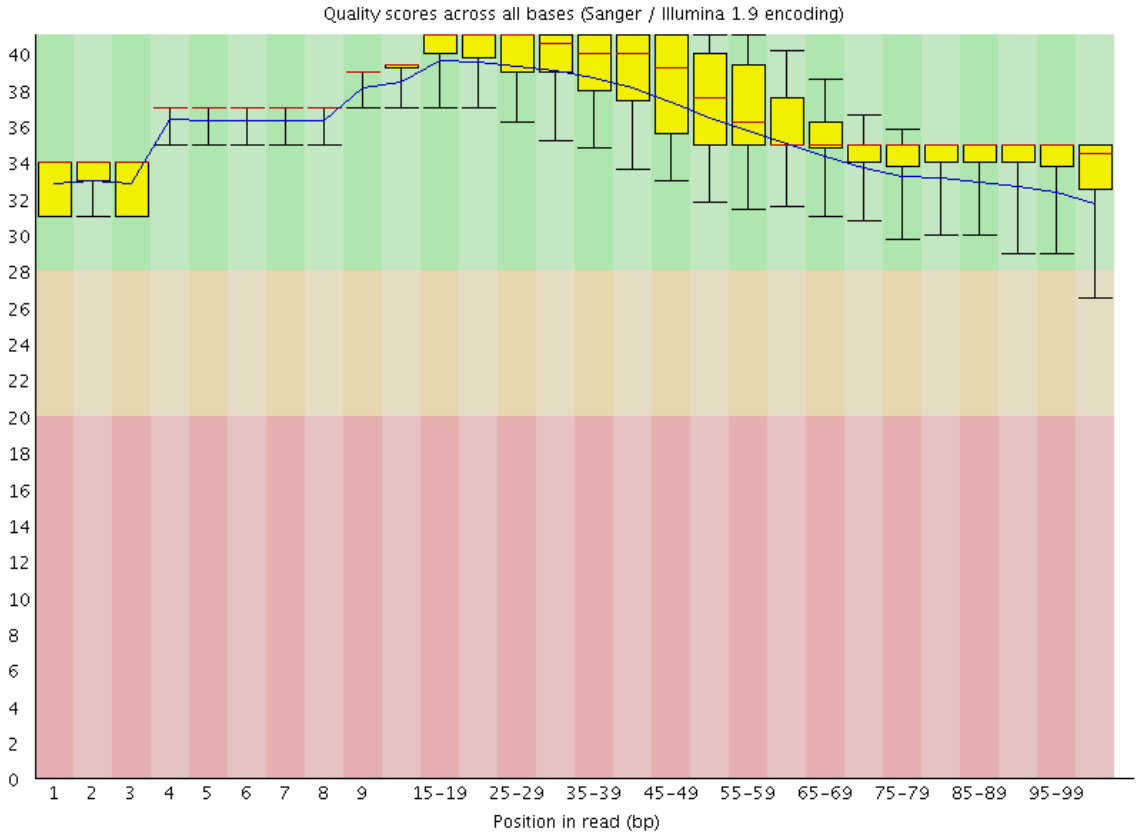

Figure 3-5. Example fastQC per-base quality report. Each bar represents the range of quality scores for the given positional range of nucleotides, beginning with the first nucleotide in each adapter-trimmed read and proceeding to the end of the read. The quality score lower error bar was used to determine if this sample was of sufficient quality to proceed. No further trimming was required for this (or any) sample as all positions of all reads were above the quality cutoff score of 20.

3.4 Gene counts

The number of reads mapping to each gene feature in the corresponding annotation files was determined using HTSeq (66). The output of this tool was a count table with the number of reads mapping to each gene in the reference annotation. These "count tables" formed the basis for all further analysis.

Count table data for human and parasite samples collected in this study was combined with data from our previous work. These combined data were checked for consistency between replicates using a variety of global inter- and intra-sample correlation analyses, as described for the human data in Chapter 4 and for the parasite in Chapter 5.

Chapter 4. Human Data

4.1 Library size

To visualize the depth of sequencing of the various samples, we plotted the raw library size of all samples as shown in Figure 4-1. All further analysis is performed on library-size normalized data to account for these differences in sequencing depth.

An inherent assumption for quantile normalization is that the count data has the same underlying distribution. To evaluate the data against this assumption, a density plot of library size-normalized gene counts (82) for each sample was generated as shown in Figure 4-2. This density analysis showed the estimated underlying distribution of abundance of reads mapping to each gene feature. The near superimposed curves showed that the distribution of counts was consistent within human samples. Human gene expression is frequently regulated at the transcriptional initiation level, with many genes turned off under different conditions (83), and this is reflected in the spike at zero count. This spike does not affect the downstream data analysis because all low-count genes are dropped prior to data normalization (68). Other than this spike, the remainder of the expression pattern has the same underlying distribution, so our assumption for quantile normalization was met.

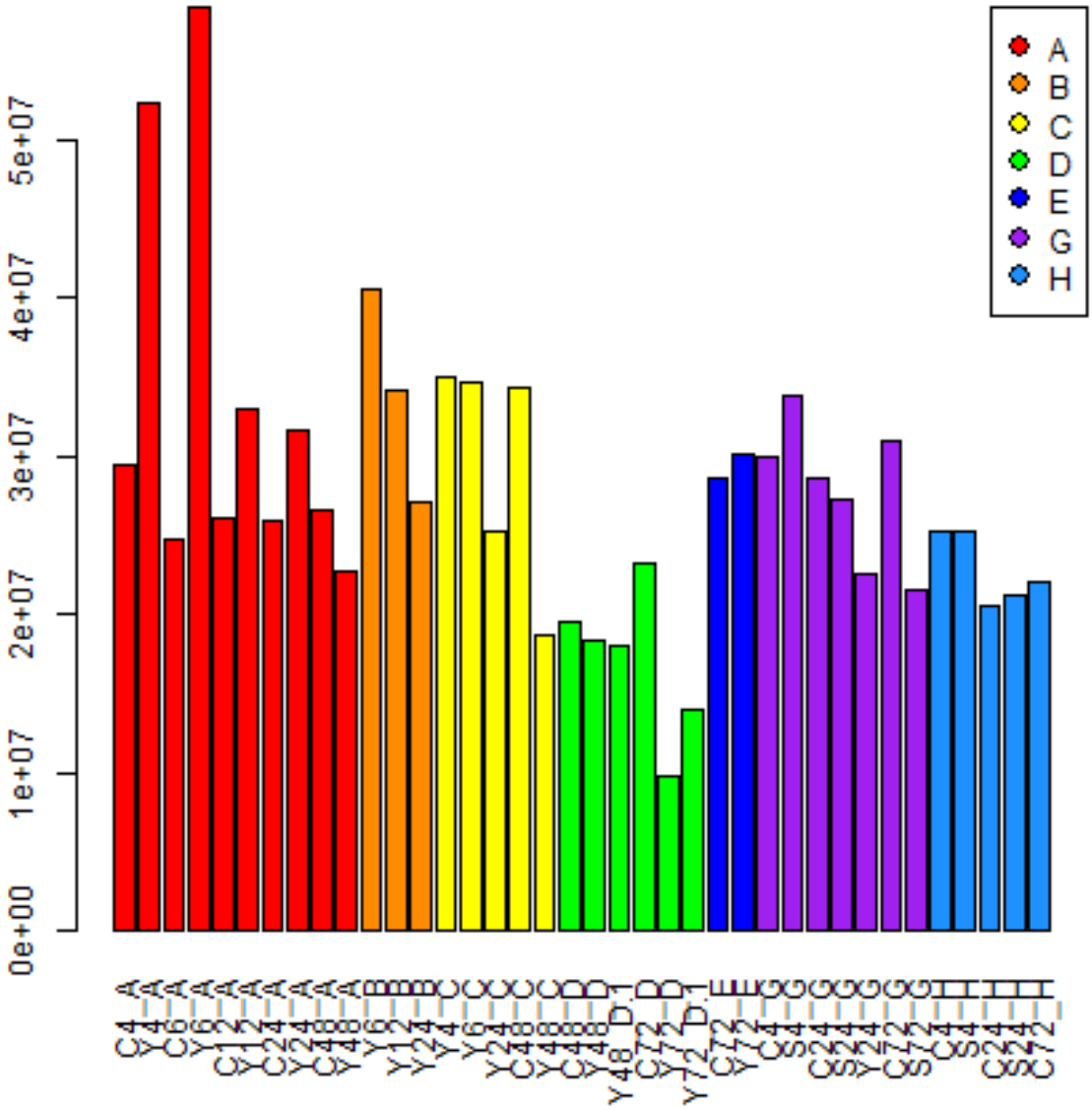

Figure 4-1. Barplot of raw library sizes for all human samples. Each bar represents the number of reads for a single sample in the combined dataset. Color represents a experimental batch, from experiments performed starting on the same day. Batches G and H were produced as part of this research.

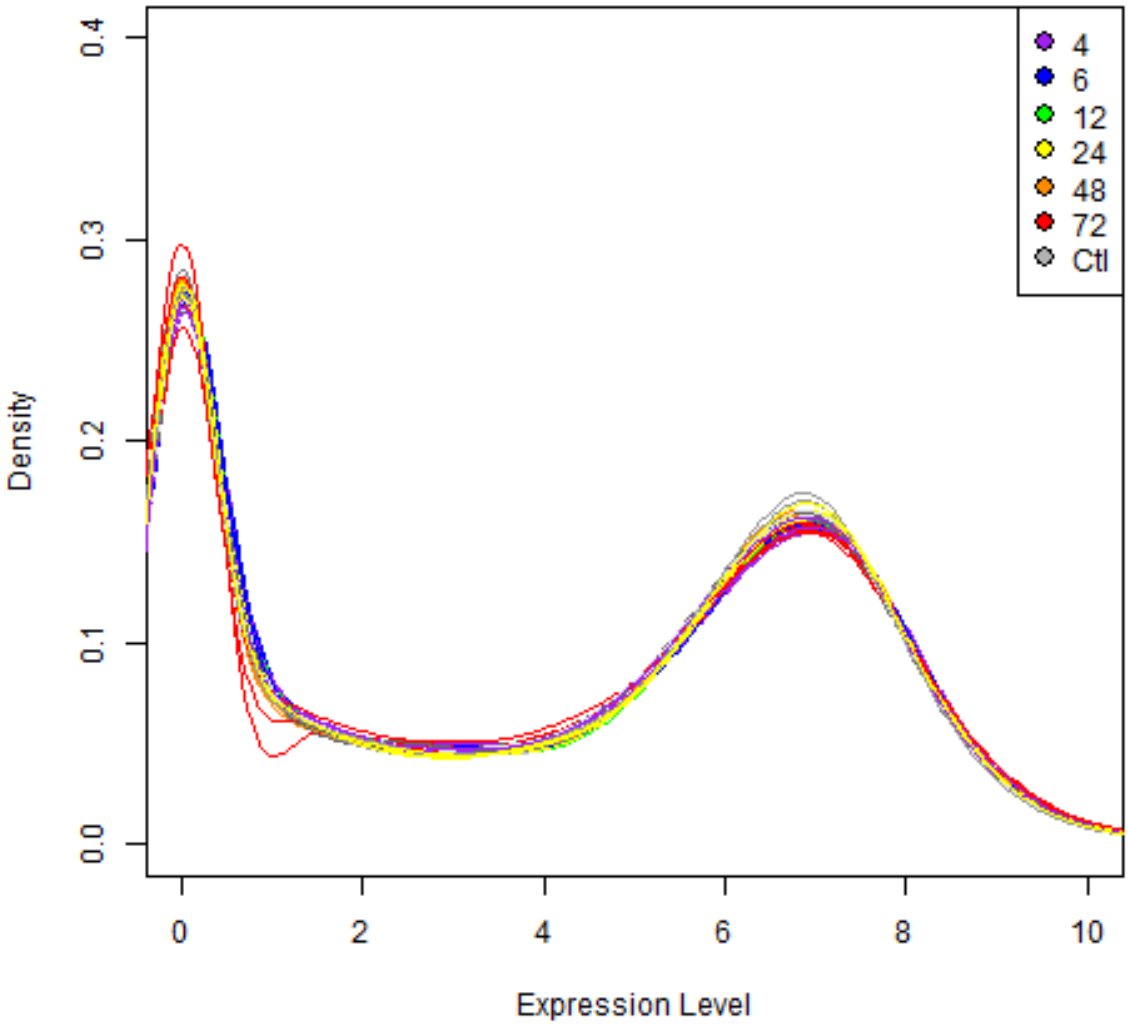

Figure 4-2. Density plot of all human samples. Each curve represents an individual sample in the combined dataset. Ideally, all curves would be superimposed, and the data would show the same underlying distribution. Counts (expression level) are library size-normalized to address the differences in sequencing depth across batches, and $\log_2$ transformed.

A different view of the same data, focusing on the mean count, is shown in Figure 4-3. The library size normalized mean count and range did not differ substantially between samples.

4.2 Identification of outlier samples

A Pearson correlation was computed for all samples against all samples in the combined dataset. A heatmap showing the resulting correlation matrix is shown in Figure 4-4.

There was generally very good inter-sample correlation across the entire dataset. The dendrogram highlights that there were clusters of correlation, which roughly correspond to samples taken at particular timepoints. Early timepoints (4-12 hpi) tended to cluster together, as did middle (24 hpi) and late (48-72 hpi) timepoints. The two datapoints at the left/bottom of the chart (red) had the lowest correlation, and were dropped after subsequent analysis.

A standard method for identifying outliers was applied to the data samples (67). The median pairwise correlation (MPC) for the first quartile across samples was computed (Q1), as was the inter-quartile range (IQR). Briefly, any sample whose MPC was less than Q1-1.5(IQR) was

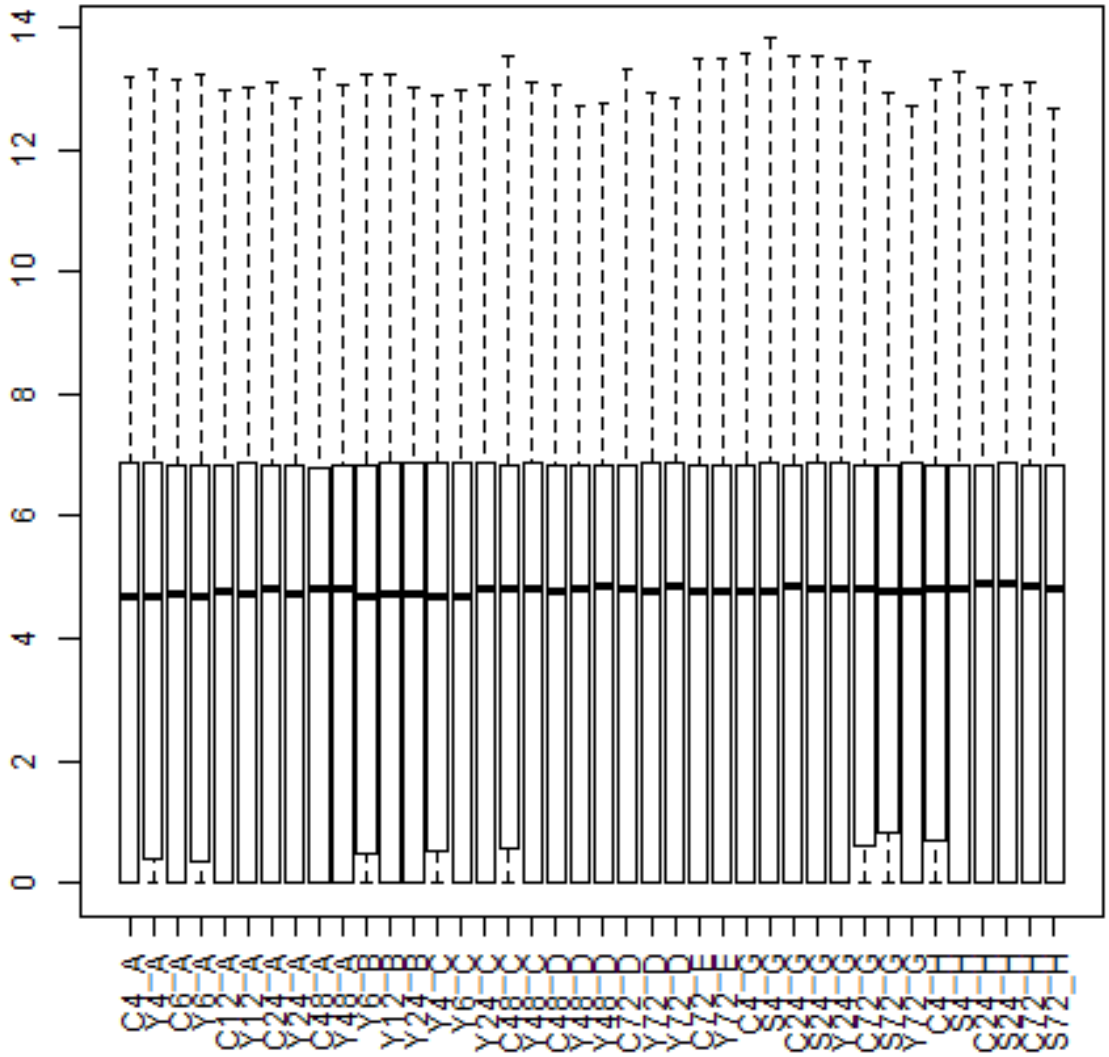

Figure 4-3. Boxplot of library size normalized counts by sample. Each box represents an individual sample in the combined dataset. Ideally, all boxes would have exactly the same mean (horizontal line) and range. Counts (expression level) are library size-normalized to address the differences in sequencing depth across batches, and $\log_2$ transformed

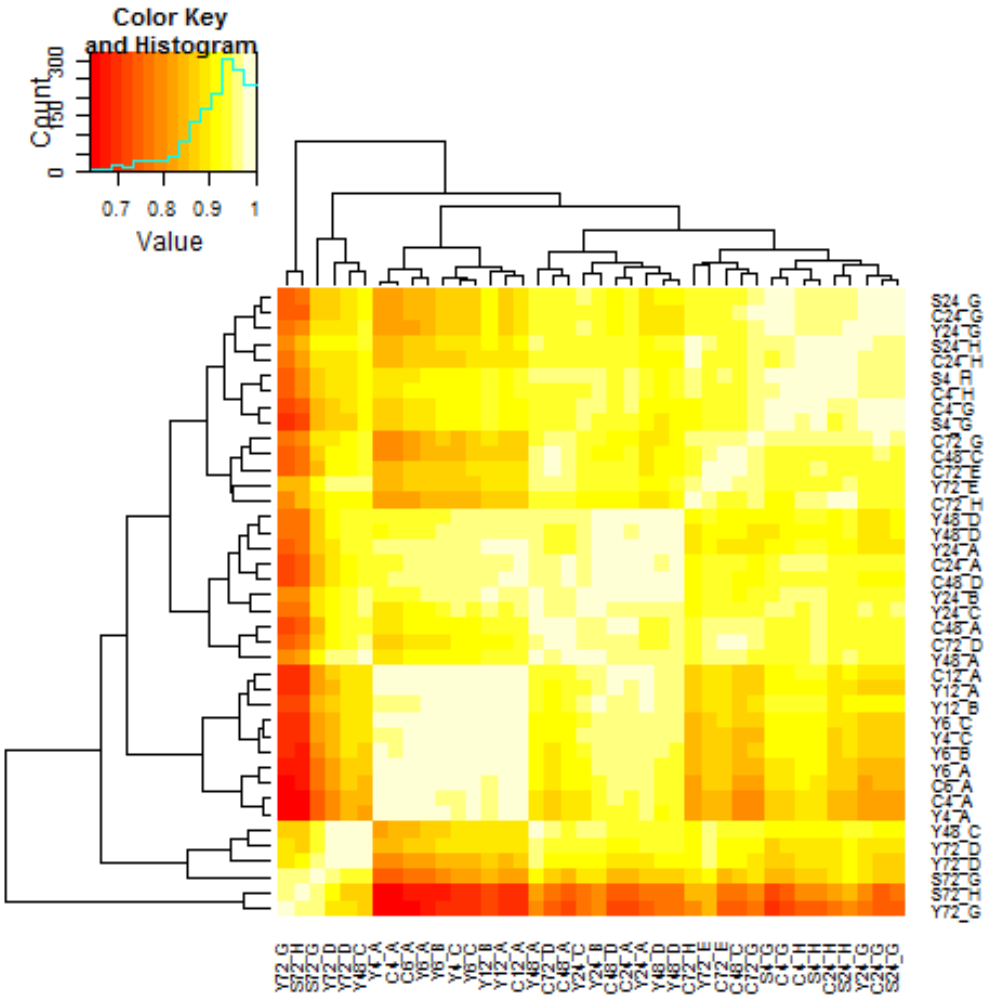

Figure 4-4. Pearson correlation heatmap. This heatmap visualizes the Pearson correlation matrix for each sample compared to every other sample in the combined dataset. Color at the intersection of each sample pair represents the Pearson correlation coefficient for that sample pair.

removed as an outlier. A scatterplot of median pairwise correlation for each sample is shown in Figure 4-5.

Two samples (Y72_G and S72_H) were removed from the human data based on this criterion.

4.3 Pairwise comparisons

To evaluate variance between pairs of samples, we produced a series of pairwise MA plots. Examples of these plots are shown in Figure 4-6.

These plots compare the ratio of expression levels for each gene (log count per million) between two samples versus the average expression. Essentially, the further the ratio is from the center line, the more variance there is in that sample. Figure 4-6A is an example of two samples with low variance per gene. Figure 4-6B is an example of two samples with considerable variance. There were a few pairs of samples that showed a large amount of dispersion, particularly early timepoints compared to late ones, however most sample pairs showed a low variance pattern.

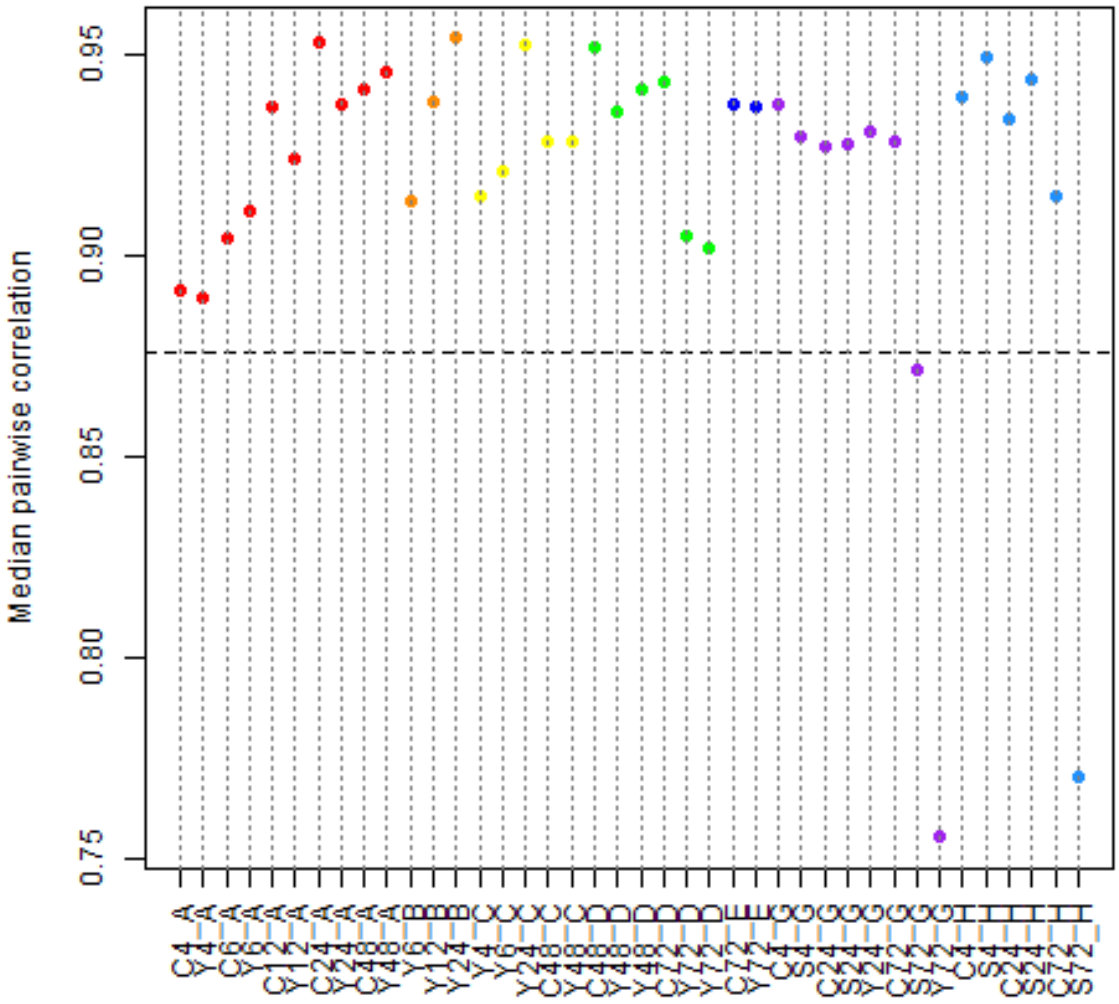

Figure 4-5. Scatterplot of median pairwise Pearson correlation. Each point represents the median pairwise correlation for that sample with all other samples in the dataset. The horizontal line is the interquartile range cutoff, computed as median pairwise correlation (MPC) for the first quartile across samples – 1.5 times the inter-quartile range (IQR).

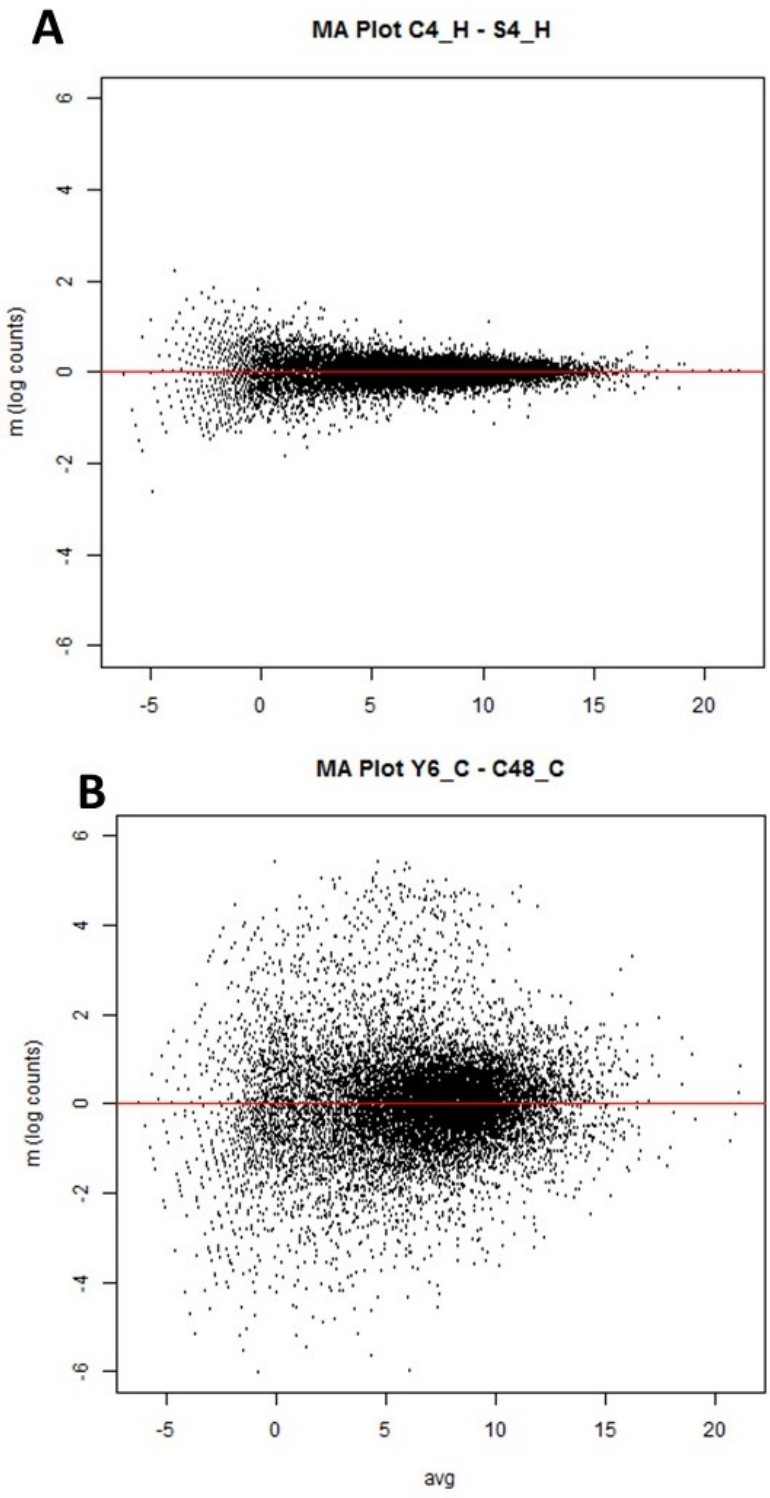

Figure 4-6. Sample MA plots for evaluating pairwise data dispersion between samples. A. Comparison of a control and infected human sample at 4hpi in the same batch. B. Comparison of a control sample at 48 hpi compared to an infected sample at 6 hpi.

4.4 Principal component analysis

A principal component analysis was performed on low-count filtered, quantile normalized per-gene counts from both species, incorporating data produced in this study with the Y strain data from our previous study. Principal Component analysis extracts important information from a table of data, combines it into a set of "principal components" which represent patterns of similarity and variability in the data, and allow that similarity/variability to be visualized (84). This allowed us to evaluate any potential batch effects introduced by the different labs and the different experimental timeframes. Batch effects are confounding variables, such as date of experiment, lab or technician, which cause variation in the data completely separate from the underlying biology (72). These batch effects can confound the results of biological studies and should be identified and corrected. (85)

Low count, or weakly expressed genes, are 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 et al., 2013) (here n= 3 for human samples). These low-count reads were removed from subsequent analyses. Of 21036 genes with Hugo annotations (protein coding genes) in the count tables, 13063 (62%) passed the low-count filter and continued on to principal component analysis.

The low count filtered data were quantile normalized and $\log_2$ transformed (69).

The singular value decomposition of the normalized and transformed combined dataset was computed. The resulting covariance matrix of the genes forms the basis of the principal component scores (86). To evaluate which genes had the biggest impact on

principal component 1, which explains the largest amount (35%) of the variance, sorted loadings were plotted across all genes as shown in Figure 4-7.

Of 2206 low-count filtered genes, 13063 (17%) had a loading greater than 0.01 (knee of the curve). The variance matrix was then evaluated to determine the variance of each principal component and how that variance correlates with experimental batch and experimental condition (Control or infected combined with hpi). The resulting table is shown in Table 4-1.

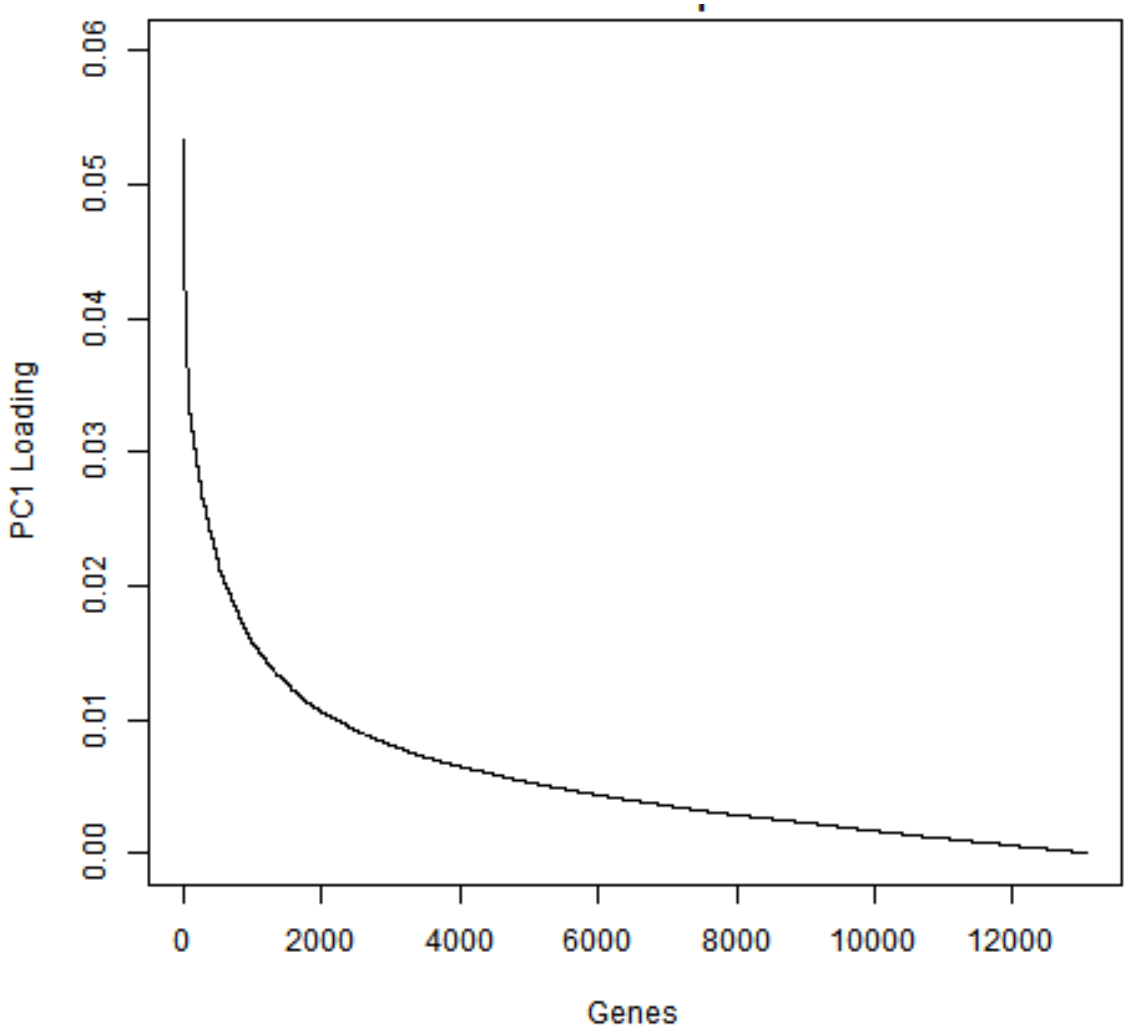

Figure 4-7. Loading analysis of uncorrected PC1. Sorted loadings were plotted across all genes.

Table 4-1. Variance for each principal component without batch correction. For each principal component, the proportion of total variance in the dataset explained by that PC, and the contribution of condition (control or infected and timepoint) and experimental batch for that PC are reported.

PC	Proportion Variance	Condition Contribution	Batch Contribution
1	35.44	88.55	45.59
2	27.65	68.34	90.49
3	11.4	83.88	11.26
4	5.02	81.96	16.9
5	3.24	34.77	72.87
6	2.3	15.9	64.48
7	1.88	63.05	31.15
8	1.51	48.76	31.49
9	1.24	27.58	19.19
10	1.03	26.3	59.34

We then plotted the Principal component scores for PC1 against PC2 for each sample as shown in Figure 4-8.

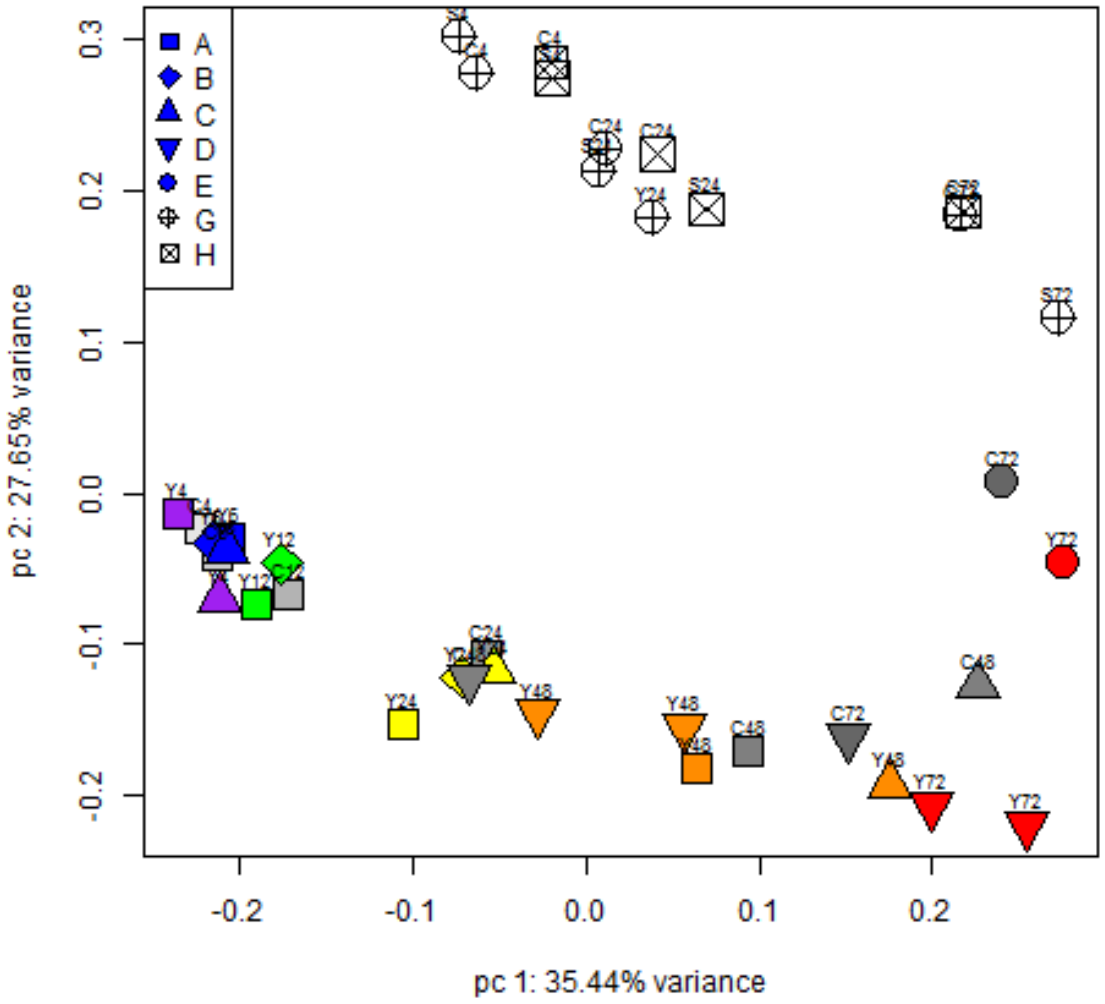

Figure 4-8. Uncorrected principal component analysis for human data. Principal component score for PC1 is plotted against the principal component score for PC2 for each sample. Color represents time hpi for infected samples. Control samples are shown in increasingly dark grayscale for increasing time hpi. Shape corresponds to experimental batch as shown in the key. Batches G and H were produced as part of this research, and batches A-E came from our previous work.

In the uncorrected PCA, principal component 1 is associated with time hpi, with early timepoints clustering toward the left and later timepoints clustering on the right. Principal component 2 appears highly associated with experimental batch, with samples produced in our previous work on the bottom and samples produced as part of this work on the top. This is evidence of batch effect in this data, so we included batch as a covariant in the linear model incorporated into the limma program (87) used to evaluate differential expression.

One of the underlying assumptions in these linear models is the independence of mean and variance in gene counts. However, in this type of count-based data, variance is often larger at lower mean counts. To include batch in the linear model, and to also correct for the mean-variance relationship, we used the voom package (71) to weight each gene. The mean-variance trend computed by voom as part of data weighting is shown in Figure 4-9.

After the data was weighted and batch was included in the linear model, the principal component analysis was repeated. The variance matrix was recomputed, and a new singular value decomposition was produced to determine the new variance of each principal component for the corrected data. The resulting table is shown in Table 4-2.

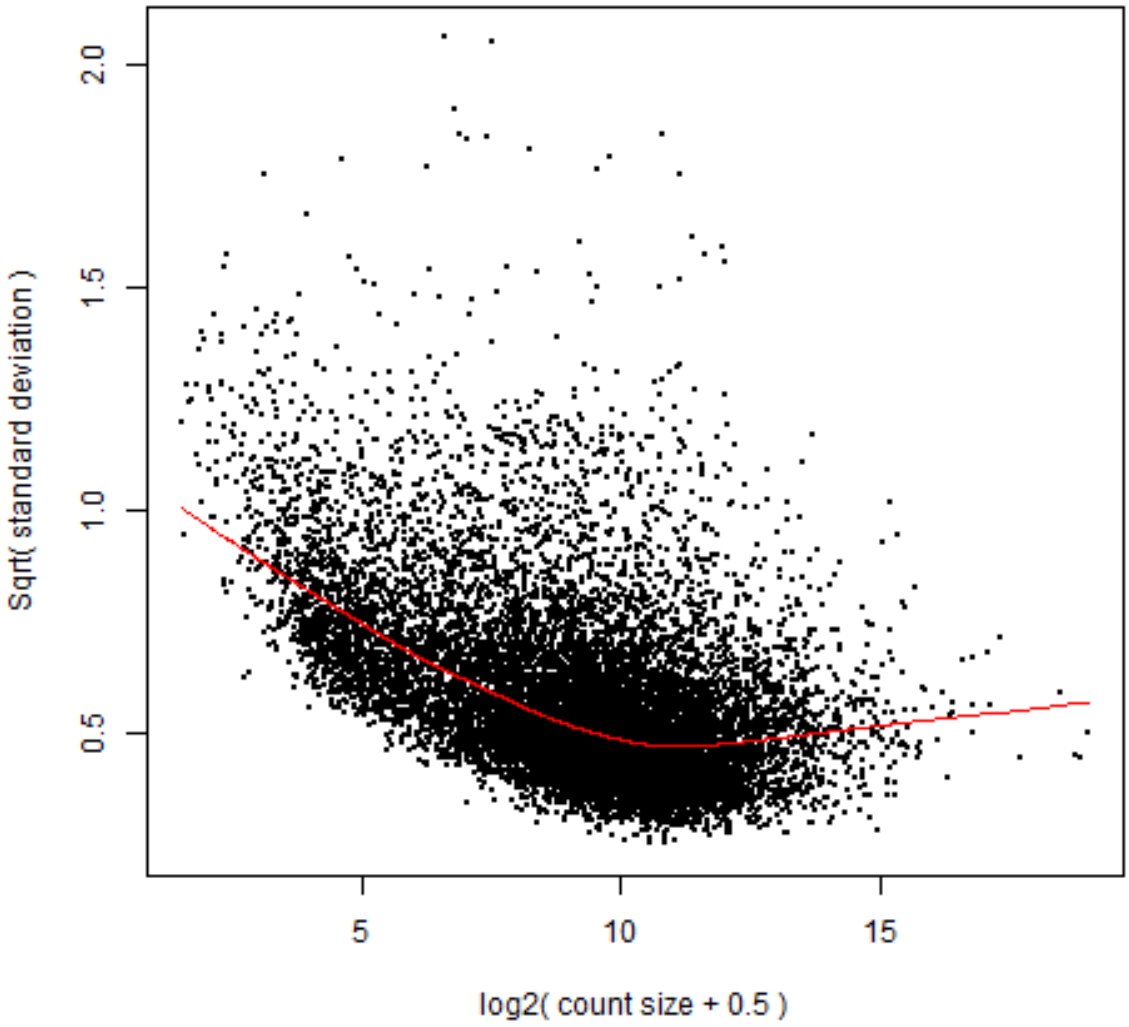

Figure 4-9. Mean-variance trend in uncorrected combined Human data. The trendline is produced by the voom package as part of computing weights to correct for mean-variance relationship in count-based data.

Table 4-2. Variance for each principal component with batch included in limma model. For each principal component, the proportion of total variance in the dataset explained by that PC, and the contribution of condition (control or infected and timepoint) for that PC are reported. The influence of batch has been removed by including it as a covariant in the model.

PC	Proportion Variance	Condition Contribution
1	44.3	58.33
2	20.49	82.89
3	9.38	85.7
4	3.68	62.37
5	2.67	27.55
6	2.56	43.27
7	1.86	52.42
8	1.57	52.32
9	1.4	68.41
10	1.17	30.38

We then plotted the new Principal component scores for PC1 against PC2 for each sample as shown in Figure 4-10.

Principal component 1 (PC1) showed a progression of time post-infection, with early timepoints clustering on one end and later timepoints on the other. Control (uninfected) samples clearly separated from infected samples along PC2, with the distance between control and infected cells growing larger as the infection progressed. Datapoints from this study, experimental batches G and H, grouped tightly with the data produced previously despite the experiments being run in two different collaborating laboratories more than a year apart. This analysis also convinced us that including batch in the limma model was sufficient to account for batch effects, and that further batch correction was unnecessary for differential expression analysis.

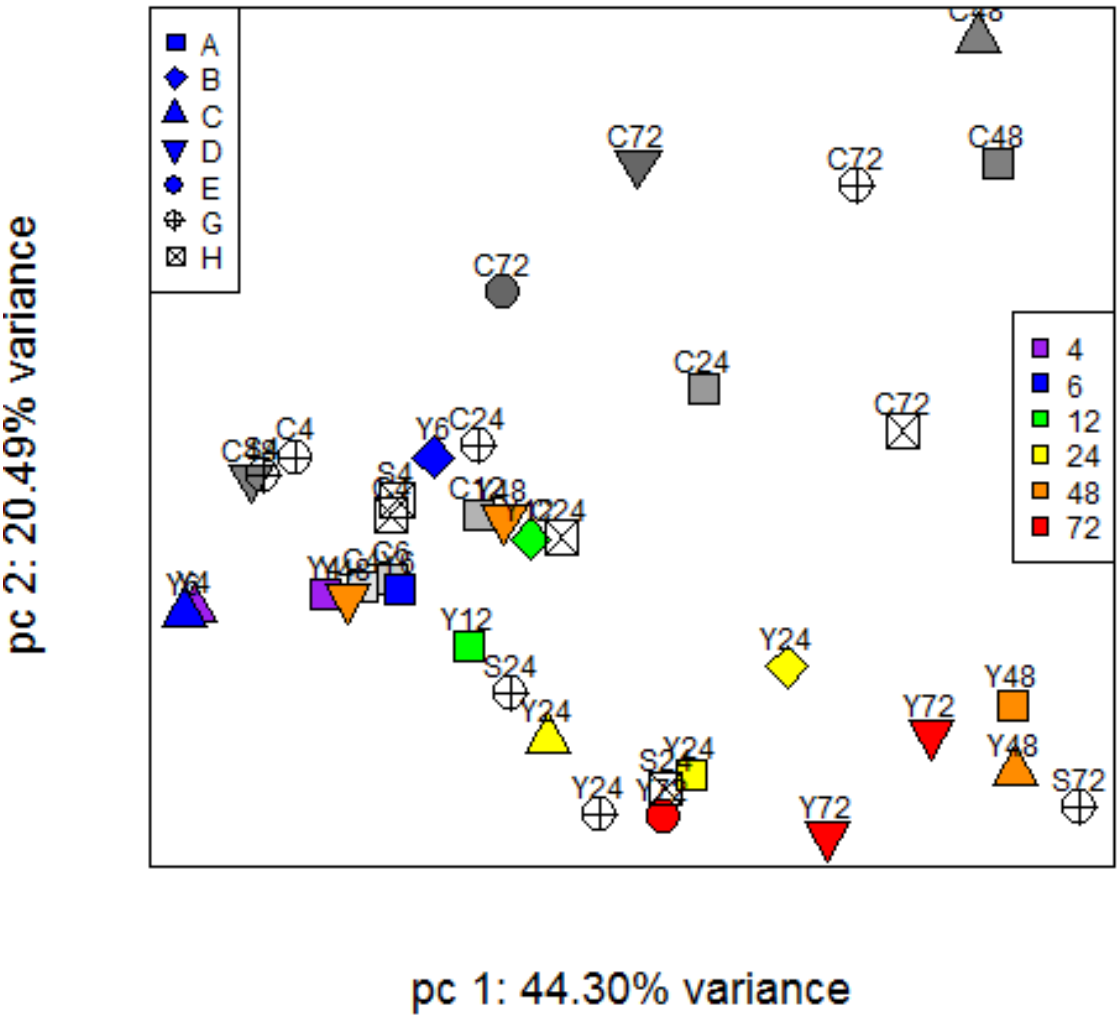

Figure 4-10. Principal component analysis, batch in limma model. Principal Component score for PC1 is plotted against the Principal component score for PC2 for each sample. Color represents time hpi for infected samples as shown in key. Control samples are shown in increasingly dark grayscale for increasing time hpi. Shape corresponds to experimental batch as shown in the key. Batches G and H were produced as part of this research, and batches A-E came from our previous work.

4.5 Differential expression analysis

There were 13892 a total of expressed (non-zero) genes in the human dataset. Differential expression for the human data was computed for each timepoint subtracting the effects of the control cells because, as seen in the principal component analysis, HFF cells change over time in culture. Differential expression was computed on these control-corrected gene counts.

The limma topTable function (73) was utilized to provide summary statistics on a per-gene basis at each timepoint for differential expression analysis. Log_2(Fold Change) was returned from the linear model to represent difference in expression between infected and control HFFs. The primary method of determining significance for each gene is the moderated t-statistic. The moderated t-statistic uses increased degrees of freedom in connection with underlying smoothing across genes. Both p-values as well as adjusted p-values (using Benjamini-Hochberg multiple-testing adjustment) are provided. Finally, the B statistic is provided which represents the log odds that the gene is differentially expressed. This value usually orders genes the same way as p-value.

The q-value, which is an adjusted p-value using prior probabilities of each gene not being differentially expressed (Bayesian posterior p-value), was also computed (88).

The number of Differentially Expressed genes for the human data is shown in Table 4-3.

A graphical representation of the similarities and differences in the host response to infection with the two different strains of the *T. cruzi* parasite is shown in the form of a scatter plot of log_2 fold change to visualize the shared signature (Figure 4-11).

Overall, a striking similarity was seen between the host response to Sylvio and Y strains of *T. cruzi* when comparing genes that are significantly DE in both infections, and had a fold change greater than 2 (89). At 24 hpi, HFFs infected with Sylvio or Y shared a common set of 113 genes that were significantly upregulated (Figure 4-11A**,** quadrant I). There were no significantly DE genes co-downregulated or upregulated exclusively in Sylvio or Y infections at 24 hpi. There was a marked increase in the number of shared genes significantly DE between 24 hpi and 72 hpi. By 72 hpi (Figure 4-11B), the number of genes that were significantly DE in HFFs infected with either Sylvio or Y in the same direction increased to 358. Of these, 291 (81%) were commonly upregulated and 67 (19%) were downregulated (Figure 4-11B, quadrants I and III respectively). An examination of these shared genes revealed that of the 113 genes significantly DE at 24 hpi, 102 (90.3%) were still significantly DE at 72 hpi. We conclude that most of the early response signature remained over the course of the infection with an additional modulation of host cell genes as the infection progressed.

Table 4-3. Summary of differential expression analysis for human data. Summary of differentially expressed genes with their direction (up- or downregulated) at various fold change cutoffs at various timepoints. To be considered differentially expressed, a gene must have a q-value < 0.05.

24 hpi	Sylvio DE genes	Y DE genes	Shared DE genes, no FC cutoff		Shared DE genes, -1<log2FC<1 (pink)		Shared DE genes, |log2 FC| > 1 (red)	
Significant DE Genes	316	7690	289		176		113	
Both Upregulated (Q1)	260	3512	243	84.1%	130	73.9%	113	100.0%
Y Up, Sylvio Down (Q2)			0	0.0%	0	0.0%	0	0.0%
Both Downregulated (Q3)	56	4178	46	15.9%	46	26.1%	0	0.0%
Sylvio Up, Y Down (Q4)			0	0.0%	0	0.0%	0	0.0%
72 hpi								
Significant DE Genes	6526	4301	2227		1828		399	
Both Upregulated (Q1)	2972	1930	944	42.4%	653	35.7%	291	72.9%
Y Up, Sylvio Down (Q2)			207	9.3%	184	10.1%	23	5.8%
Both Downregulated (Q3)	3554	2371	891	40.0%	824	45.1%	67	16.8%
Sylvio Up, Y Down (Q4)			185	8.3%	167	9.1%	18	4.5%
	no FC cutoff				|log2 FC| > 1			
Shared Over Time	24 hpi	72 hpi	Merge		24 hpi	72 hpi	Merge	
Both Upregulated (Q1)	243	944	206	84.8%	113	291	102	90.3%
Both Downregulated (Q3)	46	891	9	19.6%	0	67	0	

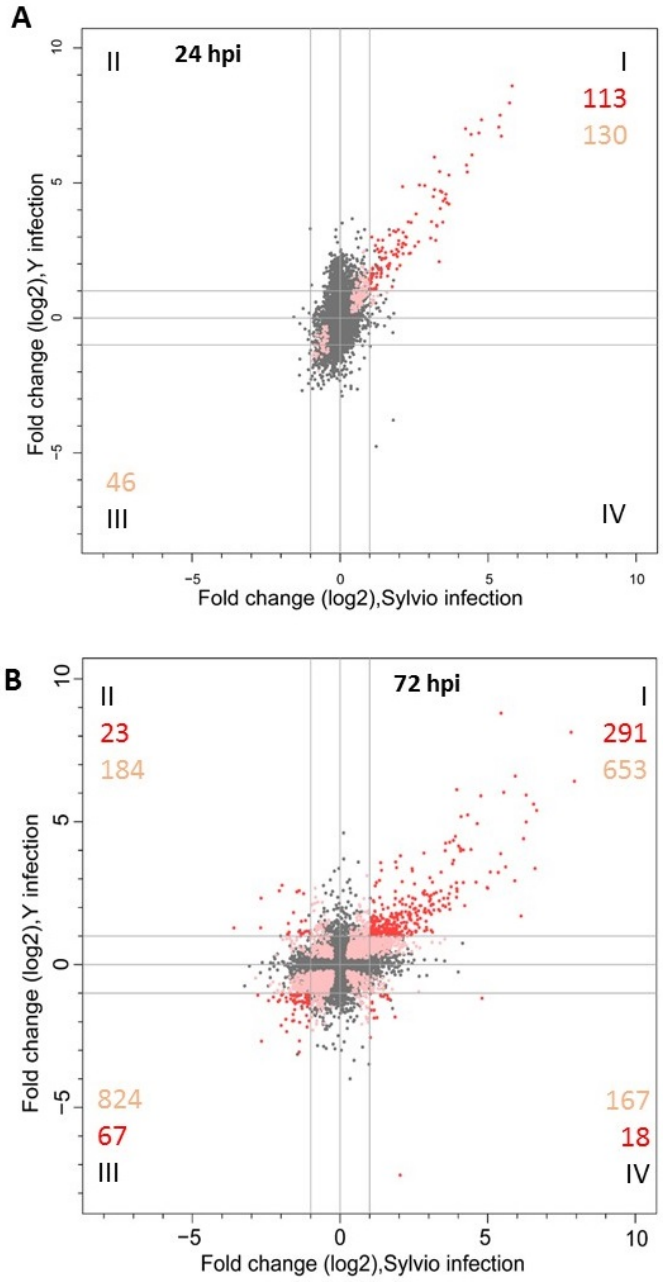

Figure 4-11. Comparative analysis of the host transcriptome response to infection with Sylvio versus Y strain. Scatterplot of log_2 fold change in expression of human genes in HFF cells infected with Sylvio (x-axis) versus Y (y-axis) at 24 hpi (A) and 72 hpi (B). Genes that were not significantly differentially expressed are shown as gray dots. Pink dots represent significantly DE genes in infections with both strains but had log_2 FC in one or both strains that was below our cutoff value of 1. Red dots depict genes that were significantly DE and had absolute value of log_2 FC>1 in both strains. The numbers in each quadrant correspond to the total number of significantly DE genes that met the log2 FC cutoff of 1 (red) or were significantly DE but did not meet the FC cutoff (pink).

4.6 Gene ontology analysis

To evaluate the biological activity being performed by the host cell for this shared signature, the GO Term Finder (77) was applied to the human data using the server at go.princeton.edu. Ensembl IDs were converted to HUGO Gene Nomenclature Committee Database (HGNC) IDs with the BioMart ID Conversion tool, then GO Term Finder was applied with the "GOA + HGNC Xrefs- H. sapiens (human)" annotation. The human data was also interrogated with g:Profiler (78) to simultaneously identify GO, KEGG and Reactome term enrichment (90, 91).

As shown in Table 4-4, GO analysis of the co-upregulated genes shows a general innate immune response at 24 hpi, with Type I interferon response being the most vigorous (three out of the top ten GO terms).

A similar enrichment was observed at 72 hpi as shown in Table 4-5.

A KEGG analysis showed an upregulation in sterol biosynthesis pathway at 72 hpi. A GO analysis of the significantly DE co-downregulated genes (Figure 4-11B, Quadrant III) showed a downregulation of the cell cycle. All of these observations were consistent with our previous work (62).

Table 4-4. Top ten GO terms for host genes upregulated in infections with both Sylvio and Y strains at 24 hpi. A list of all significantly upregulated genes in both Sylvio and Y strains which also had a log2 Fold Change > 1 were provided to the GO Term Finder

GO ID	GO TERM	CORRECTED P-VALUE	NUMBER OF GENES	INPUT LIST SIZE
GO:0051607	defense response to virus	9.79E-53	39	111
GO:0009615	response to virus	3.38E-52	41	111
GO:0045087	innate immune response	2.44E-51	56	111
GO:0006955	immune response	1.73E-48	62	111
GO:0006952	defense response	4.88E-48	61	111
GO:0098542	defense response to other organism	7.27E-47	41	111
GO:0060337	type I interferon signaling pathway	2.95E-43	28	111
GO:0071357	cellular response to type I interferon	2.95E-43	28	111
GO:0034340	response to type I interferon	4.44E-43	28	111
GO:0002376	immune system process	5.35E-43	65	111

Table 4-5. Top twenty GO terms for host genes upregulated in infections with both Sylvio and Y strains at 72 hpi. A list of all significantly upregulated genes in both Sylvio and Y strains which also had a $\log_2$ Fold Change > 1 were provided to the GO Term Finder

GO ID	GO TERM	CORRECTED P-VALUE	NUMBER OF GENES	INPUT LIST SIZE
GO:0002376	immune system process	2.74E-59	121	281
GO:0006955	immune response	3.25E-58	104	281
GO:0045087	innate immune response	6.82E-56	86	281
GO:0006952	defense response	3.28E-55	100	281
GO:0043207	response to external biotic stimulus	7.33E-44	63	281
GO:0051707	response to other organism	7.33E-44	63	281
GO:0009607	response to biotic stimulus	1.53E-42	63	281
GO:0034097	response to cytokine	2.32E-42	64	281
GO:0009615	response to virus	7.10E-41	47	281
GO:0051607	defense response to virus	5.88E-40	43	281
GO:0006950	response to stress	7.03E-40	126	281
GO:0009605	response to external stimulus	1.95E-38	93	281
GO:0010033	response to organic substance	1.05E-37	100	281
GO:0098542	defense response to other organism	6.00E-37	48	281
GO:0060337	type I interferon signaling pathway	6.77E-37	32	281
GO:0071357	cellular response to type I interferon	6.77E-37	32	281
GO:0034340	response to type I interferon	8.57E-37	32	281
GO:0071345	cellular response to cytokine stimulus	4.36E-36	56	281
GO:0051704	multi-organism process	5.68E-36	85	281
GO:0019221	cytokine-mediated signaling pathway	2.81E-35	52	281

We infer from our observations that the host cell transcriptomes when infected with either Sylvio or Y strain of *T. cruzi* are remarkably consistent.The signature behavior of the host when infected with *T. cruzi* includes upregulation of the innate immune response, specifically the Type I interferon response, an upregulation in the sterol biosynthesis pathway, and a downregulation in the mitotic cell cycle. All of these are consistent with our previous work, and are also consistent with observed behavior in these and previous published experiments.There were a few detectable differences in the host cell response to the two strains which did not appear until 72 hpi. A total of 41 genes were upregulated in only one strain and downregulated in the other. Of those, 23 genes were upregulated in Y strain infections and downregulated in Sylvio infections (Figure 4-11B, quadrant II) and do not reflect any discernible signature. The remaining 18 genes were upregulated in Sylvio infections and downregulated in Y strain infections (Figure 4-11B, quadrant IV). While we did not identify any GO categories that were significantly overrepresented among those genes, we note that the genes encoding IL-8 and IL-32 behave very differently in host cells infected by the two different *T. cruzi* strains. IL-8 is a chemokine which attracts monocytes to an area of injury or infection. IL-32 causes the production of pro-inflammatory cytokines and chemokines, including IL-8. One can speculate that the upregulation of these two cytokines in Sylvio infected cells could offer a partial explanation of the difference in the overall clinical picture between the infections with the two strains by attracting immune cells to respond to host cells infected with Sylvio. This biological relevance of this observation can only be ascertained by testing a greater number of replicates *in vitro*, testing additional parasite strains as well as

establishing that similar expression patterns can be observed in patients infected with the two strains.

We compared the top differentially expressed genes in Sylvio to the list of top differentially expressed genes in a study of Tulahuen strain (TcVI) *T. cruzi* infecting HeLa cells (53). The results of shared DE genes are shown in Table 4-6. These shared genes show the same interferon-induced picture as the Sylvio and Y strain, and also show strong upregulation of IL-8 as seen in Sylvio, but not Y strain.

Table 4-6. Top ten shared genes upregulated in HeLa cells infected with Tulahuen strain of *T. cruzi*. Gene ID in the original study, gene function, Log2 Fold change in Tulahuen, Ensembl ID for the gene in this study and Log2 Fold Change in Sylvio infected HFF cells.

Gene	Description	log2FC	Ensembl ID	log2FC
IL8	interleukin 8	2.9	ENSG00000169429	4.8
IL6	interleukin 6 (interferon, beta 2)	2.9	ENSG00000136244	2.4
HERC5	hect domain and RLD 5	2.6	ENSG00000138646	5.9
IFI44	interferon-induced protein 44	2.5	ENSG00000137965	3.6
IFIT2	interferon-induced protein with tetratricopeptide repeats 2	2.4	ENSG00000119922	6.7
PHLDA1	pleckstrin homology-like domain, family A, member 1	2.3	ENSG00000107679	-0.8
IFIT3	interferon-induced protein with tetratricopeptide repeats 3	2.3	ENSG00000119917	4.1
SLC22A18AS	solute carrier family 22 member 18 antisense	2.2	ENSG00000110628	1.0
IFIT1	interferon-induced protein with tetratricopeptide repeats 1	2.2	ENSG00000185745	4.4
GLRX	glutaredoxin (thioltransferase)	2.2	ENSG00000173221	0.8

Chapter 5. Parasite Data

5.1 Parasite orthologs

Because the annotation and closure of the *T. cruzi* genomes are considerably less mature than the human genome, and the gene naming scheme between CL Brener Esmeraldo-Like (proxy for Y strain) and Sylvio strain was completely different, we generated a set of orthologs to allow mapping of genes between the two strains. Version 8.0 genome fasta files for *T. cruzi* Sylvio strain and *T. cruzi* CL Brener strain (Esmeraldo-like haplotype) were used to generate these orthologs. The FASTA program ggsearch36 (76), which does a global alignment, was used with default parameters (E-value <= 0.001, score > 0). Reciprocal best hits provided 6273 orthologs from Sylvio to Esmeraldo-like, and 6346 orthologs in the opposite direction. This is comparable to the 6094 bidirectional best hits BBH reported between *T. cruzi* Dm28c and Sylvio strains, and 5267 BBH between Dm28c and CL Brener. (51) They are also comparable to the number of reported orthologs between T. *cruzi marinkellei* and Sylvio strain (6283), Esmeraldo-like haplotype of CL Brener (5441) and non-Esmeraldo-like haplotype (5617) (52).

The orthologs lists were then further refined by removing genes in the four largest multi-gene families (MASP, trans-sialidase, retrotransposon hot-spot and dispersed gene family protein), thereby avoiding the problems introduced by assembly of highly similar and repetitive genes.

Altogether there were 4,659 unique orthologs, of which 3,279 (70%) were annotated as hypothetical proteins. This set of genes is referred to as the “core” set of parasite genes. This list allowed us to unambiguously compare expression levels for orthologous genes in both strains despite the different naming conventions for genes in Sylvio and Y strains. The process and gene counts during the various steps of developing the core set of orthologs is shown in Figure 5-1.

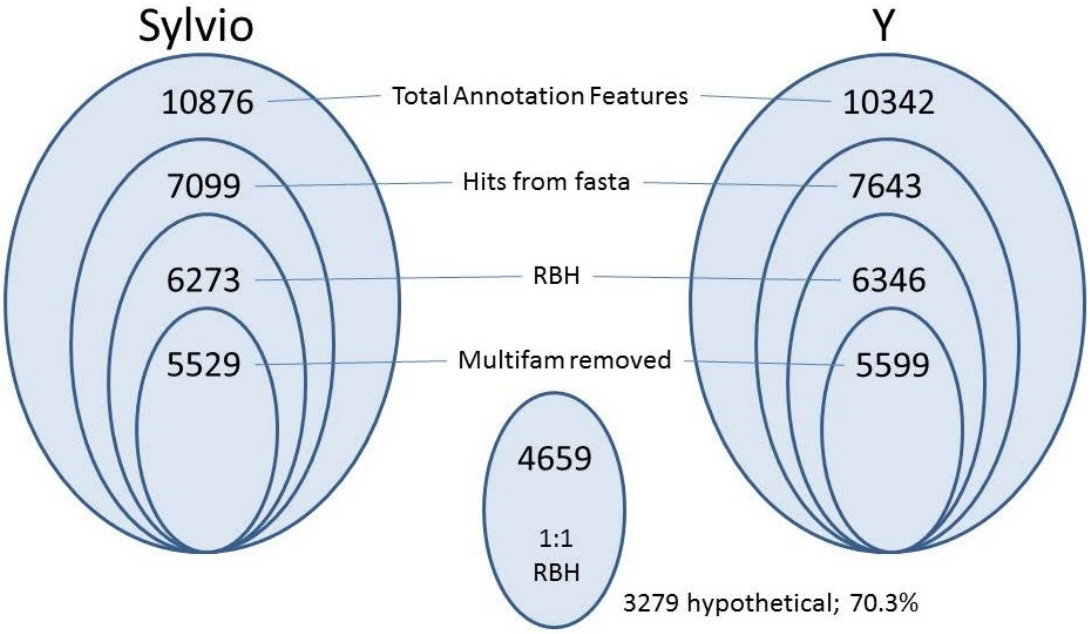

Figure 5-1. Process of developing core set of orthologs. Each set of ovals represents a step in the process of developing the core ortholog set. “Total annotation features” are the protein coding genes in the respective .gff file from TriTrypDB. “Hits from fasta” are the genes that had any number of orthologs between the two strains using ggsearch36. “RBH” are the genes that had mutual best hit pairs. “Multifamily Removed” was the process of removing the top four multigene families by name. The last step finds unique reciprocal best hits in both strains. Only the top four hits were returned, the E score had to be less than 0.001, and the match score had to be greater than 0 for ggsearch36 to return a hit.

5.2 Library size

To visualize the depth of sequencing of the various samples, we plotted the raw library size of all samples as shown in Figure 5-2. All further analysis is performed on library-size normalized data to account for differences in sequencing depth.

For the parasite data, batches G and H were combined into one batch because the second replicate of Sylvio data being in its own batch confounded batch with strain. The combined batch G was produced as part of this research. Batches A-E were produced in our prior work.

5.3 Count data statistics across samples

To verify that the data has the same underlying distribution in the parasite data, a density plot of library size-normalized gene counts for each parasite sample was generated as shown in Figure 5-3.

Figure 5-3B highlights different aspects of gene expression regulation in human and parasite genes. Human gene expression is frequently regulated at the transcriptional initiation level, with many genes turned off under different conditions leading to the spike at zero count. The *T. cruzi* parasite (like other trypanasomatids) transcribes the majority of its genes in polycistronic units (92) and fine-tunes steady-state mRNA levels post-transcriptionally. A narrow range of mRNA abundances is observed with a single peak.

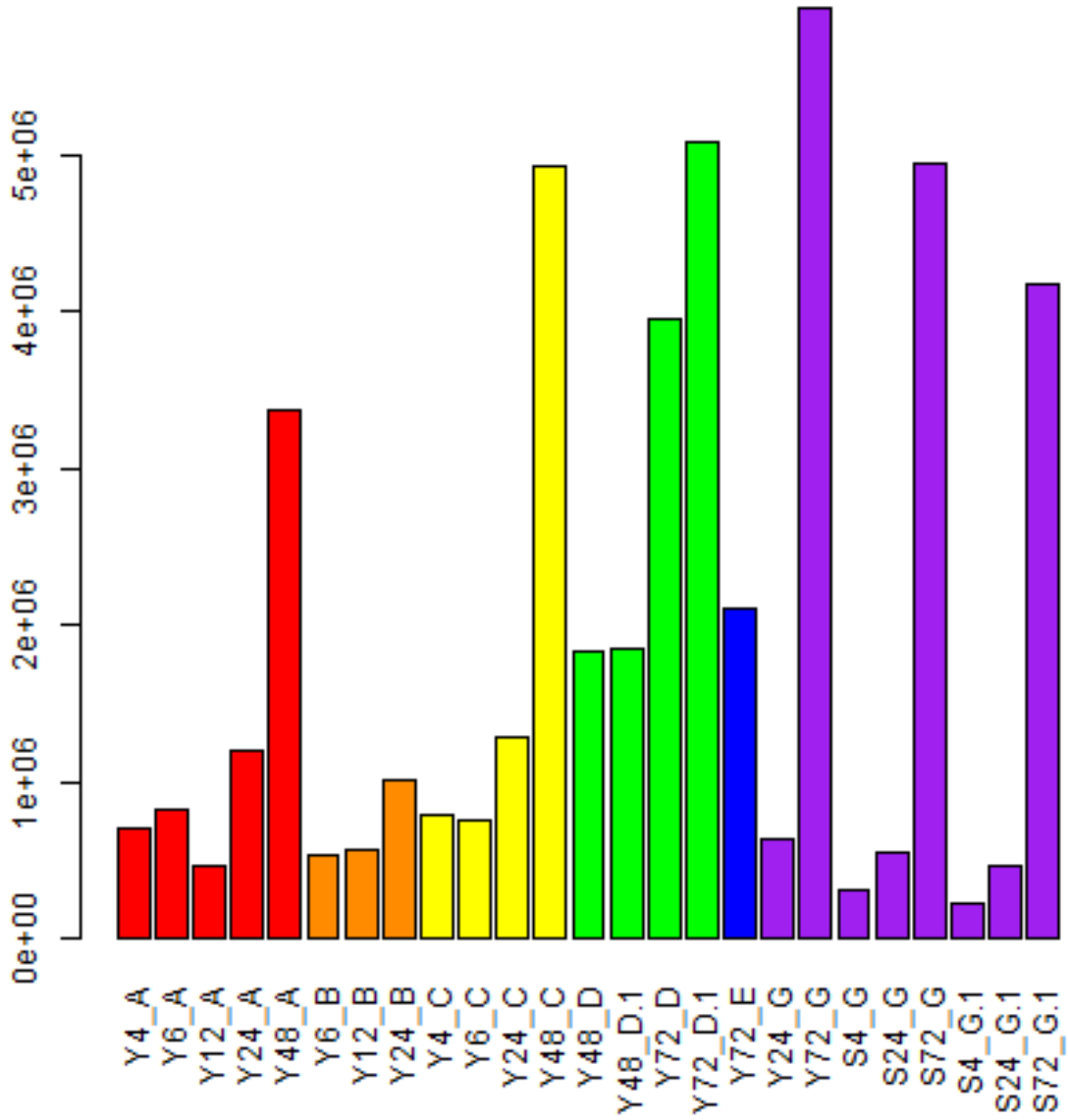

Figure 5-2. Barplot of raw library sizes for all parasite samples. Each bar represents the number of reads for a single sample in the combined dataset. Color represents a sequencing batch, generally from experiments performed starting on the same day.

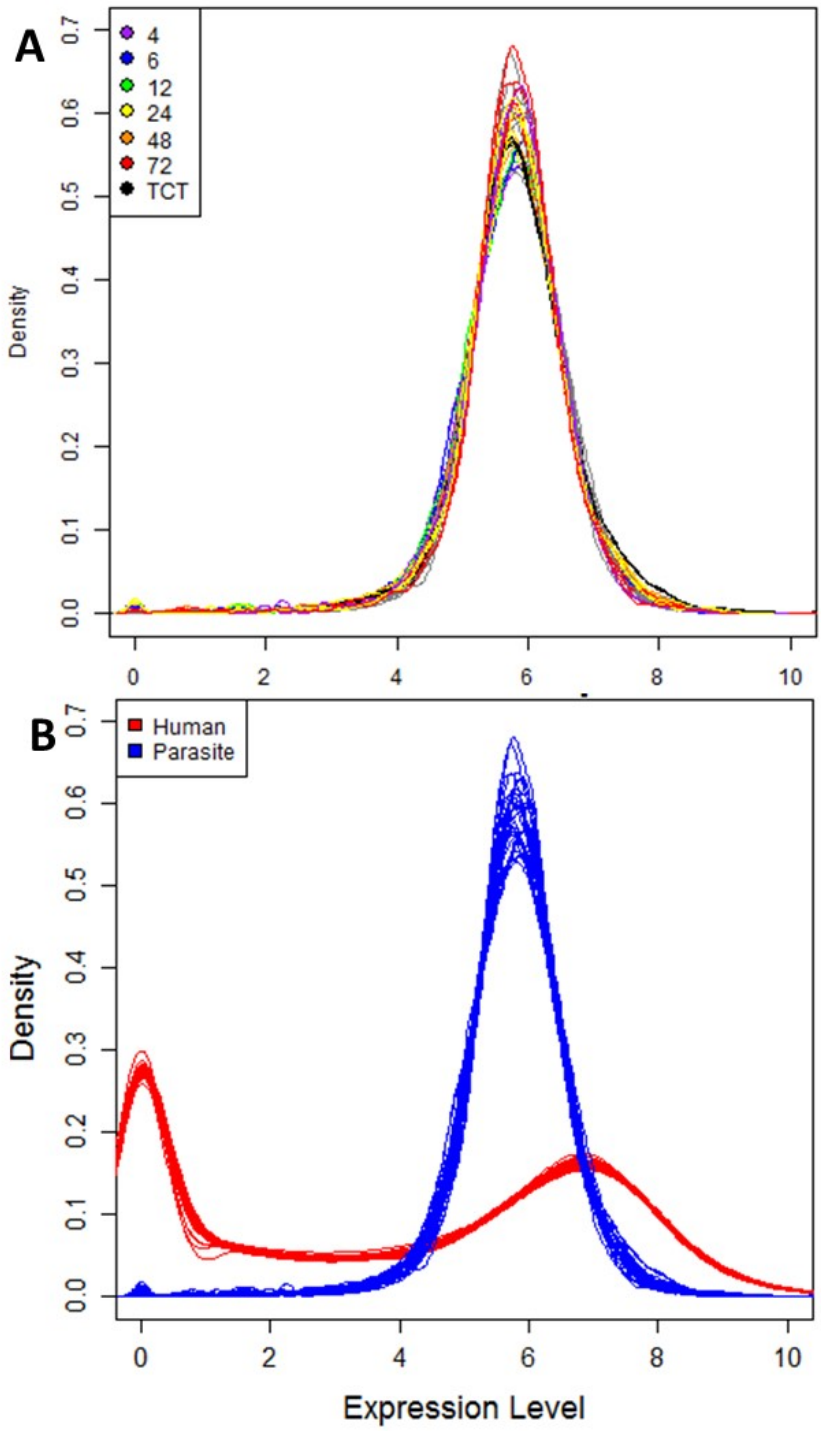

Figure 5-3. Density plot of all parasite samples and comparison with human density. A. Each curve represents an individual sample in the combined dataset. Ideally, all curves would be superimposed, and the data would show the same underlying distribution. Counts (expression level) are library size-normalized to address the differences in sequencing depth across batches, and $\log_2$ transformed. B. Parasite and Human data superimposed.

A different view of the same data, focusing on the mean count, is shown in Figure 5-4.

The library size normalized mean count did not differ substantially between samples. There was a difference between the Sylvio samples and the Y strain samples (including the ones produced in this study) at the high expression end (log_2 expression > 9). We evaluated the genes that were missing in those samples. The four genes that were very highly expressed in Y strain but not in Sylvio are shown in Table 5-1. The functions of these four genes did not indicate any functional or biochemical pattern, so no further analysis was performed.

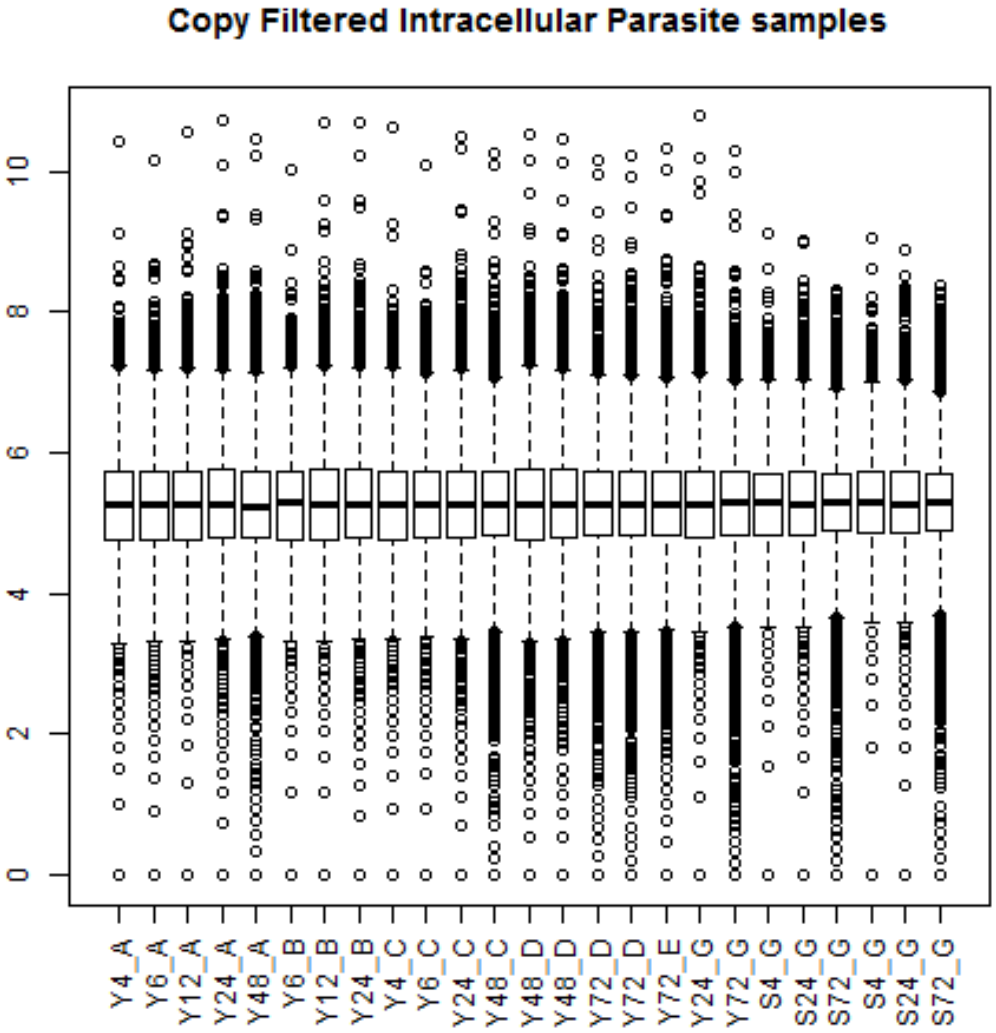

Figure 5-4. Boxplot of library size normalized counts by sample. Each box represents an individual sample in the combined dataset. Ideally, all boxes would have exactly the same mean (horizontal line) and range. Counts (expression level) are library size-normalized to address the differences in sequencing depth across batches, and $\log_2$ transformed.

Table 5-1. Genes highly expressed in Y strain but not Sylvio.

Y Gene ID	Sylvio Gene ID	Function	Avg Y	Avg S
TcCLB.506563.40	TCSYLVIO_007352	Beta tubulin putative	9.636244	3.44944
TcCLB.507713.30	TCSYLVIO_008621	Heat shock protein 85 putative	10.41224	8.726829
TcCLB.511211.160	TCSYLVIO_009847	glucose-regulated protein 78, putative	9.196091	5.267464
TcCLB.511211.170	TCSYLVIO_010042	Heat shock protein 70 (hsp70) putative	9.101139	7.417375

5.4 Identification of outlier samples

A Pearson correlation was computed for all samples against all samples in the combined parasite dataset. A heatmap showing the resulting correlation matrix is shown in Figure 5-5.

There was generally very good inter-sample correlation across the entire dataset for Y strain samples, including the two Y strain datapoints produced in this research. The dendrogram clearly separates Sylvio samples from Y strain samples. Within strain, separation was between early timepoints (4 hpi-24 hpi) and late timepoints (24 hpi-72 hpi).

A standard method for identifying outliers was applied to the data samples (67). The median pairwise correlation (MPC) for the first quartile across samples was computed (Q1), as was the inter-quartile range (IQR). Briefly, any sample whose MPC was less than Q1-1.5(IQR) was removed as an outlier. A scatterplot of median pairwise correlation for each sample is shown in Figure 5-6.

The Y strain parasite data showed good correlation between biological replicates. With the full set of all possible orthologs (Figure 5-6A), the Sylvio strain data showed considerably lower correlation with the Y strain data than when limited to the uniquely matching “core” orthologs (Figure 5-6B). No parasite samples were removed as outliers.

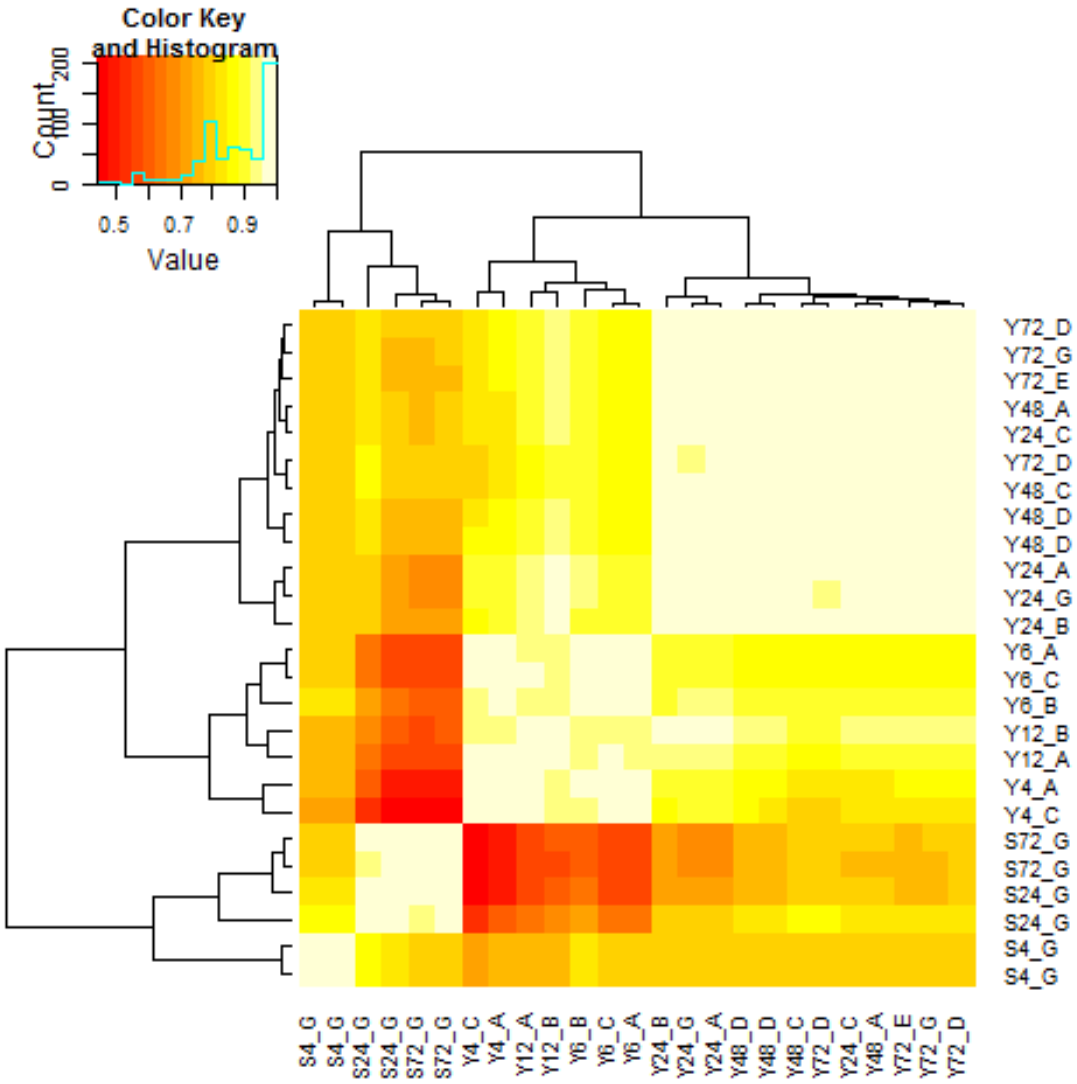

Figure 5-5. Pearson correlation heatmap. This heatmap visualizes the Pearson correlation matrix for each sample compared to every other sample in the combined dataset. Color at the intersection of each sample pair represents the Pearson correlation coefficient for that sample pair.

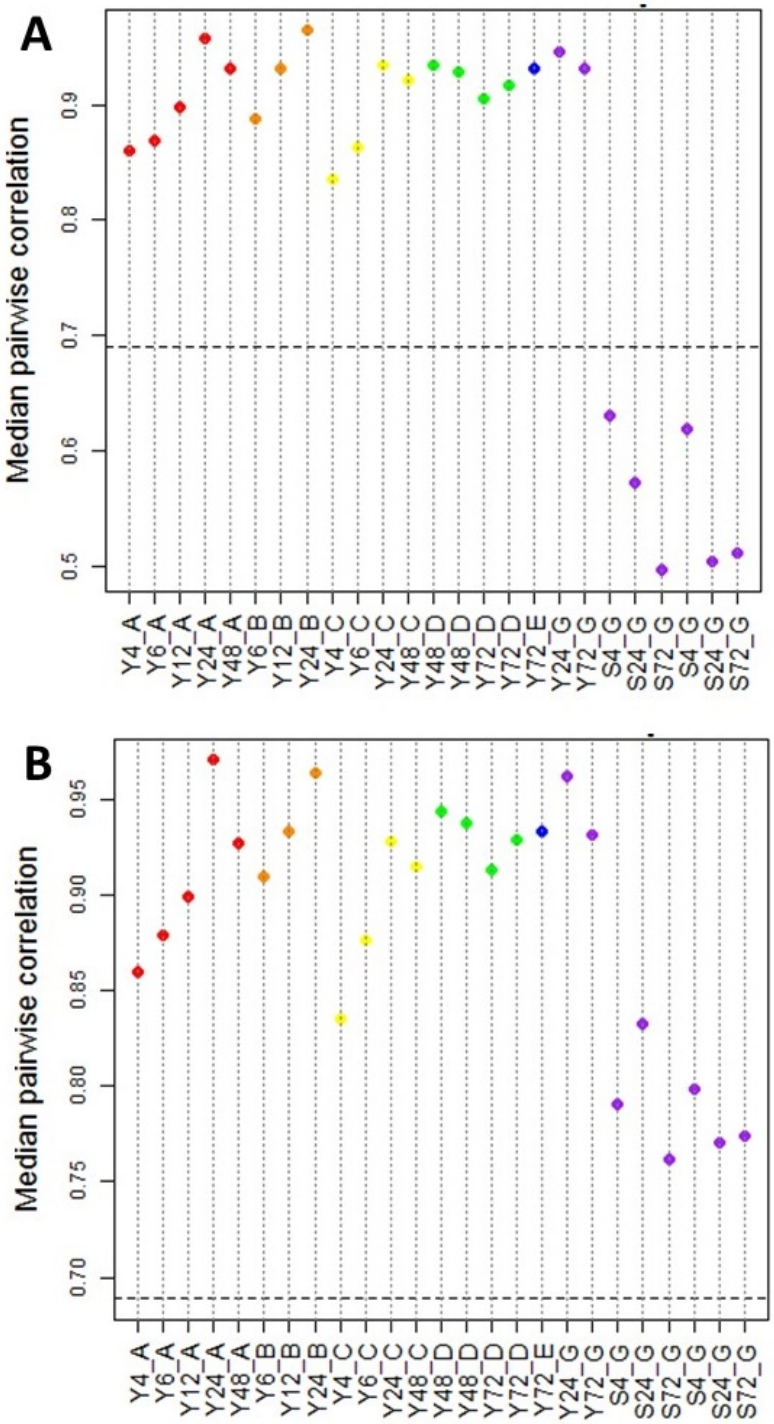

Figure 5-6. Scatterplots of median pairwise Pearson correlation. Each point represents the median pairwise correlation for that sample with all other samples in the dataset. The horizontal line is the interquartile range cutoff, computed as median pairwise correlation (MPC) for the first quartile across samples – 1.5 times the inter-quartile range (IQR). A. Median Pairwise Correlation of all genes with any ortholog. B. Median Pairwise limited to the "core" gene set of unique reciprocal best hits.

5.5 Principal component analysis

A principal component analysis was performed on quantile normalized per-gene counts from both strains incorporating data produced in this study with the Y strain data from our previous study. This allowed us to evaluate any potential batch effects introduced by the different labs and the different experimental timeframes.

Because transcription of genes in trypanosomatids is not turned off at the individual gene level, no parasite genes were removed from processing by the low count filter. Data were quantile normalized and $\log_2$ transformed (69).

As described for the human data, the singular value decomposition of the normalized and transformed combined dataset was computed. The resulting covariance matrix of the genes forms the basis of the principal component scores (86). To evaluate which genes had the biggest impact on principal component 1, which explains the largest amount (40%) of the variance, sorted loadings were plotted across all genes as shown in Figure 5-7.

Of 318 low-count filtered genes, 4653 (7%) had a loading greater than 0.025 (knee of the curve).

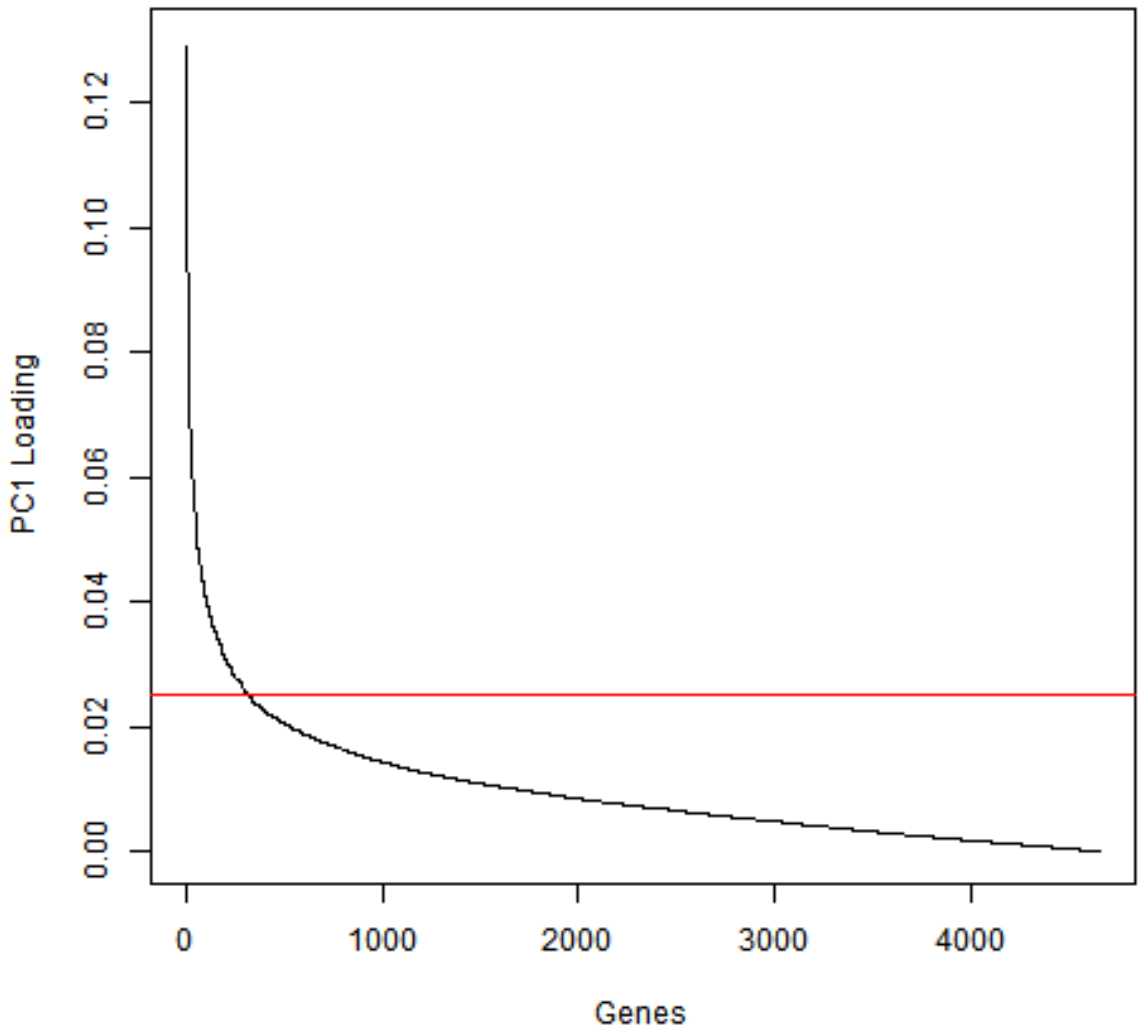

Figure 5-7. Loading analysis of uncorrected PC1. Sorted loadings were plotted across all genes. The red line identifies the value at the knee of the curve.

The variance matrix was then evaluated to determine the variance of each principal component and how that variance correlates with batch and experimental condition (strain combined with hpi). The resulting table is shown in Table 5-2.

We then plotted the Principal component scores for PC1 against PC2 for each sample as shown in Figure 5-8.

Table 5-2. Variance for each principal component without batch correction. For each principal component, the proportion of total variance in the dataset explained by that PC, and the contribution of condition (strain and timepoint) and batch for that PC are reported.

PC	Proportion Variance	Condition Contribution	Batch Contribution
1	39.52	99.61	66.6
2	29.49	98.86	22.57
3	4.98	93.56	7.48
4	4.21	97.04	22.25
5	3.92	17.63	79.52
6	2.39	50.89	13.74
7	2.17	43.09	24.8
8	2.05	7.2	4.32
9	1.42	15.81	29.14
10	1.34	57.48	12.19

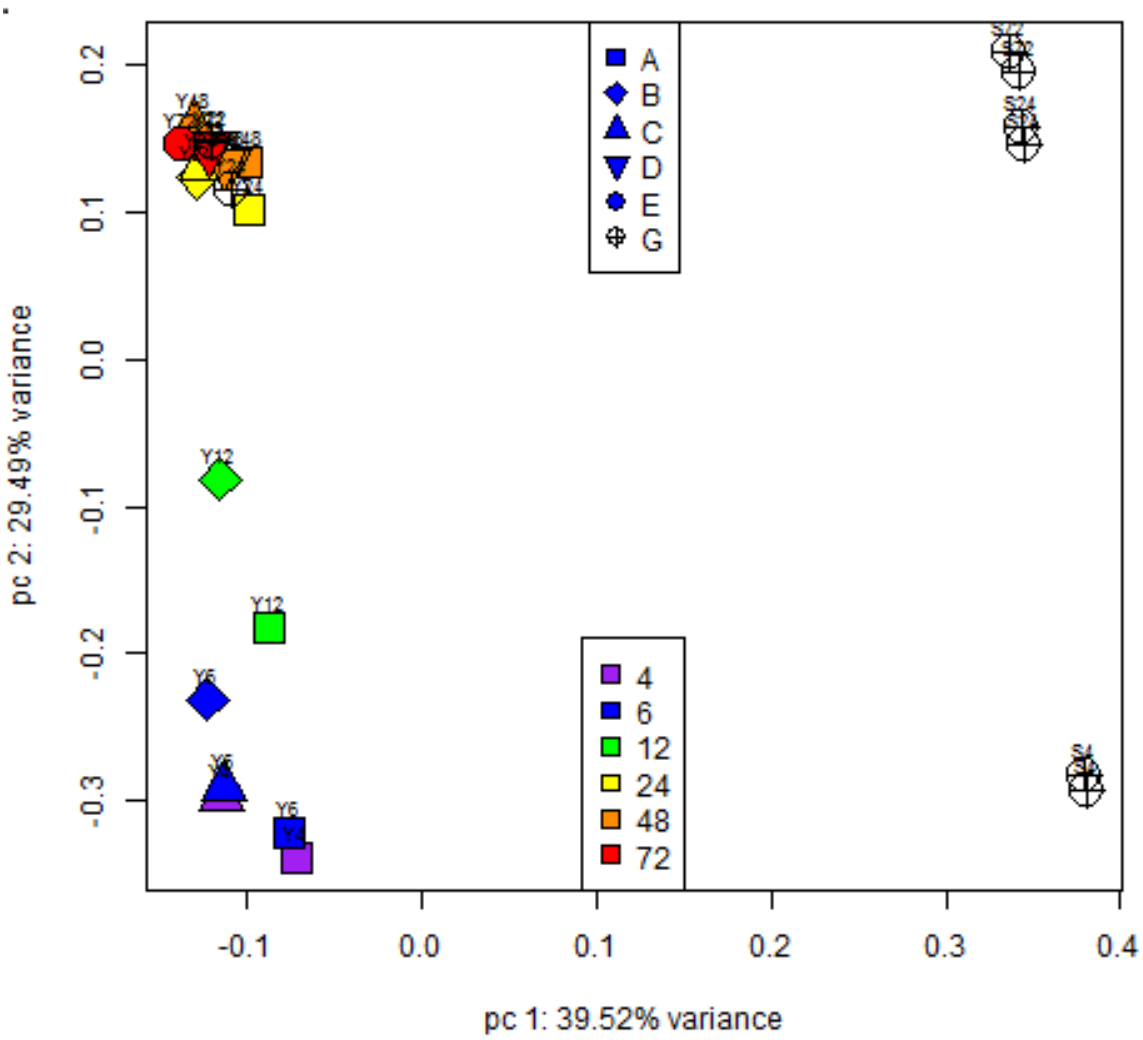

Figure 5-8. Uncorrected principal component analysis for parasite data. Principal component score for PC1 is plotted against the principal component score for PC2 for each sample. Color represents time hpi for infected samples. Shape corresponds to batch as shown in the key. Batch G was produced as part of this research, and batches A-E came from our previous work.

In the uncorrected PCA, principal component 1 is associated with strain, and principal component 2 is associated with time hpi, with early timepoints clustering at the bottom and later timepoints clustering at the top. The two Y strain datapoints produced in this work cluster nicely with their timepoint-matched datapoints from our previous work. This is evidence that batch effect is not a major contributor to variance in the parasite data.

To correct for the mean-variance relationship, we used the voom package (71) to weight each gene. The mean-variance trend computed by voom as part of data weighting is shown in Figure 5-9.

After the data was weighted and batch was included in the linear model, the principal component analysis was repeated. The variance matrix was recomputed, and a new singular value decomposition was produced to determine the new variance of each principal component for the corrected data. The resulting table is shown in Table 5-3.

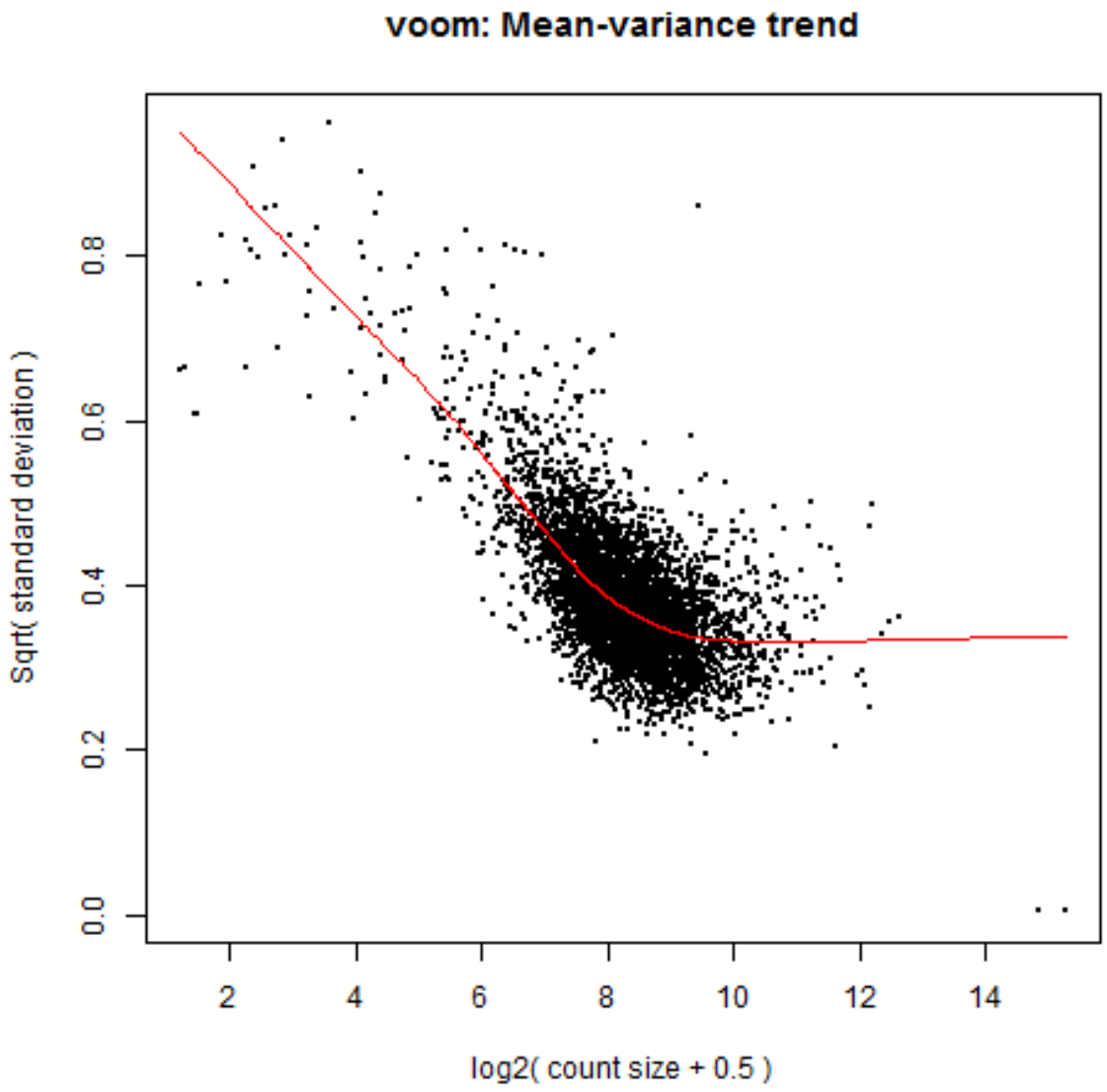

Figure 5-9. Mean-variance trend in uncorrected combined human data. The trendline is produced by the voom package as part of computing weights to correct for mean-variance relationship in count-based data.

Table 5-3. Variance for each principal component with batch included as a covariate in limma model. For each principal component, the proportion of total variance in the dataset explained by that PC, and the contribution of condition (control or infected and timepoint) for that PC are reported. The influence of batch has been removed by including it as a covariant in the model.

PC	Proportion Variance	Condition Contribution
1	49.44	99.96
2	26.99	99.29
3	8.25	99.19
4	5.3	97.72
5	1.85	28.79
6	1.63	94.69
7	1.16	8.57
8	0.97	41.28
9	0.81	33.72
10	0.67	77.26

We then plotted the new Principal component scores for PC1 against PC2 for each sample as shown in Figure 5-10.

The parasite data showed a clear separation along PC1 between *T. cruzi* Sylvio and Y strains, with the 4 hour timepoint for Sylvio strain relatively close to the corresponding Y strain datapoints and the distance increasing as the infection progressed. PC2 reflected the progression of time post-infection. The early timepoints (4-6 hpi) clustered together, when the parasite is still expected to be in the parasitophorous vacuole and beginning its transition from the infective trypomastigote into the intracellular replicative form. The datapoints from 12 hpi, when the parasite would have left the parasitophorous vacuole, are clustered between the early and late timepoints. Finally, the data from the later timepoints, 24 to 72 hpi, when the parasite would be dividing among other things, clustered together tightly within strain. Similarly to the human data, the Y strain datapoints produced in this study grouped with the corresponding Y strain timepoints from the existing data.

The separation of Y and Sylvio strain was more than expected considering that the parasite PCA analysis was limited to a core orthologous gene set defined in our study. This reduced dataset excluded many of the multi-gene family members which include many of the virulence factors associated with attachment and invasion and have generally been presumed to drive biological differences between strains (27).

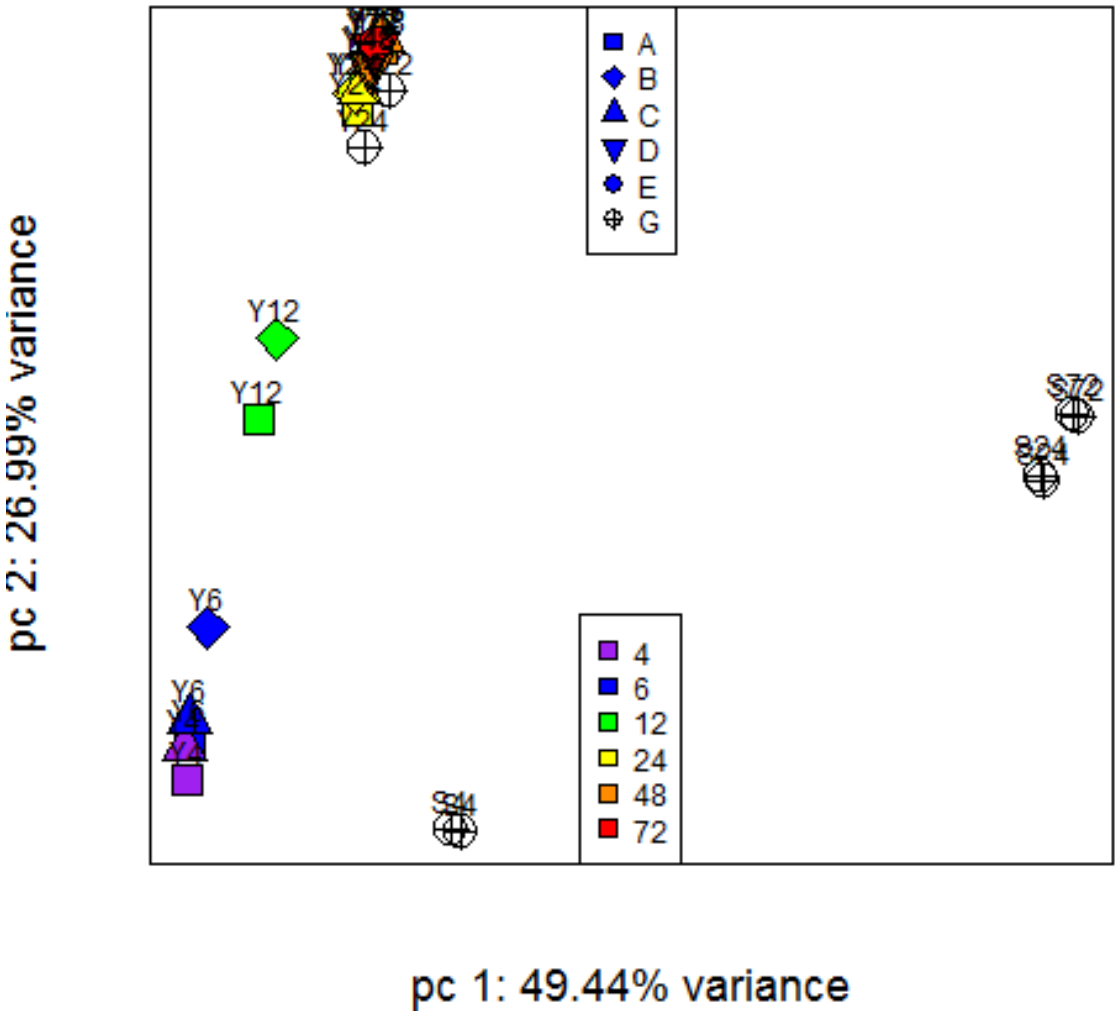

Figure 5-10. Principal component analysis, batch included as a covariate in limma model. Principal Component score for PC1 is plotted against the Principal component score for PC2 for each sample. Color represents time hpi for infected samples as shown in key. Shape corresponds to batch as shown in the key. Batch G was produced as part of this research, and batches A-E came from our previous work.

5.6 Differential expression analysis

We examined the parasite transcriptome during the infection of a human host cell by generating DE profiles for *T. cruzi* Sylvio and comparing them to the DE lists generated for *T. cruzi* Y infection in our previous study (62). All *T. cruzi* Sylvio-infected samples (no outliers) and the full set of genes in the Version 8.0 TriTrypDB Sylvio annotation were used for differential expression analysis. The analysis performed was identical to the one carried out on the human data as described in Chapter 4, with some additional limitations inherent to the parasite genome(s). While the human genome is finished and well annotated, the parasite reference genomes used in this study are not fully assembled because of their high repeat content and the multigene families. The CL Brener strain of *T. cruzi*, the original sequenced genome (46) represents the most 'finished' genome including chromosome-assigned scaffolds (93), whereas the Sylvio genome assembly consists of 19,000 contigs (50). In addition, a large proportion (~60%) of the gene products in *T. cruzi* are annotated as 'hypothetical' and the gene ontology labels and KEGG pathway assignments are sparse.

Similar to our observations in the host cell response to the two strains, a striking similarity was seen between the response of Sylvio and Y strains of *T. cruzi* when the significantly DE core genes were compared.

A summary of the number of Differentially Expressed genes for the parasite data is shown in Table 5-4.

A graphical representation of the similarities and differences between the two strains as the parasite transitions from the 4 to 24 hpi is shown in the form of a scatter plot (Figure 5-11).

There was a common set of 119 core genes that were significantly DE at greater than 2 fold. Of these, 88 were commonly upregulated (Figure 5-11, quadrant I), and 31 were commonly downregulated (Figure 5-11, quadrant III) in both strains. Between 24 and 72 hpi, only 4 shared genes were significantly DE at greater than 2 FC (Table 5-4). The substantially larger number of genes differentially expressed between 4 and 24 hpi than between 24 and 72 hpi are consistent with our current understanding of the biology of the parasite during infection and our previous observations in *T. cruzi* Y strain . After invading the host cell, *T. cruzi* begins its transformation from the invasive trypomastigote form into the intracellular, replicative amastigote form. Once it completes this transformation within the first 24 hpi, it continues dividing approximately every 12 hours until it bursts the host cell sometime between 96 and 120 hpi.

Table 5-4. Summary of differential expression analysis of parasite data. Summary of differentially expressed core genes with their direction (up- or downregulated) at various fold change cutoffs for early versus late timepoints. To be considered differentially expressed, a gene must have a q-value < 0.05.

Sylvio 24 vs 4 hpi	Sylvio DE genes	Y DE genes	Shared DE genes, no FC cutoff		Shared DE genes, -1< log2 FC<1 (pink)		Shared DE genes, |log2 Fold Change| >=1 (red)	
Significant DE Genes	3470	5403						
Core Significant DE Genes	1769	2747	1280		1159		121	
Both Upregulated (Q1)			540	42.2%	452	39.0%	88	72.7%
Y Up, Sylvio Down (Q2)			42	3.3%	41	3.5%	1	0.8%
Both Downregulated (Q3)			649	50.7%	618	53.3%	31	25.6%
Sylvio Up, Y Down (Q4)			49	3.8%	48	4.1%	1	0.8%
72 v 24								
Significant DE Genes	308	139						
Core Significant DE Genes	145	71	20		16		4	
Upregulated	63	32	8	40.0%	7	43.8%	1	25.0%
Downregulated	82	39	12	60.0%	9	56.3%	3	75.0%

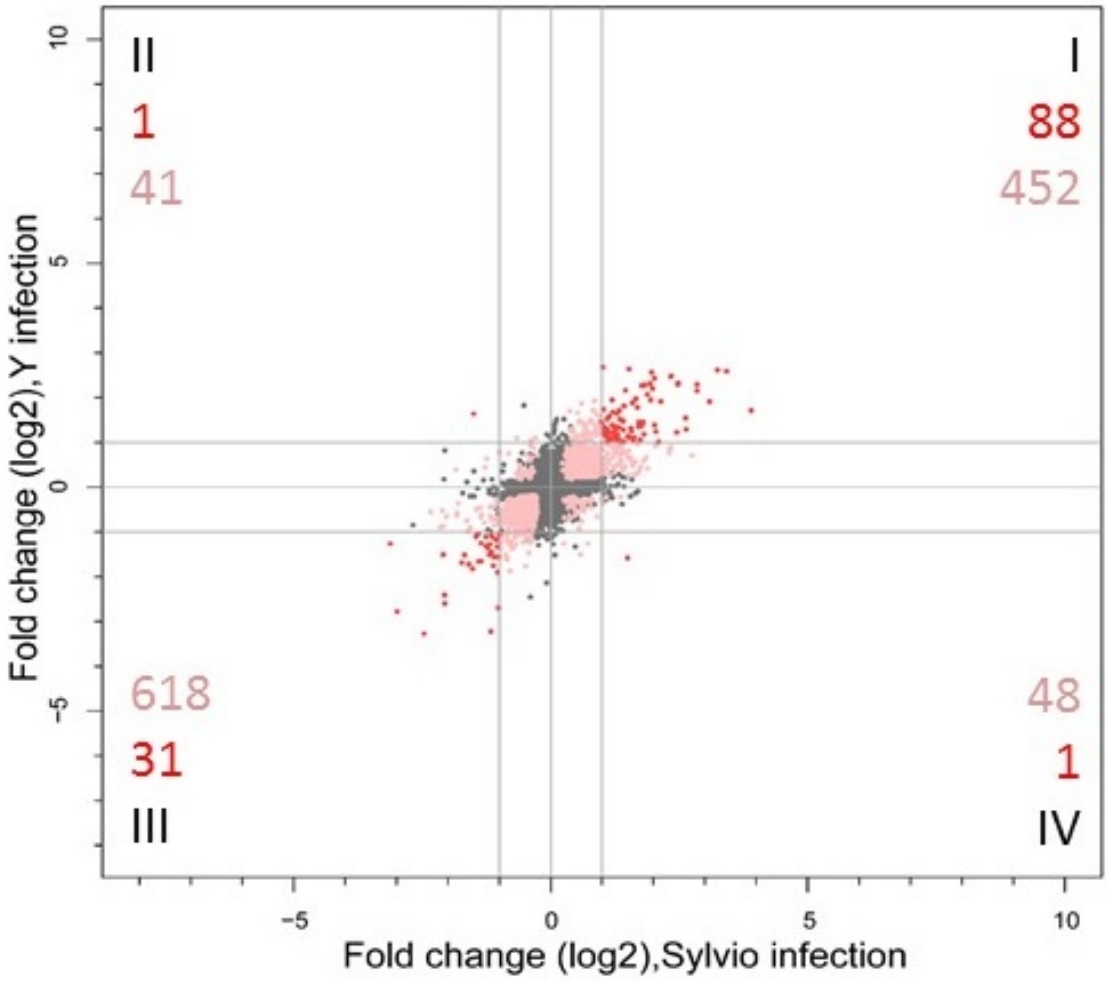

Figure 5-11. Comparative analysis of *T. cruzi* parasite Sylvio versus Y strain transcriptome response. Scatterplot of log2 fold change in expression of a subset of orthologous T. cruzi parasite genes in Sylvio (x-axis) versus Y strain (y-axis) at 24 hpi versus 4 hpi

5.7 Gene ontology analysis

The GO Term Finder (77) was applied to the parasite data using the server at go.princeton.edu. The GO terms for *T. cruzi* were extracted from the Version 9.0 TriTrypDB text file into a standard gene association file (GAF format) and provided in the advanced input options.

We examined the shared parasite signature for Gene Ontology enrichment in the subset of genes significantly DE between 4 and 24 hpi. Analysis of all co-upregulated genes revealed an enrichment of lipid and sterol biosynthesis pathways, consistent with our previous observations in *T. cruzi* Y strain (62). We also noted an enrichment of the one-carbon metabolism process (GO:0006730). This process produces the precursors for the production of trypanothione, a key metabolite responsible for addressing oxidative stress in trypanosomes, as trypanothione reductase mutants have been shown to be avirulent (94).

The one carbon/redox biochemical pathway has a large number of differentially expressed genes in the Sylvio and Y strain parasite. These genes are overlayed onto a pathway diagram in Figure 5-12. Taken together, the genes that are overexpressed generally appear to be increasing production of precursors to glutathione (GSH), and generally those that are underexpressed appear to be rescuing components of glutathione from catabolism. Glutathione is the precursor to trypanothione, a central molecule in the antioxidant defense system of the parasite (96). In fact, the drug currently in use to treat Chagas disease, Benznidazole, appears to act by covalently binding to trypanothione, thereby rendering it unavailable to the parasite. (97)

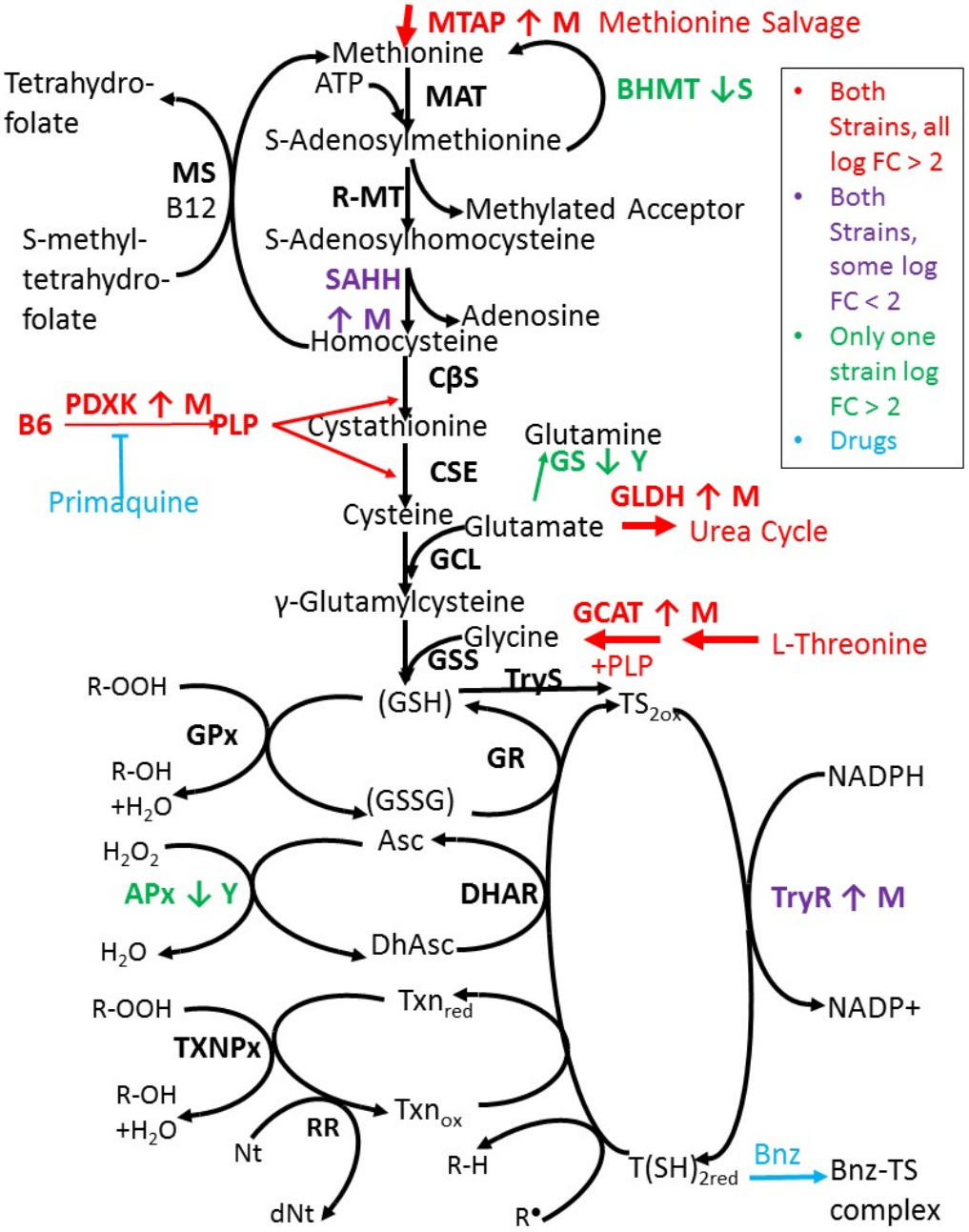

Figure 5-12. Differentially expressed genes in *T. cruzi* parasite and their role in redox chemistry. One carbon metabolism (95) leading into redox chemistry (96) is shown. Genes in red represent genes which are differentially expressed in both strains of parasite. Genes in purple met the fold-change criterion in one strain of parasite but not the other. Genes in green are differentially expressed in only one strain of parasite. Blue shows the current understanding of the actions of drugs that are used or studied in Chagas Disease. Bolded items are enzymes, otherwise, metabolites

Table 5-5. Enzymes involved in one carbon metabolism. The most common abbreviation, the enzyme name, the Enzyme Commission number, and whether this gene is differentially expressed in the parasite transcriptome.

Abbrevi-ation	Enzyme Name	Enzyme Commission Number	DE
APx	ascorbate peroxidase	EC 1.11.1.11	Yes
BHMT2	homocysteine S-methyltransferase	EC 2.1.1.10	Yes
CSE	γ-cystathionase	EC 4.4.1.1	No
CβS	cystathione β-synthase	EC 4.2.1.22	No
DHAR	dehydroascorbate reductase (aka glutathione dehydrogenase)	EC 1.8.5.1	No
GCAT	glycine C-acetyltransferase	EC 2.3.1.29	Yes
GCL	γ-glutamylcysteine synthetase	EC 6.3.2.2	No
GLDH	glutamate dehydrogenase	EC 1.4.1.2	Yes
GPx	Glutathione peroxidase	EC 1.11.1.9	No
GR	glutathione reductase	EC 1.8.1.7	No
GSS	GSH synthetase	EC 6.3.2.3	No
MAT	methionine adenosyltransferase	EC 2.5.1.6	No
MS	methionine synthase	EC 2.1.1.13	No
MTAP	methylthioadenosine phosphorylase	EC 2.4.2.28	Yes
PDXK	pyridoxal kinase	EC 2.7.1.35	Yes
R-MT	transmethylation reactions		
RR	Ribonucleotide Reductase	EC 1.17.4.1	No
SAHH	S-adenosylhomocysteine hydrolase	EC 3.3.1.1	Yes
TryR	trypanothione reductase	EC 1.8.1.12	Yes
TryS	trypanothione synthase	EC 6.3.1.9	No
TXNPx	tryparedoxin peroxidase	UniProt O61000	No

Table 5-6. Metabolites in one carbon metabolism.

Abbreviation	MetaboliteName
Asc	Ascorbate
dhAsc	Dehydroascorbate
dNt	deoxyribonucleotides
GSH	Glutathione (reduced)
GSSG	Glutathione disulfide (oxidized)
Nt	ribonucleotides
R-OOH	hydroperoxides
T(SH)2	Trypanothione disulfide (reduced)
TS	Trypanothione (oxidized)
Txn	tryparedoxin

Targeting the one carbon/trypanothione metabolic pathway appears to be a very active area of current drug research in Chagas Disease. Currently, there are active lines of research looking at: Primaquine, an antimalarial drug which inhibits peroxidal kinase (PDXK), thereby reducing availability of vitamin B6, a central enzymatic cofactor in many metabolic reactions (98, 99); Quinoxalone derivatives, which target trypanothione reductase (TryR) thereby reducing the pool of trypanothione (100); Use of Vitamin B12 (cofactor for methionine synthase (MS)), alone or in combination with trypanocidal drugs (101); and Tryparedoxin (Txn) as a drug target (102)

Significantly DE co-downregulated genes showed no significant GO enrichment. Only two significant DE genes meeting the fold change cutoff were unique to one strain or the other. One was a hypothetical protein (TcCLB.507143.60/TCSYLVIO_002455) upregulated ~3-fold in Y but downregulated ~2.8-fold in Sylvio. The other was an amino acid permease (TcCLB.511325.50/TCSYLVIO_010159) upregulated ~2.8-fold in Sylvio and downregulated ~3-fold in Y.

Chapter 6. Concluding Remarks/Future Prospects

Our analysis found a remarkable consistency between both parasite and host responses to infections with both strains of parasite. Our findings are consistent with previously published work, and also published reports looking at these individual pathway responses. This work, combined with our previous work provides a deep dive into the transcription behavior of both host and *T. cruzi* parasite, and a comparison in the host-pathogen interaction between infections with two different strains of *T. cruzi*. In addition, we have made the full set of host and parasite transcription data available in public repositories, unlike previous transcriptome studies which only reported on a subset of genes.Finally, we identified a potentially significant difference in host response regarding IL-8, which could potentially lead to an explanation for the difference in clinical outcomes in chronic Chagas Disease when infected with DTU TcI vs. DTU TcII strains.

The similarity in host response to the two strains of parasite could be explained by the fact that the *in vitro* system used in this study necessarily removes all influences of the immune system, which would be seen in an *in vivo* experiment, and this interaction may show a marked difference in either host or parasite response to infection. An *in vivo* experiment using PBMC or similar types of circulating cells in infected humans in the future would be beneficial to interrogate this interaction.

Another plausible explanation for the marked similarity is that differences may lie within multigene families which were necessarily omitted in this analysis due to the poor current state of assembly of these largely repetitive sequences. It is likely that many of the virulence factors (such as trans-sialidases and MASPs) lie within these multigene

families. Using technologies that produce longer sequence reads will lead to a better assembly and annotation, which will resolve differences in multigene families and improve similar analyses in the future. There is work currently ongoing in other labs to finish the Sylvio genome into chromosomes using these technologies.

The organs most affected in chronic Chagas disease are the heart and GI tract. This study uses human foreskin fibroblasts, not only to facilitate comparison with our previous work, but because these cells multiply in culture without being genetically modified. Fully differentiated non-cancer cardiomyocytes and GI smooth muscle cells do not multiply in culture, and are expensive to purchase and notoriously difficult to work with on the bench. Nevertheless, using more relevant cell lines might provide additional tissue-specific insights into the host-pathogen interaction of these parasite strains.

Finally, differences at the proteome level may provide better insights into possible translational regulatory mechanisms not interrogated in this work.

A unique benefit of using RNA-Seq in the context of this work is the ability to characterize the early behavior of both host and parasite simultaneously, even when parasite RNA levels are proportionally very small during the early timepoints of an infection. This study provides a unique look at the host-pathogen interaction between *T. cruzi* Sylvio strain and a human host, and the first in-depth comparative transcriptome analysis of two different strains of parasite and their host response.

References

1. WHO. Chagas disease in Latin America: an epidemiological update based on 2010 estimates. Wkly Epidemiol Rec. 2015;90(6):33–44.
2. Rassi A, Marin-Neto JA. Chagas disease. Lancet. 2010;375(9723):1388–402.
3. Longo DL, Bern C. Chagas' Disease. N Engl J Med [Internet]. 2015;373(5):456–66. Available from: http://www.nejm.org/doi/abs/10.1056/NEJMra1410150
4. Lee BY, Bacon KM, Bottazzi ME, Hotez PJ. Global economic burden of Chagas disease: a computational simulation model. Lancet Infect Dis. United States; 2013 Apr;13(4):342–8.
5. Coura JR, Dias JCP. Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. Mem Inst Oswaldo Cruz. 2009;104 Suppl (i):31–40.
6. Medical Pictures Info, Centers for Disease Control. Chagas Disease [Internet]. 2016 [cited 2016 Jan 1]. Available from: http://www.cdc.gov/parasites/chagas/
7. Tarleton RL, Reithinger R, Urbina JA, Kitron U, Gurtler RE. The challenges of Chagas disease - Grim outlook or glimmer of hope? PLoS Med. 2007;4(12):1852–7.
8. Villar JC, Marin-Neto JA, Ebrahim S, Yusuf S. Trypanocidal drugs for chronic asymptomatic *Trypanosoma cruzi* infection. Cochrane database Syst Rev. England; 2002;(1):CD003463.
9. Wilkinson SR, Kelly JM. The role of glutathione peroxidases in trypanosomatids. Biol Chem. Germany; 2003 Apr;384(4):517–25.
10. Prata A. Clinical and epidemiological aspects of Chagas disease. Lancet Infect Dis. United States; 2001 Sep;1(2):92–100.
11. Teixeira ARL, Nitz N, Guimaro MC, Gomes C, Santos-Buch CA. Chagas disease. Postgrad Med J. England; 2006 Dec;82(974):788–98.
12. Williams-Blangero S, VandeBerg JL, Blangero J, Corrêa-Oliveira R. Genetic Epidemiology of Chagas Disease. Adv Parasitol. England; 2011;75:147–67.
13. Tostes SJ, Bertulucci Rocha-Rodrigues D, de Araujo Pereira G, Virmondes RJ. Myocardiocyte apoptosis in heart failure in chronic Chagas' disease. Int J Cardiol. Ireland; 2005 Mar;99(2):233–7.
14. Petherick A. Campaigning for Chagas disease. Nature. England; 2010 Jun;465(7301):S21–2.
15. Dias J, Silveira A, Schofield C. The Impact of Chagas Disease Control in Latin America - A Review. Mem Inst Oswaldo Cruz Rio Janeiro. 2002;97(5):603–12.
16. Rassi A, Rassi A, Marcondes de Rezende J, Rassi Jr. A, Rassi A, Marcondes de RJ. American trypanosomiasis (Chagas disease). InfectDisClinNorth Am. 2012 Jun;26:275–91.

17. Morillo CA, Marin-Neto JA, Avezum A, Sosa-Estani S, Rassi AJ, Rosas F, et al. Randomized Trial of Benznidazole for Chronic Chagas' Cardiomyopathy. N Engl J Med [Internet]. 2015;373(14):1295–306. Available from: http://dx.doi.org/10.1056/NEJMoa1507574

18. Villalta F, Dobish MC, Nde PN, Kleshchenko YY, Hargrove TY, Johnson CA, et al. VNI Cures Acute and Chronic Experimental Chagas Disease. JInfectDis. Department of Biochemistry, School of Medicine, Vanderbilt University, Nashville, Tennessee; 2013 Aug;208(1537-6613 (Electronic)):504–11.

19. Fernandes MC, Andrews NW. Host cell invasion by *Trypanosoma cruzi*: A unique strategy that promotes persistence. FEMS Microbiol Rev. 2012;36(3):734–47.

20. Control C for D. Parasites - American Trypanosomiasis (also known as Chagas Disease) [Internet]. Centers for Disease Control Website. 2013. Available from: http://www.cdc.gov/parasites/chagas/index.html

21. Atwood J A, Weatherly DB, Minning T A, Bundy B, Cavola C, Opperdoes FR, et al. The *Trypanosoma cruzi* proteome. Science. 2005;309(5733):473–6.

22. cram.com. Trypanosomes Flashcards - Cram.com [Internet]. 2015 [cited 2015 Jun 1]. Available from: http://www.cram.com/flashcards/0113100930-mfm-ii-trypanosomes-kunigelis-962955

23. Ullu E, Tschudi C. Trans splicing in trypanosomes requires methylation of the 5' end of the spliced leader RNA. ProcNatlAcadSciUS A. Yale MacArthur Center for Molecular Parasitology, New Haven, CT; 1991 Nov 15;88(0027-8424 (Print)):10074–8.

24. Daniels J-P, Gull K, Wickstead B. Cell biology of the trypanosome genome. Microbiol Mol Biol Rev. 2010;74(4):552–69.

25. Kim KS, Teixeira SMR, Kirchhoff L V., Donelson JE. Transcription and editing of cytochrome oxidase II RNAs in *Trypanosoma cruzi*. J Biol Chem. 1994;269(2):1206–11.

26. Burleigh BA, Andrews NW. The mechanisms of *Trypanosoma cruzi* invasion of mammalian cells. Annu Rev Microbiol. 1995;49:175–200.

27. de Souza W, de Carvalho TM, Barrias ES. Review on *Trypanosoma cruzi*: Host Cell Interaction. IntJCell Biol. 2010;2010.

28. Butler CE, Tyler KM. Membrane traffic and synaptic cross-talk during host cell entry by *Trypanosoma cruzi*. Cell Microbiol. 2012 Sep;14(1462-5822 (Electronic)):1345–53.

29. Barrias ES, de Carvalho TM, de SW. *Trypanosoma cruzi*: Entry into Mammalian Host Cells and Parasitophorous Vacuole Formation. Front Immunol. 2013;4(1664-3224 (Electronic)):186.

30. Albertti LA, Macedo AM, Chiari E, Andrews NW, Andrade LO. Role of host lysosomal associated membrane protein (LAMP) in *Trypanosoma cruzi* invasion

and intracellular development. MicrobesInfect.; 2010 Sep;12(1769-714X (Electronic)):784–9.

31. Villalta F, Madison MN, Kleshchenko YY, Nde PN, Lima MF. Molecular analysis of early host cell infection by *Trypanosoma cruzi*. Front Biosci. 2008;13(1093-4715 (Electronic)):3714–34.

32. Fernandes AB, Neira I, Ferreira AT, Mortara RA. Cell invasion by *Trypanosoma cruzi* amastigotes of distinct infectivities: studies on signaling pathways. ParasitolRes. 2006 Dec;100(0932-0113 (Print)):59–68.

33. Ferreira ER, Bonfim-Melo A, Mortara RA, Bahia D. Trypanosoma cruzi extracellular amastigotes and host cell signaling: more pieces to the puzzle. Front Immunol.; 2012;3(1664-3224 (Electronic)):363.

34. Mortara RA, Andreoli WK, Fernandes MC, da Silva C V, Fernandes AB, L'Abbate C, et al. Host cell actin remodeling in response to *Trypanosoma cruzi*: trypomastigote versus amastigote entry. SubcellBiochem.; 2008;47(0306-0225 (Print)):101–9.

35. Ley V, Andrews NW, Robbins ES, Nussenzweig V. Amastigotes of *Trypanosoma cruzi* sustain an infective cycle in mammalian cells. JExpMed. 1988 Aug 1;168(0022-1007 (Print)):649–59.

36. Zingales B, Andrade SG, Briones MRS, Campbell D a., Chiari E, Fernandes O, et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: Second revision meeting recommends TcI to TcVI. Mem Inst Oswaldo Cruz. 2009;104(7):1051–4.

37. Cosentino RO, Agüero F. A simple strain typing assay for *Trypanosoma cruzi*: Discrimination of major evolutionary lineages from a single amplification product. PLoS Negl Trop Dis. 2012;6(7).

38. Zingales B, Miles M A., Campbell D A., Tibayrenc M, Macedo AM, Teixeira MMG, et al. The revised *Trypanosoma cruzi* subspecific nomenclature: Rationale, epidemiological relevance and research applications. Infect Genet Evol. 2012 Mar;12(2):240–53.

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

40. Camargos ERS, Franco DJ, Garcia CMMG, Dutra AP, Teixeira A.L. J, Chiari E, et al. Infection with different *Trypanosoma cruzi* populations in rats: myocarditis, cardiac sympathetic denervation, and involvement of digestive organs. AmJTropMed Hyg. 2000 May;62(0002-9637 (Print)):604–12.

41. Duz ALC, Vieira PMDA, Roatt BM, Aguiar-Soares RDO, Cardoso JMDO, Oliveira FCB De, et al. The TcI and TcII *Trypanosoma cruzi* experimental infections induce distinct immune responses and cardiac fibrosis in dogs. Mem Inst Oswaldo Cruz. 2014;109(8):1005–13.

42. Ragone PG, Pérez Brandán C, Padilla AM, Monje Rumi M, Lauthier JJ, Alberti D'Amato AM, et al. Biological behavior of different *Trypanosoma cruzi* isolates circulating in an endemic area for Chagas disease in the Gran Chaco region of Argentina. Acta Trop. Netherlands; 2012 Sep;123(3):196–201.

43. Higo H, Miura S, Horio M, Mimori T, Hamano S, Agatsuma T, et al. Genotypic variation among lineages of *Trypanosoma cruzi* and its geographic aspects. Parasitol Int. 2004;53(4):337–44.

44. Zingales B, Souto RP, Mangia RH, Lisboa C V., Campbell D a., Coura JR, et al. Molecular epidemiology of American trypanosomiasis in Brazil based on dimorphisms of rRNA and mini-exon gene sequences. Int J Parasitol. 1998;28(1):105–12.

45. Zingales B, Miles MA, Moraes CB, Luquetti A, Guhl F, Schijman AG, et al. Drug discovery for chagas disease should consider *Trypanosoma cruzi* strain diversity. Mem Inst Oswaldo Cruz. 2014;109(6):828–33.

46. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, et al. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science. 2005;309(5733):409–15.

47. Zingales B, Pereira MES, Almeida KA, Umezawa ES, Nehme NS, Oliveira RP, et al. Biological Parameters and Molecular Markers of Clone CL Brener - The Reference Organism of the *Trypanosoma cruzi* Genome Project. Mem Inst Oswaldo Cruz [Internet]. 1997;92(6):811–4. Available from: http://www.bioline.org.br/request?oc97152

48. De Pablos LM, Osuna A. Multigene families in *Trypanosoma cruzi* and their role in infectivity. InfectImmun. 2012 Jul;80(1098-5522 (Electronic)):2258–64.

49. Weatherly DB, Boehlke C, Tarleton RL. Chromosome level assembly of the hybrid *Trypanosoma cruzi* genome. BMC Genomics. 2009;10:255.

50. Franzén O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles M A., et al. Shotgun sequencing analysis of *Trypanosoma cruzi* Sylvio X10/1 and comparison with T. cruzi VI CL Brener. PLoS Negl Trop Dis. 2011;5(3):1–9.

51. Grisard EC, Teixeira SMR, de Almeida LGP, Stoco PH, Gerber AL, Talavera-López C, et al. *Trypanosoma cruzi* Clone Dm28c Draft Genome Sequence. Genome Announc [Internet]. 2014;2(1):2–3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3907723&tool=pmcentrez&rendertype=abstract

52. Franzén O, Talavera-López C, Ochaya S, Butler CE, Messenger L A, Lewis MD, et al. Comparative genomic analysis of human infective *Trypanosoma cruzi* lineages with the bat-restricted subspecies T. cruzi marinkellei. BMC Genomics. 2012;13:531.

53. Shigihara T, Hashimoto M, Shindo N, Aoki T. Transcriptome profile of *Trypanosoma cruzi*-infected cells: simultaneous up- and down-regulation of

proliferation inhibitors and promoters. ParasitolRes. 2008 Mar;102(0932-0113 (Print)):715–22.

54. Costales J A, Daily JP, Burleigh B A. Cytokine-dependent and-independent gene expression changes and cell cycle block revealed in *Trypanosoma cruzi*-infected host cells by comparative mRNA profiling. BMC Genomics. 2009;10:252.

55. Adesse D, Garzoni LR, Meirelles MDN, Iacobas D A., Iacobas S, Spray DC, et al. Transcriptomic signatures of alterations in a myoblast cell line infected with four distinct strains of *Trypanosoma cruzi*. Am J Trop Med Hyg. 2010;82(5):846–54.

56. Goldenberg RCDS, Iacobas D A., Iacobas S, Rocha LL, da Silva de Azevedo Fortes F, Vairo L, et al. Transcriptomic alterations in *Trypanosoma cruzi*-infected cardiac myocytes. MicrobesInfect [Internet]. 2009 Dec;11(1769-714X (Electronic)):1140–9. Available from: http://dx.doi.org/10.1016/j.micinf.2009.08.009

57. Udoko AN, Johnson CA, Dykan A, Rachakonda G, Villalta F, Mandape SN, et al. Early Regulation of Profibrotic Genes in Primary Human Cardiac Myocytes by *Trypanosoma cruzi*. PLoS Negl Trop Dis [Internet]. 2016;10(1):1–23. Available from: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L607985752

58. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog. 2010;6(9):1–15.

59. Minning T A, Weatherly DB, Atwood J, Orlando R, Tarleton RL. The steady-state transcriptome of the four major life-cycle stages of *Trypanosoma cruzi*. BMC Genomics. 2009;10:370.

60. Diaz ML, Solari A., Gonzalez CI, Díaz ML, Solari A., González CI. Differential expression of *Trypanosoma cruzi* I associated with clinical forms of Chagas disease: overexpression of oxidative stress proteins in acute patient isolate. JProteomics [Internet]. 2011 Aug 24;74(1876-7737 (Electronic)):1673–82. Available from: http://dx.doi.org/10.1016/j.jprot.2011.05.001

61. Duran-Rehbein GA, Vargas-Zambrano JC, Cuéllar A, Puerta CJ, Gonzalez JM. Mammalian cellular culture models of Trypanosoma cruzi infection: a review of the published literature. Parasite [Internet]. 2014;21(c):38. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4118624&tool=pmcentrez&rendertype=abstract

62. Li Y, Shah-Simpson S, Okrah K, Belew AT, Choi K, Caradonna KL, et al. Transcriptome remodeling in *Trypanosoma cruzi* and human cells during intracellular infection. PLoS Pathog. 2016;In Press.

63. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

64. Anders S. Babraham Bioinformatics FastQC Page [Internet]. 2015 [cited 2013 Nov 10]. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

65. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. BioMed Central Ltd; 2013;14(4):R36.

66. Anders S, Pyl PT, Huber W. HTSeq A Python framework to work with high-throughput sequencing data. bioRxiv. 2014;31(2):002824.

67. Hoaglin DC, Mosteller F, Tukey JW. Understanding robust and exploratory data analysis. New York: Wiley; 1983. 447 p.

68. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765–86.

69. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11(12):220.

70. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94.

71. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.

72. ’t Hoen P A C, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol [Internet]. 2013;31(11):1015–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24037425

73. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7).

74. Rapaport F, Khanin R, Liang Y, Krek A, Zumbo P, Mason CE, et al. Comprehensive evaluation of differential expression analysis methods for RNA-seq data. arXiv. 2013;1–21.

75. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5.

76. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991;11(3):635–50.

77. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710–5.

78. Reimand J, Arak T, Vilo J. G:Profiler - A web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011;39(SUPPL. 2):1–9.

79. Cruz L, Vivas A, Montilla M, Hernández C, Flórez C, Parra E, et al. Comparative study of the biological properties of *Trypanosoma cruzi* I genotypes in a murine experimental model. Infect Genet Evol [Internet]. Elsevier B.V.; 2015;29:110–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1567134814004158

80. Poveda C, Fresno M, Gironès N, Martins-Filho O a., Ramírez JD, Santi-Rocca J, et al. Cytokine profiling in chagas disease: Towards understanding the association with infecting *Trypanosoma cruzi* discrete typing units (A benefit trial sub-study). PLoS One. United States; 2014;9(3):1–8.

81. Burleigh BA, Woolsey AM. Microreview Cell signalling and *Trypanosoma cruzi* invasion. Cell Microbiol. 2002;4(11):701–11.

82. Love M, Anders S, Huber W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. 2013;1–32.

83. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 2006;7:29–59.

84. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat [Internet]. John Wiley & Sons, Inc.; 2010;2(4):433–59. Available from: http://dx.doi.org/10.1002/wics.101

85. Leek JT, Scharpf R, Bravo HC, Simcha D, Langmead B, Johnson W, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.

86. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A. 2003;100(6):3351–6.

87. Smyth G. limma: Linear Models for Microarray Data. Bioinforma Comput Biol Solut Using R Bioconductor [Internet]. 2005;(2005):397–420. Available from: http://dx.doi.org/10.1007/0-387-29362-0_23

88. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. Ann Stat. 2003;31(6):2013–35.

89. Mccarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25(6):765–71.

90. Chen CM, Shih TH, Pai TW, Liu ZL, Chang MDT. RNA-seq coverage effects on biological pathways and GO tag clouds. 2012 IEEE 6th Int Conf Syst Biol ISB 2012. 2012;(Cc):240–5.

91. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.

92. Andersson B, ÅSlund L, Tammi M, Tran A-N, Hoheisel JD, Pettersson U.

Complete sequence of an 93.4 kb contig from chromosome 3 of *Trypanosoma cruzi* containing a strand switch region. Genome Res. 1998;8:809–16.

93. Souza RT, Lima FM, Barros RM, Cortez DR, Santos MF, Cordero EM, et al. Genome size, karyotype polymorphism and chromosomal evolution in *Trypanosoma cruzi*. PLoS One. 2011;6(8).

94. Krieger S, Schwarz W, Ariyanayagam MR, Fairlamb AH, Krauth-Siegel RL, Clayton C. Trypanosomes lacking trypanothione reductase are avirulent and show increased sensitivity to oxidative stress. Mol Microbiol. 2000 Feb;35(3):542–52.

95. Lu SC. Regulation of hepatic glutathione synthesis: current concepts and controversies. FASEB J [Internet]. 1999;13(10):1169–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10385608

96. Irigoín F, Cibils L, Comini M A., Wilkinson SR, Flohé L, Radi R. Insights into the redox biology of *Trypanosoma cruzi*: Trypanothione metabolism and oxidant detoxification. Free Radic Biol Med. 2008;45(6):733–42.

97. Trochine A, Creek DJ, Faral-Tello P, Barrett MP, Robello C. Benznidazole Biotransformation and Multiple Targets in *Trypanosoma cruzi* Revealed by Metabolomics. PLoS Negl Trop Dis. 2014;8(5).

98. Kimura T, Shirakawa R, Yaoita N, Hayashi T, Nagano K, Horiuchi H. The antimalarial drugs chloroquine and primaquine inhibit pyridoxal kinase, an essential enzyme for vitamin B6 production. FEBS Lett [Internet]. Federation of European Biochemical Societies; 2014;588(20):3673–6. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0014579314006127

99. Jones DC, Alphey MS, Wyllie S, Fairlamb AH. Chemical, genetic and structural assessment of pyridoxal kinase as a drug target in the African trypanosome. Mol Microbiol. England; 2012 Oct;86(1):51–64.

100. Villalobos-Rocha JC, Sánchez-Torres L, Nogueda-Torres B, Segura-Cabrera A, García-Pérez C a., Bocanegra-García V, et al. Anti-*Trypanosoma cruzi* and anti-leishmanial activity by quinoxaline-7-carboxylate 1,4-di-N-oxide derivatives. Parasitol Res. Germany; 2014 Jun;113(6):2027–35.

101. Ciccarelli AB, Frank FM, Puente V, Malchiodi EL, Batlle A, Lombardo ME. Antiparasitic effect of vitamin B12 on *Trypanosoma cruzi*. Antimicrob Agents Chemother. United States; 2012 Oct;56(10):5315–20.

102. González-Chávez Z, Olin-Sandoval V, Rodíguez-Zavala JS, Moreno-Sánchez R, Saavedra E, Gonzalez-Chavez Z, et al. Metabolic control analysis of the *Trypanosoma cruzi* peroxide detoxification pathway identifies tryparedoxin as a suitable drug target. Biochim Biophys Acta - Gen Subj [Internet]. Netherlands: Elsevier B.V.; 2015 Feb;1850(2):263–73. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0304416514003651