ABSTRACT

Title of Thesis: DEVELOPMENT AND OPTIMIZATION OF TOOLS FOR CO-EXPRESSION NETWORK ANALYSES OF HOST-PATHOGEN SYSTEMS

Vincent Keith Hughitt, Doctor of Philosophy, 2018

Thesis directed by: Dr. Najib M. El-Sayed, Ph.D.

Department of Cell Biology and Molecular Genetics

High-throughput transcriptomics has provided a powerful new approach for studying host-pathogen interactions. While popular techniques such as differential expression and gene set enrichment analysis can yield informative results, they do not always make full use of information available in multi-condition experiments. Co-expression networks provide a novel way of analyzing these datasets which can lead to new discoveries that are not readily detectable using the more popular approaches.

While significant work has been done in recent years on the construction of co-expression networks, less is known about how to measure the quality of such networks. Here, I describe an approach for evaluating the quality of a co-expression network, based

on enrichment of biological function across the network. The approach is used to measure the influence of various data transformations and algorithmic parameters on the resulting network quality, leading to several unexpected findings regarding commonly-used techniques, as well as to the development of a novel similarity metric used to assess the degree of co-expression between two genes. Next, I describe a simple approach for aggregating information across multiple network parameterizations, in order to arrive at a robust “consensus” co-expression network. This approach is used to generate independent host and parasite networks for two host-trypanosomatid transcriptomics datasets, resulting in the detection of both previously known disease pathways and novel gene networks potentially related to infection. Finally, a differential network analysis approach is developed and used to explore the impact of infection on the host co-expression network, and to elucidate shared transcriptional signatures of infection by different intracellular pathogens.

The approaches developed in this work provide a powerful set of tools and techniques for the rigorous generation and evaluation of co-expression networks, and have significant implications for co-expression network-based research. The application of these approaches to several host-pathogen systems demonstrates their utility for host-pathogen transcriptomics research, and has resulted in the creation of a number of valuable resources for understanding systems-levels processes that occur during the process of infection.

DEVELOPMENT AND OPTIMIZATION OF TOOLS FOR CO-EXPRESSION NETWORK ANALYSES OF HOST-PATHOGEN SYSTEMS

by

Vincent Keith Hughitt

Thesis 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
2018

Advisory Committee:
Professor Najib M. El-Sayed, Chair
Professor Volker Briken
Professor Héctor Corrada Bravo
Professor David M. Mosser
Professor Karen Carleton, Dean's Representative

Dedication

I dedicate my degree to my grandparents, Phyllis and Jerry, who always inspire me to be a better person than I am, and without whose support I could never have made it this far, and to Ajax, for getting in the way of work just enough to keep me sane sometimes 😸.

Acknowledgements

I would like to thank mentor, Najib El-Sayed, for providing me with amazing educational, career, and life experience these past five years. You challenged me when I needed to be challenged, criticized me when I needed to be criticized, and probably praised me a fair bit more than I deserve. I respect your honesty, transparency, and your unwavering insistence on scientific rigor and precision in all research endeavors. I've learned a lot from working with you and seeing how you think about problems, and while I'm still far from fluent, I think I've even made some progress in deciphering the ancient handwriting script that you have been using over the years to try and communicate with me. Thank you for an amazing graduate experience.

I would like to thank my committee members: Héctor Corrada Bravo, Volker Briken, Karen Carleton, and David Mosser. Your expertise, keen eye for weaknesses, and insightful suggestions benefitted nearly all aspects of research efforts over the past years. I've learned from a lot from each of you, and am immensely appreciative of you for taking the time to help guide my research.

I would also like to thank all of my labmates, past and present. It's been a lot of fun work with you and learning from over these past years. I would especially like to thank Trey Belew for your unfailing kindness, and for your willingness to stop what you are doing and help me with whatever problem I may be having - no matter how big or small. I owe you at least two beers.

I would like to thank my undergraduate advisor, Stephen Freeland, whose introductory ecology and evolution and bioinformatics seminar courses are entirely

responsible for igniting my love of science, and for setting me on the path I'm currently on. And I would like to thank my former mentors and at Goddard, Jack Ireland and Steven Christe, whose guidance, friendship, and support ruled out any possible escape from science.

I would like to thank my family for all of their support over these years, especially my grandmother who thought I would be in school forever (it's not too to apply to med school!) I would like to thank my mom for too many reasons to list here, and I would like to thank my niece, Amber, whose wisdom, tenacity, and work ethic never ceases to amaze me - you are going to make an amazing teacher.

I would like to thank the BISI and CBCB communities. It's been a pleasure learning with all of you over these years. Thank you Chuck, Michelle, and Gwen for helping me navigate the sometimes murky waters of graduate life, and for always supporting me in whatever extracurricular endeavors I proposed.

Finally, I would like to thank all of the incredible friends I have in my life, who make it all worthwhile. Thank you Ariane for all of your support over the years - it couldn't have happened without you. Thank you to my beautiful and loving roommates, Anita and Samson -- Ajax and I could not have asked for better friends and housemates. And thank you to you all the amazing friends I've made over the past year - Marko, Soraya, Jeff and John - you all inspire me and keep me sane in ways you can't even imagine, and I feel extremely lucky to have met each of you.

Table of Contents

List of Abbreviations

A	adenine
dj. *P*-value	adjusted *P*-value
bicor	biweight midcorrelation
C	cytosine
CDS	coding sequence
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
cor	Pearson correlation
CPDB	ConsensusPathDB
CPM	counts-per-million
DE	differentially expressed
DEA	differential expression analysis
DiNA	differential network analysis
G	guanine
GLM	generalized linear model
GO	Gene Ontology
GPI	Glycosylphosphatidylinositol
GRN	gene regulatory network
HFF	human foreskin fibroblasts
hpi	hours post-infection
HsAll	*H. sapiens* all samples
HsLb	*H. sapiens* infected with *L. braziliensis*
HsLm	*Homo sapiens* infected with *Leishmania major*
HsLmUI	*Homo sapiens* infected with *Leishmania major* (uninfected samples)
HsTc	*H. sapiens* infected with *T. cruzi*
HsTcUI	*H. sapiens* infected with *T. cruzi* (Uninfected samples)
IFN	Interferon
KEGG	Kyoto Encyclopedia of Genes and Genomes
LmAll	*L. major* all samples
LmHs	*L. major* infecting *H. sapiens*
LmMm	*L. major* infecting *M. musculus*
logFC	log2 fold change
MmLm	*M. musculus* infected with *L. major*
nt	nucleotides
ORF	open reading frame

P/S	primary site over the secondary site
PC	principle component
PCA	principle component analysis
Pol II	polymerase II
PolyPy	polypyrimidine
qRT-PCR	Real-Time Quantitative Reverse Transcription PCR
SL	spliced leader
SRA	Sequence Read Archive
TcAll	*T. cruzi* all samples
TcHs	*T. cruzi* infecting *H. sapiens*
TF	transcription factor
TOM	topological overlap matrix
TriTryp	*T. brucei*, *T. cruzi*, and *L. major*
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
UTR	untranslated region
WGCNA	Weighted Gene Co-expression Network Analysis

Chapter 1

Introduction

A number of serious parasitic diseases, including sleeping sickness, Chagas disease, and Leishmaniasis, are caused by unicellular eukaryotic parasites of the order *Trypanosomatida*. Sleeping sickness is a severe disease which results from infection by the *Trypanosoma brucei*, an extracellular parasite which is able to successfully evade the immune system and survive in the bloodstream. Eventually, the parasite makes its way to the brain, resulting in the deep lethargy that gives this disease its name (Legros et al. 2002; WHO 2017). *T. brucei* parasites are transmitted to individuals through the bite of an infected Tsetse fly. If left untreated, the disease is nearly always fatal. Sleeping sickness is found wherever its insect vector lives which includes 36 countries in sub-Saharan Africa (WHO 2017). Despite large decreases in incidence over the past fifty years resulting from aggressive control strategies, thousands of individuals are still infected every year, and rates of infection have started going back up in several countries (Stich, Barrett, and Krishna 2003; Bilbe 2015).

Chagas disease is found throughout South and Central America with an estimated 8-9 million cases in 2002 (Committee 2002). The causative agent of Chagas disease, *Trypanosoma cruzi*, is transmitted to humans and other mammals by insects in the family *Reduviidae*, and is capable of infecting a range of host cells including macrophages, muscle cells, and fibroblasts (Andrade and Andrews 2005; de Souza, de Carvalho, and Barrias 2010) (**Figure 1**). The disease progresses through two stages: an acute stage and a

chronic stage. While the acute stage is generally short-lived, and may be asymptomatic, about 10-30% of individuals suffer severe and often fatal heart, colon, or esophagus complications many years after infection.

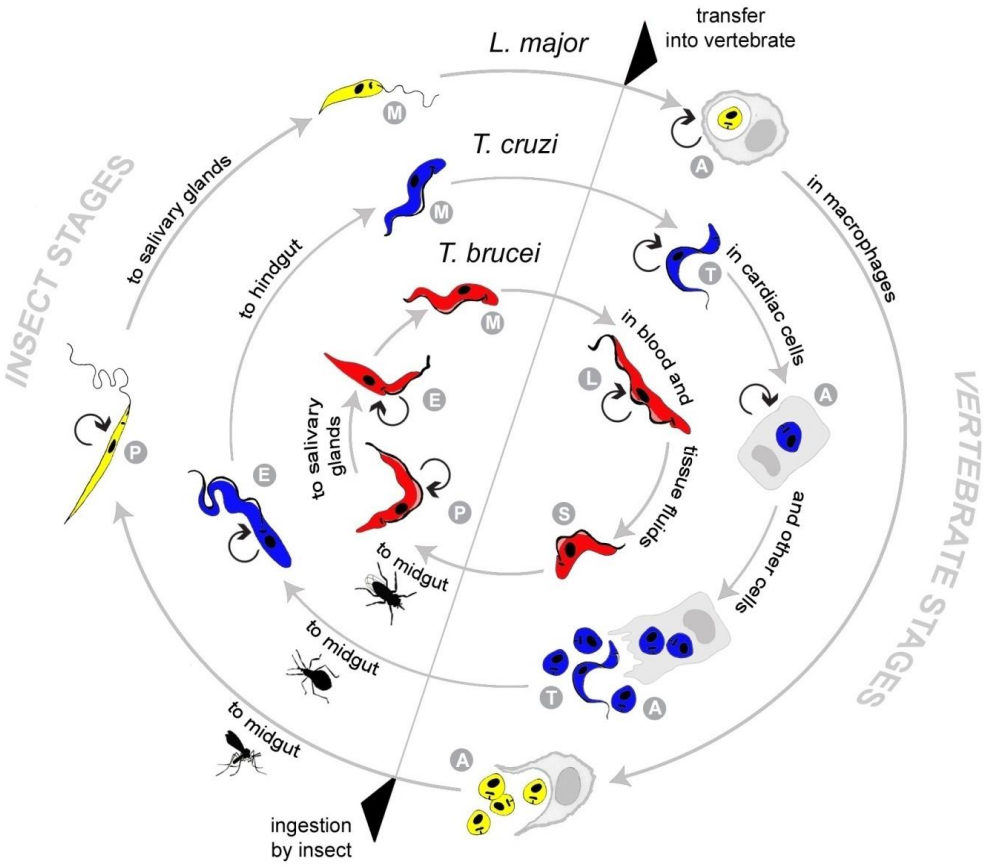

Figure 1: Lifecycles of the TriTryp parasites.

The digenetic lifecycles of the three major trypanosomatid disease agents, *T. brucei*, *T. cruzi*, and *L. major* are shown. Each parasite switches between an insect vector and a mammalian host, undergoing numerous transcriptional and morphological changes along the way in order to adapt to the very different environments and immune systems of their insect and mammalian hosts. While *T. cruzi* and *L. major* are both intracellular parasites which invade, or are phagocytosed by mammalian host cells, *T. brucei* has adapted to survive in the extracellular environment of the bloodstream, using a thick glycosylated protein coat to evade immune detection. (source: Najib El-Sayed)

Leishmaniasis is spread by bites of female sandflies in the genus *Phlebotomus* and *Lutzomyia* and is endemic in 98 countries. The disease is spread by over 20 different species of *Leishmania*, with symptoms ranging from self-healing lesions (cutaneous leishmaniasis) to a visceralizing and potentially fatal form (visceral leishmaniasis). It is

estimated that there are 0.7 to 1.2 million new cases of cutaneous leishmaniasis and 0.2 to 0.4 million new cases of visceral leishmaniasis each year (Alvar et al. 2012).

Despite many years of intensive research efforts, including the sequencing of the genomes of several high-profile trypanosomatid species, with few exceptions (Khare et al. 2016), effective treatments for these diseases remains elusive (J. Clayton 2010; Stich, Ponte-Sucre, and Holzgrabe 2013; Santos et al. 2008; Peña et al. 2015). A major hurdle to drug development efforts is the lack of annotation for many trypanosomatid genes (Gardner et al. 2002; Downing et al. 2011; Dharia et al. 2010; Pink et al. 2005). The large evolutionary distance separating trypanosomatids from the rest of Eukaryota has limited the effectiveness of traditional homology-based annotation, resulting in a majority of genes lacking meaningful annotation (Simpson, Stevens, and Lukes 2006; Choi and El-Sayed 2012). An incomplete understanding of the precise mechanisms of expression regulation has further hindered efforts to develop new treatments (Palenchar and Bellofatto 2006; Kramer 2012).

In recent years, the enormous growth of genomic and transcriptomic datasets, coupled with the continued sophistication of the computational tools and techniques used to analyze them, have led to countless discoveries, impacting all areas of biological research (Binnewies et al. 2006; Koboldt et al. 2013), including parasitology (Andersson 2010; Brehm, Carlton, and Hoffmann 2012; Forrester and Hall 2014) and host-pathogen interactions (Diehn and Relman 2001; König et al. 2008). We now have an unprecedented opportunity to probe the molecular interactions that occur between host and parasite cells during infection. Furthermore, there has been an increasing appreciation for the need to take an integrative approach to studying biological systems,

resulting in a wealth of new systems biology resources and methods (Ng et al. 2006; Chuang, Hofree, and Ideker 2010; Ghosh et al. 2011). Among these, co-expression networks have emerged as a particularly powerful approach to analysing high-throughput transcriptomics datasets (D'haeseleer, Liang, and Somogyi 2000; Zhang and Horvath 2005; Saha et al. 2017).

In this thesis, tools and techniques based on co-expression network analysis and related approaches are developed and are used to improve the structural and functional annotations of trypanosomatid genomes, as well as to elucidate important host and parasite interactions relating to infection. In Chapter 2, a bioinformatics pipeline for the annotation of trypanosomatid gene structure is described, and is used to annotate and explore important features in the *T. cruzi* genome including untranslated region (UTR) boundaries, and alternative trans-splicing and polyadenylation events. In Chapter 3, a generalized framework for co-expression network construction from RNA-Seq data is developed, and the role of various data transformation and network construction parameters are explored. Complementary methods for evaluating network quality and combining information across multiple network parameterizations to improve robustness are described. These techniques are then used to construct four high-quality host and parasite co-expression networks, and important features of each of the individual networks are explored. In Chapter 4, a simple differential network analysis is developed, and is used to investigate the impacts of infection on host co-expression networks, as well as to detect shared transcriptional signatures of intracellular trypanosomatid infection.

The work described in this thesis was both influenced by, and supportive of, additional co-authored research not discussed here. An early version of the

trypanosomatid gene structure analysis pipeline described in Chapter 2 was used to characterize the 5' and 3' UTR boundaries for *L. major* strain Friedlin, and to detect and visualize alternative trans-splicing and polyadenylation events (Dillon, Suresh, et al. 2015). Co-expression network analysis techniques related to those described in Chapter 3 were used to investigate the relationship between host co-expression modules and sample trait data for a *Leishmania braziliensis* biopsy study (Christensen et al. 2016). Transcriptome approaches developed for analysis of host-Mycobacterium RNA-Seq data (Quigley et al. 2017) helped to ensure all software developed throughout this thesis could be easily extended to data from arbitrary species, and provided insights into the co-expression landscape for alternative host-pathogen systems. The co-expression network construction and optimization techniques described in Chapter 3 were also used to determine important parasite pathways relevant to survival in insect-stage forms of *L. major* (Inbar et al. 2017).

Several of the RNA-Seq datasets used throughout this thesis were collected by other members of the El-Sayed lab. Yuan Li generated the *T. cruzi* Y strain RNA-Seq samples, Laura Dillon and Rahul Suresh generated the *L. major* strain Friedlin promastigote and murine infection samples, and the *L. major* strain Friedlin human infection samples were generated by Cecilia Fernandes. All software used for the *T. cruzi* gene structure analysis (Chapter 2) was written by Keith Hughitt, with specific figures and tables based, in part, on an earlier version of the analysis performed by Yuan Li.

Differential expression analysis and batch adjustment code used in this thesis was heavily influenced by code written by Laura Dillon, in collaboration with Kwame Okrah and Hector Corrada Bravo (Chapter 3). The co-expression network analysis code

(Chapter 3) makes use of helper functions for basic data preprocessing, quality assurance plots, etc. written by Trey Belew and distributed as part of the hpgltools package for R. The statistical model used to visualize the relationship between various network parameter combinations and associated network scores, and the use of -log10 adjusted *P*-value scores for evaluating consensus networks (Chapter 3) was developed in collaboration with Hector Corrada Bravo.

Chapter 2

Trypanosomatid Gene Structure Analysis

Introduction

Trypanosoma cruzi is the etiological agent of American Trypanosomiasis, also known as Chagas disease. Recent estimates indicate that ~8 million people throughout Central and South America are infected with this parasite and 100 million people are at risk of infection (Coura and Dias 2009; Rassi and Marin-Neto 2010; Bern 2015). *T. cruzi* has a complex life cycle, involving insect and mammalian hosts with four distinct developmental stages: epimastigotes, metacyclic trypomastigotes, amastigotes and bloodstream trypomastigotes (Brener 1971; Miles, Feliciangeli, and de Arias 2003; Minning et al. 2009). Each developmental stage brings with it a unique set of challenges, requiring dramatic changes in the parasite's transcriptome in order to respond to changing conditions such as metabolic constraints and host immune defenses. The precise mechanism for the large-scale transcriptome remodelling which occurs at each of these developmental stages is still largely unknown. (Kramer 2012).

The draft sequence of *T. cruzi* CL Brener genome has provided a framework for gene identification and functional annotation which contributed significant insights into the biology and metabolism of the parasite, as well as an architectural landscape allowing comparative genome analyses with two other model trypanosomatids, *Trypanosoma brucei* and *Leishmania major*, etiological agents of African sleeping sickness and leishmaniasis, respectively (El-Sayed, Myler, Blandin, et al. 2005). Following the availability of the genome sequence, more than 12,000 allelic pairs of genes were

predicted, including 3,590 pseudogenes (El-Sayed, Myler, Bartholomeu, et al. 2005). Yet a systematic genome-wide identification of transcripts has not been conducted for *T. cruzi* and a major challenge remains to identify the bona fide transcripts and the exact boundaries of the genes encoding them.

Unlike other eukaryotes, trypanosome genes coding for proteins with unrelated functions are organized into co-directional clusters that undergo polycistronic transcription by RNA polymerase II (Pol II) (El-Sayed, Myler, Blandin, et al. 2005; T. Nicolai Siegel, Tan, and Cross 2005; T. Nicolai Siegel et al. 2011). Most chromosomes contain at least two polycistronic gene clusters, which can be either divergently or convergently transcribed (Weatherly, Boehlke, and Tarleton 2009). Trans-splicing, together with polyadenylation, allows polycistronic transcripts to be processed into monocistronic units ready for translation (El-Sayed, Myler, Blandin, et al. 2005; Daniels, Gull, and Wickstead 2010). In the trans-splicing event, a 39-nt splice leader (SL) sequence is transferred from SL RNA to the 5′ end of every mRNA, providing the cap structure needed (Agabian 1990). The signal directing trans-splicing events has been reported as an AG dinucleotide with an upstream polypyrimidine tract of varying length (Michaeli 2011). Trans-splicing events are spatially and temporally coordinated with the polyadenylation events (LeBowitz et al. 1993; Matthews, Tschudi, and Ullu 1994). There is no recognized consensus polyadenylation signal in the 3′ UTR. Instead, evidence from a small number of loci in the related trypanosomatid species, *T. brucei* and *L. major*, suggests a preferential usage of polyadenylation sites around groups of adenines (Benz et al. 2005; Dillon, Okrah, et al. 2015). However, similar analyses have not been performed for *T. cruzi*.

The lack of identifiable RNA pol II promoters in trypanosomatids suggests that these organisms lack precise transcriptional control over the majority of their genes (T. Nicolai Siegel et al. 2011). Thus, regulation of gene expression occurs mainly at the post-transcriptional level, through pre-mRNA processing, RNA degradation, or translational repression (Martínez-Calvillo et al. 2010; Kramer 2012). Both the 5′ UTR and 3′ UTR can be involved in stabilization-destabilization mechanisms, up-regulating and down-regulating mRNA levels in a developmentally regulated manner (Nozaki and Cross 1995; Furger et al. 1997; Di Noia et al. 2000). Indeed, an in silico investigation by De Gaudenzi et al. (2013) found evidence for enrichment of RNA structural motifs in noncoding regions of groups of functionally-related *T. cruzi* genes (De Gaudenzi et al. 2013).Without mRNA expression data, however, the authors were limited to using sets of annotated genes, with the assumption that those genes would be co-regulated.

Heterogeneity of RNA processing sites, present as alternative trans-splicing and polyadenylation sites, have been detected at high frequency across the genome in *T. brucei*, *T. vivax* and *L. major*, potentially modifying accessibility of RNA binding sites to RNA binding proteins which are increasingly believed to play an important role in trypanosomatid gene regulation (Kolev, Ullu, and Tschudi 2014; Romaniuk 2016; C. Clayton 2013; C. E. Clayton 2014; Gazestani, Lu, and Salavati 2014; Dillon, Okrah, et al. 2015) In *C. elegans*, where trans-splicing is common, the frequency of alternative trans-splicing is much lower than that observed in trypanosomes (Mangone et al. 2010). The significance of this phenomena and its potential role in posttranscriptional regulation has not been fully investigated (Kolev et al. 2010; Greif et al. 2013).

Here, we have generated the first complete transcriptome map for *Trypanosoma cruzi* using high-depth multi-stage RNA sequencing technology (RNA-Seq) data. We mapped transcribed regions at single nucleotide resolution on a genomic scale, retrieved trans-splicing and polyadenylation sites for *T. cruzi* across various developmental stages, and both enhanced and curated the current genome annotation. We also discovered a large heterogeneity of RNA processing sites across the genome, noting the preference of different primary sites in various developmental stages, potentially playing a role in stage-specific gene regulation.

Experimental samples interrogated using RNA-seq

We applied an RNA-seq approach to characterize the global transcriptome of the *T. cruzi* Y strain parasite across different life cycle stages including various time points in its intracellular life cycle (Y. Li et al. 2016). These included two extracellular forms (epimastigote and trypomastigote) and the intracellular forms (amastigotes) at 4, 6, 12, 24, 48 and 72 hrs post-invasion of human foreskin fibroblast (HFF) cells. *In vitro* infection experiments were repeated on different dates and for each of the developmental stages, sequence data was collected from two to four independent biological replicates, generating a total of 950 million pairs of 100 bp reads from 34 samples (see Table S1 in Li *et al.*, 2016). Because the goal of this study was to characterize the gene structure in *T. cruzi*, only the reads mapping to the parasite genome (~300 million) were used.

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

The significance of defining distinct transcript boundaries and gene structure (regulatory) elements for each of the 10,339 *T. cruzi* protein-coding genes in the current database is of particular relevance to the biology of this pathogen which, like other trypanosomatids, lacks tight transcriptional control of its polycistronic gene clusters. We exploited two mRNA sequence features (*trans*-splicing of a mini-exon SL RNA sequence and polyadenylation) to accurately map the 5′ and 3′ boundaries of genes. We selected subsets of reads that ended with a minimum match of three nucleotides to the 3′ portion of the spliced leader (SL) sequence or a minimum of four adenine residues. A total of 2,879,187 SL-containing and 2,733,307 poly(A)-containing reads were identified, respectively 1.01% and 0.96% of the reads that mapped to *T. cruzi* reference genome (**Table S1**). Trimming of the SL and Poly(A) sequences and mapping the remaining tags back to the genome allowed us to detect at least one unique SL-addition site for 7,813 genes (75.6% of the 10,339 annotated protein-coding genes) and at least one polyadenylation site for 7,662 genes (74.11%) in one or more of the developmental stages.

Using the coordinates of the SL-addition sites and existing start codon annotations for coding sequences (CDS), we defined the boundaries of all 5′ UTRs in the *T. cruzi* genome. The median 5′ UTR length was 68 nt with a range from 1 nt, adjacent to the initiation codon, to 12,506 nt (**Figure 2A**). The distribution of the 5′ UTR lengths was similar across different life stages of the parasite. A similar analysis using poly(A) addition sites and existing stop codon annotations allowed us to determine the boundaries

of the 3′ UTRs. The 3′ UTRs had a median length of 124 and a range from 1 to 16,764 nt (**Figure 2B**). The median mRNA length was 1,647 nt, of which the 5′ UTR, CDS, and 3′ UTR accounted for 8.4%, 77.8%, and 13.8%, respectively. The 5′ UTR, 3′ UTR, and CDS lengths were found to be weakly correlated (spearman correlation value between 0.10-0.16). (**Figure 3**). With the gene boundaries accurately mapped, we were also able to define the intergenic regions. The median value for intergenic spacing was 381 nt.

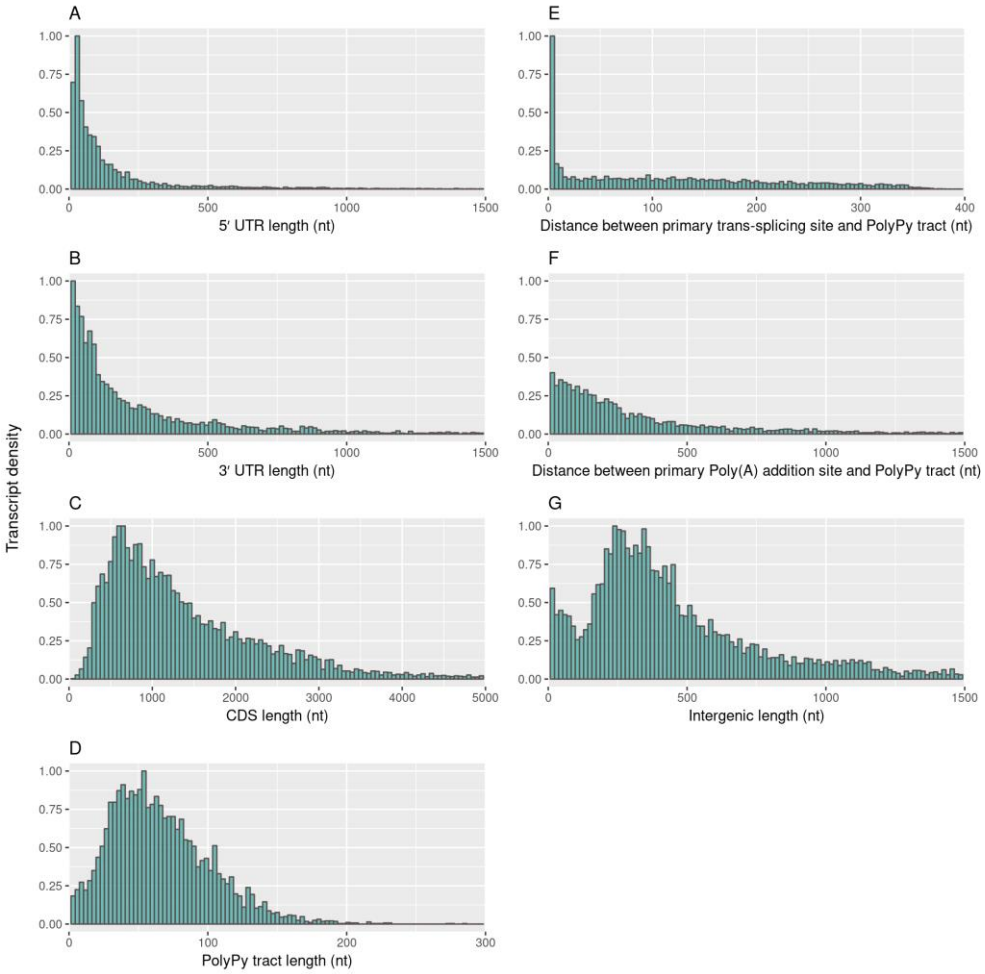

Figure 2: Length and position distribution of gene structure components in *T. cruzi*.

(A) Distribution of 5′ UTR lengths. A Spliced Leader (SL) analysis of reads obtained from all developmental stages was performed to map the exact *trans*-splicing sites associated with each CDS and to identify the coordinates and lengths of the 5′ UTRs. For each panel, "transcript density" reflects the relative proportions of transcripts at each length. Ninety genes with 5′ UTR length greater than 1.5 kb are not shown. **(B)** Distribution of 3′ UTR lengths. A polyadenylation site analysis of sequences from all developmental stages was performed to map the poly(A) addition site associated with each CDS and to identify the coordinates and lengths of the 3′ UTRs as described in Methods. Two-hundred and thirty-five genes with 3′ UTR length greater than 1.5 kb are not shown. **(C)** Distribution of CDS lengths. Start and

stop coordinates for coding sequences were retrieved for the *T. cruzi* CL Brener reference genome - Esmeraldo haplotype (TriTrypDB, version 29). Two-hundred and twenty-one genes with a CDS length greater than 5000 nt are not shown. **(D)** Distribution of polyPy tract lengths. PolyPY tracts were identified by scanning a 400 nt window upstream of the primary *trans*-splicing site to. One gene with polypyrimidine tract length greater than 300 is not shown.**(E)** Distribution of distances between primary *trans*-splicing sites and the 3′ end of each polyPy tract. **(F)** Distribution of distances between primary polyadenylation sites and downstream 5′ ends of each polyPy tract, excluding 433 genes with distances greater than 1500 nt. **(G)** Size distribution of intergenic regions. Intergenic regions were mapped between the primary polyadenylation site and the primary *trans*-splicing site of the downstream gene and lengths computed for the pooled data from all developmental stages. Five-hundred and twenty-two genes with intergenic distances greater than 1500 nt not shown.

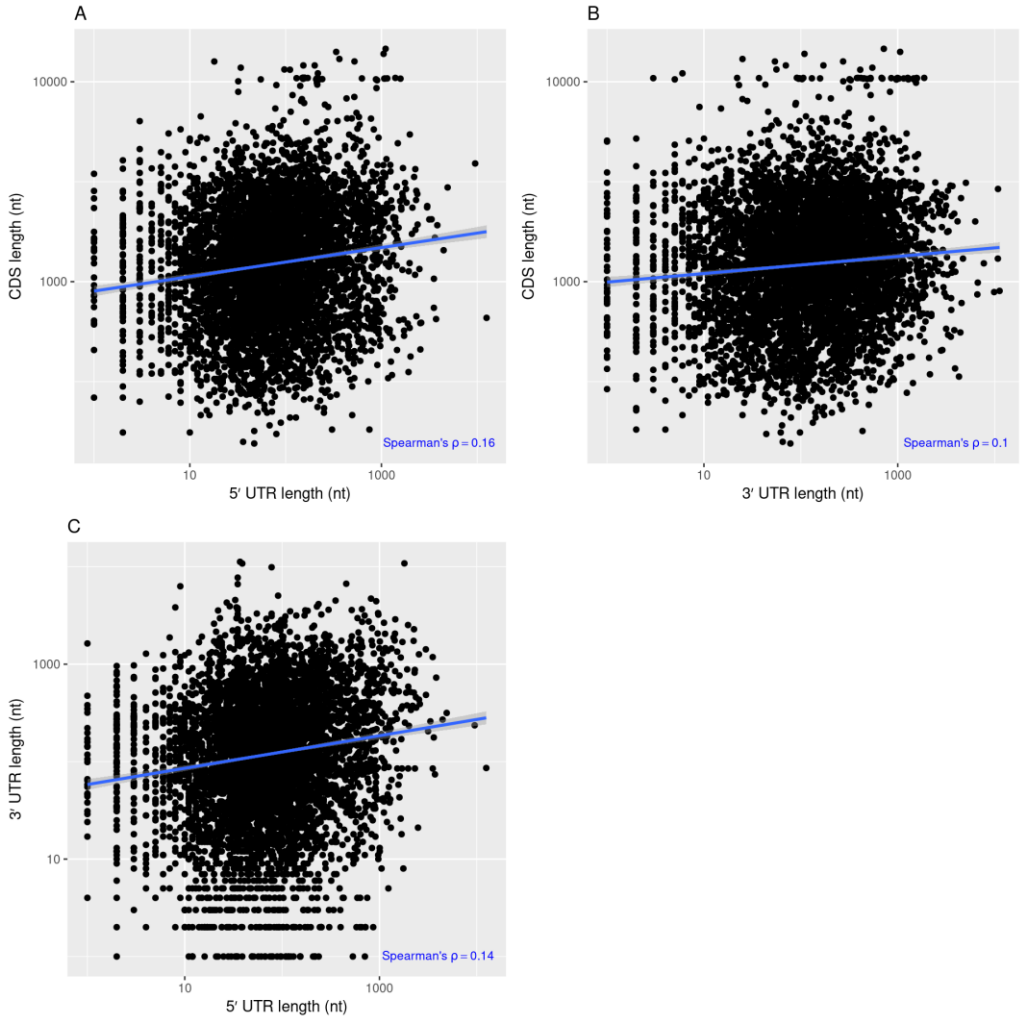

Figure 3: Correlations between CDS and UTR lengths.

(A) Scatterplot of 5′ UTR and CDS lengths, **(B)** Scatterplot of 3′ UTR and CDS lengths, **(C)** Scatterplot of 5′ UTR and 3′ UTR lengths. A regression line was fitted using the robust linear model (rlm) function. The Spearman correlation value (rho) is shown in the lower right corner of each panel.

We extended our analysis to include the examination of features known to be involved in the regulation of RNA processing events. Polypyrimidine (polyPy) tracts are

located upstream of the 5′ splice acceptor sites and provide a required signal for the *trans*-splicing machinery (Huang and Van der Ploeg 1991; Günzl 2010; T. Nicolai Siegel, Tan, and Cross 2005). We searched for the longest stretch of pyrimidine residues, interrupted by no more than two contiguous purines and located upstream of each of the (primary) *trans*-splicing sites. PolyPy tracts ranged from 1 to 311 nt in length, with a median of 66 nt (**Figure 2D**) and a clear preference for thymine (63.8 %) over cytosine (36.2%) residues. This observation is consistent with the findings in *T. brucei* and *T. vivax* (Kolev et al. 2010; Greif et al. 2013) and our logo analysis shown in **Figure S1E**. In contrast, polypyrimidine tract lengths ranged from 7 to 123 nt in *L. major*, with a median length of 21 nt and a preference for cytosine (54%) over thymine (42%) (Dillon, Okrah, et al. 2015). The distance between the polyPy tract and the upstream primary polyadenylation site exhibited a relatively tight distribution with a median value of 185 nt (**Figure 2F**) which was longer than the median distance (97 nt) between polyPy tract and the downstream primary SL-addition site (**Figure 2E**). This spacing of the polyPy tract in about two thirds of intergenic regions can also be noted in *T. brucei* (Kolev et al. 2010).

With the canonical features for the *T. cruzi* genes resolved at the single nucleotide level, we compared them to their orthologous features reported in *T. brucei* (Kolev et al. 2010) and *L. major (Dillon, Okrah, et al. 2015).* A striking aspect we observed was the shorter median length of the 5′ and 3′ UTR regions in *T. cruzi* genes when compared to either *T. brucei* or *L. major* (**Table 1**). To account for possible biases from species-specific multigene families, we further restricted our analysis to the subset of three-way clusters of orthologous genes (COGs) in the reference trypanosomatid (TriTryp) genomes

(El-Sayed, Myler, Blandin, et al. 2005) and obtained similar results.

	T. brucei	*T. cruzi*	*L. major*
5′ UTR	91	68	547
CDS	1,242	1,136	1,241
3′ UTR	388	124	729
Poly(A) - PolyPy distance	81	185	558
PolyPy tract	18	60	21
SL - PolyPy distance	43	97	64

Table 1: Comparative gene structure feature lengths for *T. cruzi, T. brucei, and L. major*.

Median lengths/distance of 5′ UTR, coding sequence (CDS), 3′ UTR, Poly(A) site - Poly(Py) tract distance, Poly(Py) tract, SL addition site - Poly(Py) for *T. cruzi*, *T. brucei* (Kolev et al. 2010), and *L. major (Dillon, Okrah, et al. 2015)*. Note that for the purposes of computing polypyrimidine tracts, a 400 nt window upstream of the primary SL site was scanned, allowing up to two consecutive purines to be present in a polypyrimidine tract. This is different from the approach used by *Dillon et al.* (*L. major*) and *Kolev et al.* (*T. brucei*), whereby a 250 nt window was scanned, allowing only a single purine interruption.

This relative compaction of the *T. cruzi* UTRs is congruous with earlier observations we made at the level of coding sequences, whereby the mean length of CDS regions in *T. cruzi* was shorter than in *T. brucei* and *L. major* (El-Sayed, Myler, Blandin, et al. 2005). Despite the fact that *T. cruzi* has a notably smaller genes (CDS + UTRs), our characterization here of the true intergenic region (and not the inter-CDS region) reveals longer intergenic regions in *T. cruzi* when compared to *T. brucei*. We infer that the higher gene density in the *T. cruzi* genome is a direct consequence of shorter UTR lengths. We also note that the GC content of CDS is the highest (52.6%), followed by 5′ UTR (44.7%) and 3′ UTR (43.5%) , with the lowest level detected in intergenic regions (40.8%).

Alternative RNA processing sites

The sequencing depth of our transcriptome profiling experiments allowed not only the identification of the SL-addition and polyadenylation sites at a single-base

resolution, but also the characterization of widespread alternative RNA processing events in *T. cruzi*. Because our study design was mainly aimed at a quantitatively unbiased profiling of the transcriptome of the parasite at various developmental stages, we did not enrich for SL- or poly(A)-containing reads during library construction and relied on deep coverage to collect large numbers of reads from both ends of transcripts (2.9 million SL- and 2.7 million poly(A)-containing reads). This approach permitted us to identify as well as quantitate differential RNA processing events.

Of the 7,814 genes with SL-addition sites detected, 93.2% used more than one trans-splicing site in at least one developmental stage (29.2% used two to four trans-splicing sites and 63.9% had five or more sites). This observation is similar to what has been reported for *T. brucei* where 89% of genes showed evidence of alternative trans-splicing events (Kolev et al. 2010). An examination of the *trans*-splicing sites revealed a propensity for usage of the canonical acceptor sequence (AG) both at the primary (85%) and alternative (32%) splicing sites (**Figure S1A**). The distribution of the distances between the primary and alternative *trans*-splicing sites both using the AG acceptor sequence revealed that a slight majority (56%) of the alternative splice sites are located downstream of the primary site (**Figure S1B**). This observation is consistent with a model that proposes that the 3′ splice site in mammalian introns is located by a scanning process that recognizes the first AG downstream of the branch point in a sequence-specific context (Mount 1983; Smith et al. 1989).

The usage of non-canonical acceptor sequences was much more pronounced in the alternative (68%) than in the primary (15%) splice sites (**Figure S1A**). Unlike for the canonical acceptor sequences, the distribution of the distances between the non-canonical

primary and the alternative splicing sites was heterogenous and showed no distinct pattern, with the exception of a peak of alternative *trans*-splicing events that occurred within the 10 nt that precede the primary site (**Figure S1C**). A similar peak can be observed for AG splicing sites, albeit less pronounced. Those peaks may reflect sloppy splicing events near the primary SL-addition site as has been suggested previously for *T. brucei* (Kolev et al. 2010; Tim Nicolai Siegel et al. 2010) and/or slippage due to the presence of tandem NAGNAG acceptor sites. We also noted a slight preference for GG and TT among both the canonical and non-canonical acceptor sequences. We are unable to rule out polymorphisms between the strain used in this study (strain Y) and the reference genome (CL Brener) as a source, at least in part, of this bias as well as some of the non-canonical sites we observe.

A significant proportion (22.4%) of SL-addition sites were detected within known annotated CDS regions, reflecting two possible scenarios: mis-annotated initiation start sites or events that have the potential to impact the resulting protein product(s). An example of the latter is the gene encoding LYT1 (TcCLB.503829.50), for which we detected three alternative SL-addition sites at positions -44, -46 and +10 relative to the start codon (+1). Of these, two of the sites precisely match coordinates which were carefully mapped in a previous study that also showed that *T. cruzi* LYT1 generates two protein products, a process mediated through stage-regulated alternative *trans*-splicing events (Manning-Cela, González, and Swindle 2002). The shorter product (28 amino acid truncation) localizes to the mitochondrial kinetoflagellar zone, whereas the longer product localizes on the plasma membrane (Benabdellah, González-Rey, and González 2007)

Of the 7,662 genes for which polyadenylation sites were identified, 7,146 use an alternative polyadenylation site in one or more developmental stages, 19% of which used two to four polyadenylation sites and 73% used more than four sites. The heterogeneity observed is similar to what we have calculated for *L. major* where 92% of genes display evidence for alternative polyadenylation events (Dillon, Okrah, et al. 2015). The distribution of the distances between primary and alternative polyadenylation sites revealed that ~25% of the minor sites were located within a 200 nt window centered around the primary site, whereas 10% were located within a 20 nt window (**Figs. S1D**). This abundant heterogeneity of polyadenylation sites at several tightly spaced positions has been observed not only in *T. brucei* (Kolev et al. 2010; Benz et al. 2005; Matthews, Tschudi, and Ullu 1994), but also in other systems, including plants, animals and human (Elkon, Ugalde, and Agami 2013; Ji et al. 2014).

Alternative trans-splicing events across parasite development

The heterogeneity of RNA processing that we detected at high frequency in the form of alternative *trans*-splicing and polyadenylation events may be the manifestation of another level of gene expression regulation. To examine the role and dynamics of RNA processing events at different stages of the development of the parasite, we examined the usage of alternative SL-addition sites across the various developmental stages. Our analyses included reads from the trypomastigote, epimastigote and combined intracellular amastigote samples.

In order to identify a subset of genes that switch the use of their primary *trans*-splicing site across different developmental stages, we plotted 5′ UTR lengths against one another for each pair of stages. The size and color of plot points scale with respect to the

minimum number of RNA-Seq reads supporting the putative UTR boundaries, and those genes which use the same primary site across stages, and thus would fall along the diagonal of the plot, are excluded while data points which are far off the diagonal and larger/redder are good candidates for stage-specific trans-splicing. (**Figure 4**). Among the genes with evidence for site switching between the trypomastigote and amastigote stages, 353 were detected at least three reads supporting each stage.

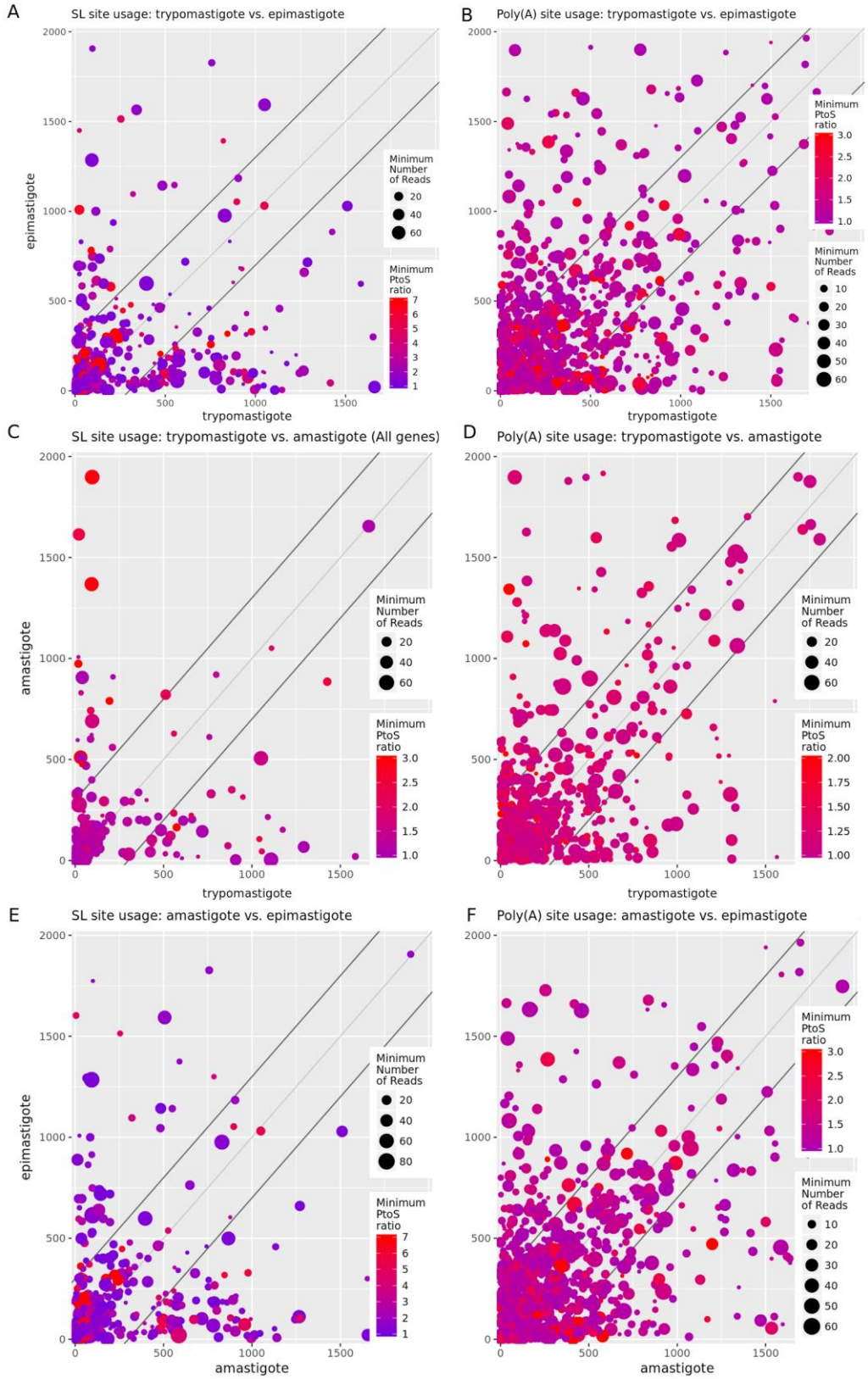

Figure 4: Alternative splicing and polyadenylation profiles across T. cruzi trypomastigote, epimastigote, and amastigote stages.

(A) 5′ UTR length is plotted in trypomastigotes (x-axis) and epimastigotes (y-axis) for each gene for which the primary site location changed by at least six nucleotides across stages, with each detected primary site having at least three supporting RNA-Seq reads. The size of each point represents the minimum read coverage for the primary boundaries across the two stages while the size of each point represents the minimum ratio of SL-containing reads mapping to the primary site over the secondary site (P/S) within each stage. (B) 3′ UTR length is plotted in trypomastigotes (x-axis) and epimastigotes (y-axis) in a similar manner. Similar comparisons are shown for trypomastigotes vs. amastigotes (C-D) and amastigotes vs. epimastigotes (E-F).

Relationship between *T. cruzi* gene expression and UTR length

One mechanism which is believed to play a role in trypanosomatid post-transcriptional gene regulation is the modulation of mRNA stability and decay rates (C. Clayton and Shapira 2007; Fadda et al. 2014; Furger et al. 1997; D'Orso, De Gaudenzi, and Frasch 2003). In particular, 3′ UTR length is thought to play a role in mRNA turnover rates for some organisms either directly (Hogg and Goff 2010; Mishima and Tomari 2016) via mechanisms capable of sensing 3′ UTR length, or indirectly through variable inclusion of regulatory elements by alternative polyadenylation events (T. Nicolai Siegel et al. 2011; Rodrigues et al. 2010; Furger et al. 1997). While much of the focus on post-transcriptional regulatory elements in trypanosomatids has focused on the longer 3′ UTR's, the presence of widespread alternative trans-splicing sites described above prompted us to investigate the possible role of both 5′ and 3′ UTR's on gene expression.

In order to get a global view of the role of UTR length on gene expression, a co-expression network was generated using the intracellular amastigote stage samples (see Chapter 3 for details), and the average 5′ and 3′ UTR lengths of each detected network module were plotted below a dendrogram representation of the network. (**Figure 5**). If expression levels were unrelated to UTR length, one would expect a random distribution of colors in the plot. The presence of a significant amount of non-random sorting by both

the 5′ and 3′ UTR length in the network plot suggests a possible relationship between gene expression and UTR length.

Next, differential expression analysis was performed (Refer to chapter 3), comparing each possible pair of amastigote timepoints. For each contrast, the distribution of 3’ UTR lengths, separated by differential expression status, were plotted. Large differences in the distributions of 3′ UTR lengths across differential expression status were observed for many of the contrasts, further supporting the idea that expression may be influenced by UTR length in *T. cruzi*. (**Figure 6**).

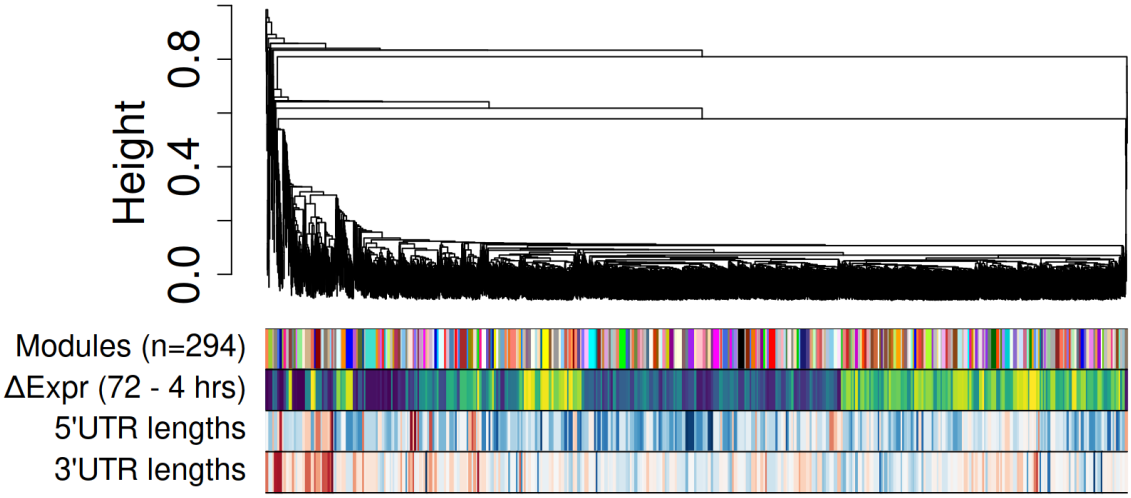

Figure 5: Relationship between gene UTR length and co-expression in *T. cruzi* infecting *H. sapiens*. A co-expression network dendrogram is shown for the optimized TcHs network. The height of the dendrogram indicates the level of dissimilarity in expression profiles between genes, with lower values indicating low dissimilarity, and thus a high degree of co-expression. The closer genes are to one another in the dendrogram, the more similar their expression patterns are across time. Below the dendrogram, co-expression module membership is shown, with each color bar indicating a separate co-expression module. The second row of annotations shows the average change in log2-CPM gene expression between 4 and 72 hrs, for each module ranging from purple (modules whose expression decreases between 4 and 72 hours) to yellow (modules whose expression increases between 4 and 72 hours). The last two annotation rows correspond to average 5′ and 3′ UTR lengths, respectively, for each co-expression module. Blue indicates a shorter UTR length, while red indicates a longer UTR length. The non-random sorting of blue and red visible in the 5′ and 3′ UTR lengths color bars hints at a possible relationship between UTR length and gene expression in *T. cruzi*.

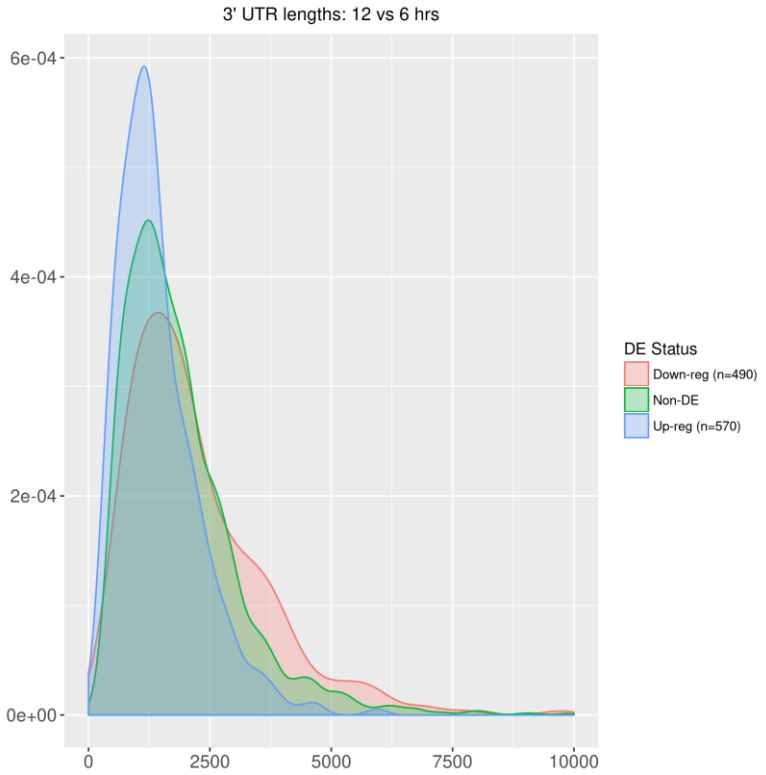

Figure 6: Distribution of 3′ UTR lengths for up- and down-regulated genes.
The distribution of 3' UTR lengths, separated by differential expression status, is shown for a representative differential expression contrast (12 vs. 6 hrs). Genes which increased in expression between 6 and 12 hours (blue) tended to have shorter 3' UTRs. Conversely, genes which decreased in expression across the same time period (red), tended to have longer 3' UTRs.

Relationship between *T. cruzi* gene expression and UTR sequence composition

In addition to exploring the relationship between UTR length and gene expression in *T. cruz*i, we also investigated whether basic sequence composition features such as GC- and CT-richness vary as a function of expression. While, in general, 5′ and 3′UTR GC- and CT-richness tended to be similar between up- and down-regulated *T. cruzi* genes, some small biases were observed. Genes which increased in expression across time tended to have less GC-rich 5′ UTRs, compared with non-differentially expressed genes or genes that decreased in expression over time. For example, the median GC-richness for the 847 down-regulated genes from 12 to 24 hours was found to be 47.2%,

while the GC-richness for the 856 genes which increased in expression over the same time period was determined to be 41.7%.

Conservation of gene structure across *Trypanosoma cruzi* strains

Finally, in order to study conservation of gene structure across strains, a similar analysis was performed using *T. cruzi* strain Sylvio RNA-Seq samples generated in our lab (Houston-Ludlam, Belew, and El-Sayed 2016). We compared the 5′ UTR lengths for a total of 5,845 genes for which the boundaries were determined in both strains. Among these genes, the median UTR length was highly similar between *T. cruzi* Y strain (59 nt) and sylvio (56 nt). For the 5,834 genes for which 3′ UTR boundaries could be determined for both strains, Sylvio genes tended to use a more distant primary polyadenylation site, and thus have slightly longer 3′ UTRs compared with Y strain (152 vs. 124 nt, respectively). Overall, 5′ and 3′ UTR lengths were moderately conserved across strains with a Pearson correlation value ranging from 0.74 (5′ UTRs) to 0.75 (3′ UTRs) (**figure 7**).

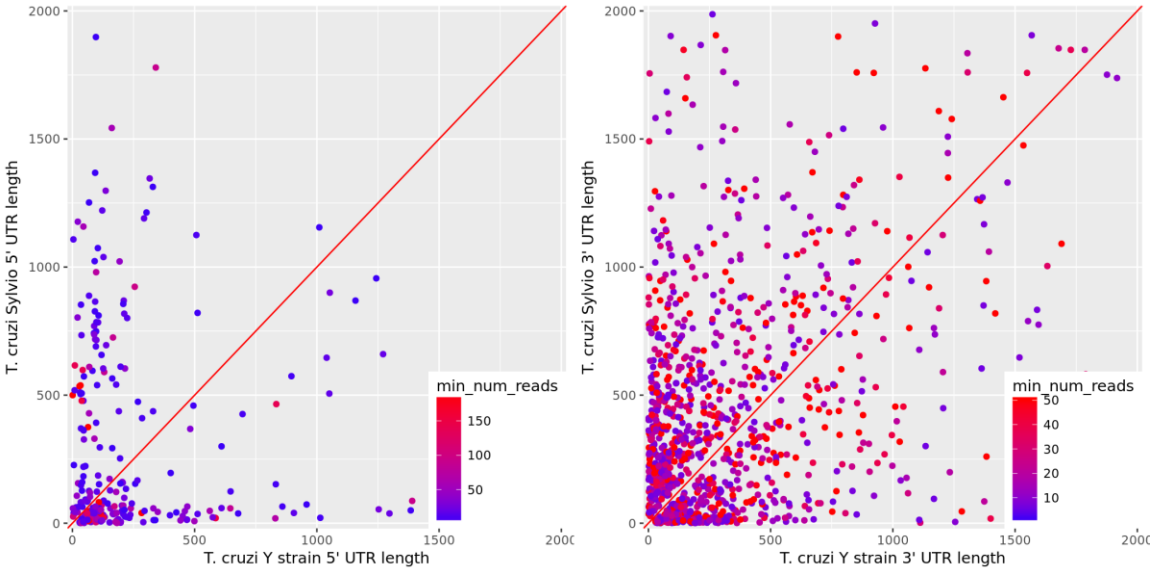

figure 7: Conservation of T. cruzi 5′ and 3′ UTR lengths across strains.

(A) 5′ UTR lengths are plotted for T. cruzi Y strain (x-axis) vs. T. cruzi sylvio (y-axis). Of the 5,845 genes whose 5′ UTR boundaries could be determined in both strains, the 1,268 genes with 5′ UTR lengths differing by five or more nucleotides across strains and having at least five supporting RNA-Seq reads for each strain are shown. The color of the points represents the minimum number of reads supporting the computed 5′ UTR boundaries across each stage. The red diagonal line indicates where genes with identical 5′ UTR lengths across strains would fall, if shown. (3) 3′ UTR lengths are plotted for T. cruzi Y strain (x-axis) vs. T. cruzi sylvio (y-axis). Of the 5,834 genes whose 3′ UTR boundaries could be determined in both strains, the 2,680 genes with 3′ UTR lengths differing by five or more nucleotides across strains and having at least five supporting RNA-Seq reads for each strain are shown. The color of the points represents the minimum number of reads supporting the computed 3′ UTR boundaries across each stage. The red diagonal line indicates where genes with identical 3′ UTR lengths across strains would fall, if shown.

Conclusion

In this chapter, we provide a detailed map of the stage-specific *T. cruzi* transcriptome at the single-nucleotide resolution. Trans-splicing acceptor sites and polyadenylation sites are detected, and UTR boundaries and additional gene structure elements are defined for most of the currently annotated *T. cruzi* genes. Consistent with what has been observed for other trypanosomatid species (Kolev et al. 2010; Dillon, Okrah, et al. 2015), we detected widespread alternative trans-splicing and polyadenylation events. One unexpected finding is an apparent relationship between 5′ and 3′ UTR length and gene expression across time. While addition work is required to verify the relationship, if true, it could have significant implications for our understand of trypanosomatid gene regulation.

In addition to providing a valuable resource for the interrogation of individual genes, the global transcriptome map described here offers an unprecedented opportunity for the discovery of important parasite regulatory elements. The precise definition of *T. cruzi* UTR boundaries, along with the characterization of the *T. cruzi* co-expression network, allows for the systematic detection of UTR sequence elements involved in expression regulation. In the next chapter, we develop an approach for inference of a

robust co-expression network built from RNA-Seq data and apply it to construct a *T. cruzi* intracellular and extracellular stage co-expression network.

Methods

Detection of trans-splicing acceptor and polyadenylation sites

RNA-Seq reads were pooled across biological replicates for each of eight different *T. cruzi* developmental stages from a dataset previously generated in our lab (Y. Li et al. 2016) (**Table S1**): epimastigotes, trypomastigotes, and intracellular amastigotes at 4, 6, 12, 24, 48 and 72 hrs post infection (hpi) of human foreskin fibroblasts (HFF). Reads were mapped to the Esmeraldo-like haplotype of the *T. cruzi* CL Brener reference genome (El-Sayed, Myler, Bartholomeu, et al. 2005; Trapnell, Pachter, and Salzberg 2009) using Tophat (v2.1.1) ((El-Sayed, Myler, Bartholomeu, et al. 2005; Trapnell, Pachter, and Salzberg 2009)) and allowing for a single mismatch, and excluding mixed and discordant reads. Reads which successfully mapped to the genome (and therefore unlikely to contain a portion of the spliced leader (SL) or poly(A) sequence) were filtered out. For amastigote samples which also include host mRNAs, reads were similarly mapped to the human hg38 reference genome (Lander et al. 2001) and excluded from consideration if successfully mapped. The ends of the remaining reads were searched for the presence of at least three nucleotides matching the 3′ end of the SL sequence (AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG) (McCarthy-Burke, Taylor, and Buck 1989) or three adenines in the expected orientation and differing from the corresponding genomic sequence by at least two nucleotides. After trimming the SL or poly(A) fragments, the remaining portion of each read was mapped back to the *T. cruzi* genome, allowing us to precisely map the location of trans-splicing and polyadenylation

sites. The site detection pipeline is freely available at https://github.com/elsayed-lab/utr_analysis (Dillon, Okrah, et al. 2015).

Determination of primary 5′ and 3′ UTR boundaries

Using the trans-splicing acceptor sites and polyadenylation sites detected in the above step, we determined the primary 5′ and 3′ untranslated region (UTR) boundaries for each gene for which sites were detected. Because of the likely existence of unannotated (short) coding regions in the genome that will confound the UTR boundary determination by giving the appearance of excessively long UTRs, we first searched for putative novel open-reading frames (ORFs) in the inter-CDS regions between current *T. cruzi* gene coordinates. Each inter-CDS region was scanned for ORFs at least 30 amino acids long and the single ORF with maximal RNA-Seq coverage density was selected. Next, each possible UTR boundary configuration was scored for pairs of neighboring genes. The scoring metric was designed to rank configurations using as criteria: the number of assigned UTR boundaries, the RNA-Seq total coverage across all chosen sites, the total RNA-Seq read density spanning any detected ORFs, and configurations which *do not* assume the presence of a novel ORF, in that order of priority. By considering pairs of neighboring genes simultaneously, we were able to avoid invalid configurations such as overlapping 5′ and 3′ UTRs, which might occur if a novel ORF is present but not detected. The code for converting SL and polyadenylation site data into UTR boundaries was written in Python and makes use of the NumPy (der Walt, Colbert, and Varoquaux 2011) and BioPython (Cock et al. 2009) libraries. It is freely available at https://github.com/elsayed-lab/gene-structure-determination.

Alternative splicing analysis

For each pair of developmental stages, alternative trans-splicing events were detected by searching for genes with different primary trans-splicing sites across the two stages. Scatterplots were generated showing all such events for which the primary trans-splicing sites were greater than 5 nt apart and each site had at least 3 reads mapped to it.

Splice acceptor site analysis

The splice acceptor site was identified for each gene by extracting the dinucleotide sequence in the genome upstream of the detected trans-splicing site. The sequence composition of the region spanning 90 nt upstream and 10 nt downstream of the *trans*-splicing sites, including both primary and minor sites, was plotted using the seqLogo Bioconductor package (Bembom 2017; R. C. Gentleman et al. 2004).

Polypyrimidine tract characterization

A window of 400 nt upstream of the primary trans-splicing site was scanned to identify polypyrimidine (polyPy) tracts. A polyPy tract was defined as the longest stretch of sequence consisting of pyrimidines, allowing interruption by no more than two purines. The length of the polyPy tract, the distance between the 3′ end of the polyPy tract and the *trans*-splicing site and the distance between the 3′ end of the pyrimidine tract and the polyadenylation site were computed . The nucleotide composition of the polyPy tract was visualized using the seqLogo Bioconductor package.

Alternative polyadenylation analysis

For each pair of developmental stages, alternative polyadenylation events were detected by searching for genes with different primary polyadenylation sites across the two stages. Scatterplots were generated showing all such events for which the primary

polyadenylation sites were greater than 5 nt apart and each site had at least 3 reads mapped to it.

Conservation of site usage across strains

RNA-Seq reads for sixteen *T. cruzi* strain sylvio samples (Houston-Ludlam, Belew, and El-Sayed 2016) were mapped to the *T. cruzi* CL Brener Esmeraldo-like reference genome, and trans-splicing acceptor sites and polyadenylation sites were detected using the approach described above. While mapping reads from both strains to the same reference genome simplifies the process of comparing site usage, it should be noted that this approach is likely to miss important structural differences such as long indels in the UTRs, and as such, should be considered a first approximation at the likely changes across strains. Despite these limitations, this approach can still be useful for detecting large shifts in site usage when both sites are conserved across strains.

Chapter 3

Co-expression Network Construction and Optimization

Introduction

Over the past decade, there has been an explosion in the generation and study of transcriptomic datasets across a wide range of experimental conditions and organisms. Particularly, with the development of RNA-Seq, our ability to gain an unbiased view of the transcriptional state of a population of cells, or more recently individual cells, has provided a powerful tool for biologists working on a diverse set of problems (Wang, Gerstein, and Snyder 2009; McGettigan 2013; Sandberg 2014). Commonly employed approaches for analyzing these datasets including differential expression analysis (DEA) (Soneson and Delorenzi 2013), functional enrichment or gene set enrichment analysis (Hung et al. 2012) and co-expression network analysis (Zhang and Horvath 2005).

While differential expression analysis is most commonly used to characterize sets of up- or down-regulated genes across pairs of experimental conditions (e.g. “infected vs. uninfected”), co-expression network analysis is useful for exploring co-expression relationships among genes across multiple experimental conditions, and, when combined with module detection techniques, allows us to detect sub-groups of co-expressed genes within the larger “up-” and “down-regulated” categories (Serin et al. 2016). A number of different approaches have been developed for constructing a co-expression network from microarray or RNA-Seq expression data (De Smet and Marchal 2010; López-Kleine, Leal, and López 2013). Typically, these methods involve measuring the similarity

between pairwise gene expression profiles and then converting those similarity scores into edge weights in a weighted network (Butte and Kohane 2000; Zhang and Horvath 2005; Song, Langfelder, and Horvath 2012; Gibson et al. 2013).

Once a co-expression network has been constructed for a given transcriptome dataset, we would like a way to evaluate its biological meaningfulness. Because of the varied interactions which ultimately form a co-expression network (Jansen, Greenbaum, and Gerstein 2002; Stuart et al. 2003; Allocco, Kohane, and Butte 2004; Michalak 2008; W. Li, Freudenberg, and Oswald 2015), it is not immediately obvious how this is best approached. Although there have been some attempts to develop general methods to validate empirical networks, often using some variation of robustness analysis (Babtie, Kirk, and Stumpf 2014; Aghdam et al. 2014; Filosi et al. 2014), less work has been done on the specific validation of co-expression networks. Carlson et al. (Carlson et al. 2006) attempted to validate a yeast co-expression network by showing that network connectivity correlated with gene essentiality, while Liang et al. (Liang et al. 2014) used qRT-PCR to show that genes from a grapevine (*Vitis vinifera*) co-expression module enriched for GO terms relating to "response to environmental stress" were co-expressed upon heat-shock. Here, we make use of the observation that "real" biological networks tend to have modules enriched for specific functions (D'haeseleer, Liang, and Somogyi 2000; Michalak 2008) in order to develop a co-expression network scoring approach. This scoring method is then used to compare alternative network parameterizations across a number of different species and experimental designs. We show that parameter choice has a significant impact on the quality of the resulting co-expression network, and that the specific influence of each parameter is highly dataset-dependent.

An alternative approach to selecting a single optimized network is to combine information across multiple network instantiations, in order to generate a robust "consensus network". For example, in the field of phylogenetics, an edge-counting and thresholding approach was used to determine optimal phylogenetic tree structure from multiple probabilistically-generated tree instantiations (Holland and Moulton 2003). To date, however, no such methodology has been developed for use with co-expression networks. Here, we describe a "consensus co-expression network" approach which combines information across multiple alternative co-expression network instantiations, followed by gene/edge pruning to remove low-confidence nodes and edges from the network.

Finally, the co-expression network construction techniques developed in the first part of this chapter are used to construct robust host co-expression networks for two dual RNA-seq datasets: *H. sapiens* infected with *L. major* (HsLm) (Dillon, Okrah, et al. 2015) and *H. sapiens* infected with *T. cruzi* (HsTc) (Y. Li et al. 2016). The networks and corresponding co-expression modules are characterized and modules related to infection response are highlighted. On the parasite side, consensus co-expression networks are constructed either using both intracellular and extracellular stage samples from a single experiment (TcAll) (Y. Li et al. 2016), or by combining intracellular and extracellular samples across several experiments (LmAll) (Dillon, Okrah, et al. 2015; Fernandes et al. 2016; Dillon, Suresh, et al. 2015; Inbar et al. 2017). While including extracellular stage samples during network construction limits our ability to investigate infection-specific network properties, by utilizing as many different samples as possible during the construction of each parasite network, we are better able to capture various stage-specific

co-regulatory relationships that exist in the data, providing a rich resource for future regulatory element detection efforts.

Co-expression network construction, module detection, and functional enrichment analysis

To begin, a generalized pipeline for the construction of co-expression networks from RNA-Seq data was developed. The pipeline is based on the WGCNA framework, and provides a way to quickly generate co-expression networks for a wide range of datasets, while still providing flexibility in terms of the specific methods and parameters used for network construction. This development of this pipeline was crucial for later steps where thousands of different co-expression networks were generated, each with different parameterizations. While there are many different possible transformations and algorithmic choices that can be applied along the way, the basic process starts by converting an $n \times p$ RNA-Seq count matrix into an $n \times n$ symmetric similarity matrix. Next, the similarity matrix is converted into an adjacency matrix, with each cell in the matrix representing the connectivity or magnitude of co-expression between two genes. Finally, hierarchical clustering and a dynamic branch cut approach are applied to partition the resulting co-expression network into multiple co-expression network modules, each representing a set of highly co-expressed genes (**figure 8).**

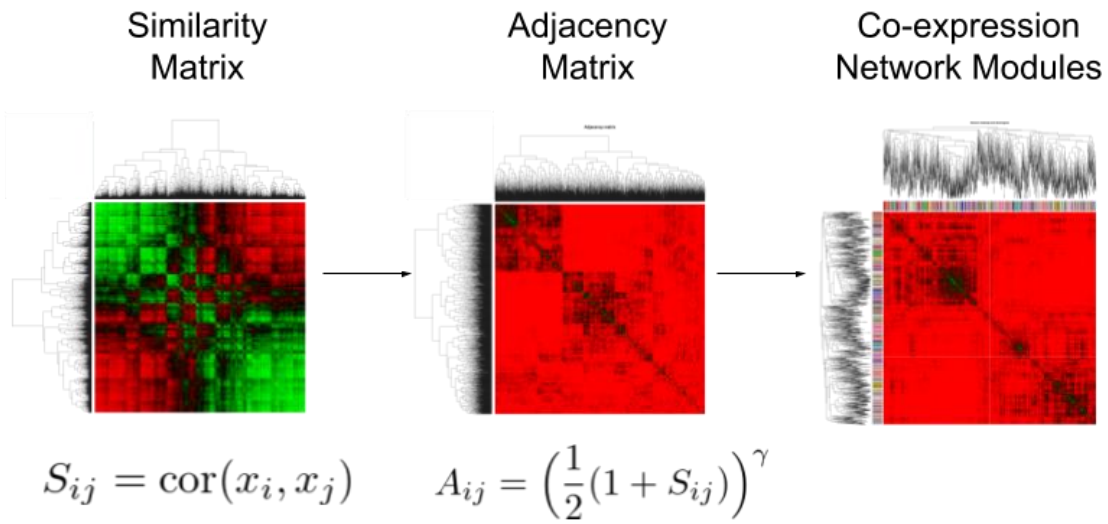

figure 8: Overview of key co-expression network construction and module detection steps. Biclustering heatmap representations of several of the key stages during co-expression network construction and module detection are shown. On the left, the similarity matrix shows a typical distribution of raw Pearson correlation scores between genes from an RNA-Seq dataset. The values range from -1 (red) to +1 (green). Low correlation scores close to zero appear black. The presence of the large amount of bright red and green in the matrix is indicative of the significant number of spurious correlations present in low sample-size RNA-seq experiments. The middle plot shows the shifted and power-transformed adjacency matrix with values ranging from 0 (red) to +1 (green). The sparseness induced by the power transformation is clear by the relatively low ratio of green to red in the plot. The rightmost plot shows an example hierarchical clustering dendrogram and module partitioning that is generated following network construction. Each color stripe along the left and top axes represents a separate co-expression module, typically containing between 10-100 genes.

Once the co-expression network has been divided into a set of modules, each module is assessed for functional enrichment (also referred to as “gene set enrichment”) using several different biological annotation resources. Using the goseq package for R (Young et al. 2010), co-expression modules for all datasets were tested for enrichment of Gene Ontology (GO) terms (Ashburner et al. 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000). Depending on the specific organism being analyzed, we additionally measured enrichment of ConsensusPathDB (CPDB) pathways (Kamburov et al. 2013), known transcription factor (TF) regulated genes (Marbach et al. 2016a), LeishCyc pathways (Doyle et al. 2009), and/or genes

predicted to encode secreted proteins (Petersen et al. 2011), transmembrane domains (Sonnhammer, von Heijne, and Krogh 1998) or GPI-anchors (Pierleoni, Martelli, and Casadio 2008; Myler and Fasel 2008). For the co-expression networks generated from infection datasets, this was useful, for example, for elucidating co-expressed modules relating to infection or immune response. Further, for the sparsely annotated trypanosomatid genomes, this functional enrichment analysis step is particularly useful since it can provide hints as to the possible roles for some of the numerous unannotated genes present in the genome. Finally, the assessment of functional enrichment across entire networks provided us with a potential criteria to begin to evaluate the quality of the networks, in turn leading the discovery of some unexpected pitfalls relating to commonly used methodologies, as described below.

During the development of this pipeline, we tested a wide range of data transformations, similarity metrics, gene filtering approaches, and module detection methods, hereafter collectively referred to as "network construction parameter choices". This led to the realization that the co-expression network construction process is highly sensitive to some parameter decisions, while at the same time being largely unaffected by others **(figure 9)**.

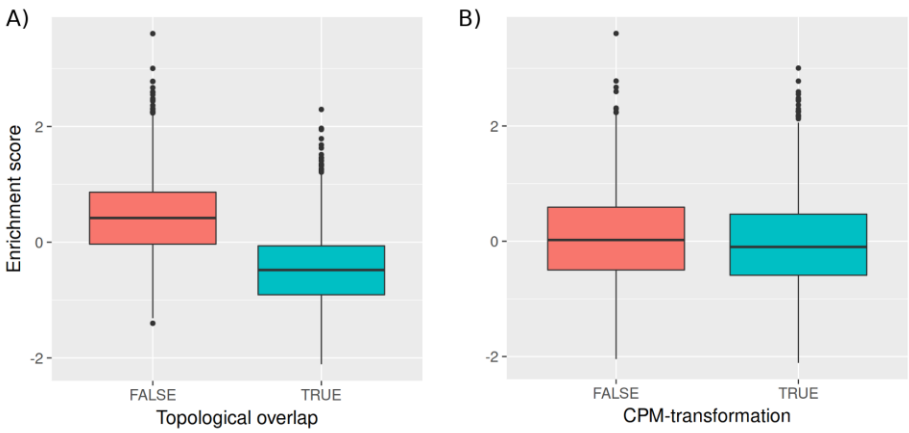

figure 9: Influence of co-expression network parameter choice on network functional enrichment. For a given dataset, there are many different parameter selections which must be made, each of which may play a role in determining the quality of the resulting network. The degree to which a parameter influences the network depends both on which parameter is being modified, whether that parameter interacts with other network parameters, and on the specific dataset for which the network is being constructed. Above, two examples of the variability of influence are shown for the same dataset, *Homo sapiens* infected with *Leishmania major* (HsLm). (A) The effect of the topological overlap transformation on overall functional enrichment of the network is shown. When the topological overlap transformation is applied to the adjacency matrix (right, blue), the resulting networks tend to have significantly lower levels of functional enrichment compared to when the step is skipped (left, red). On the other hand, for the same dataset, the counts-per-million transformation (B) has a much smaller impact on the overall network enrichment. Each individual box plot shown represents the distribution of network functional enrichment scores for 2,688 possible alternative network parameterizations, when the specific parameter choice displayed is fixed.

Moreover, it was also discovered that, at least for the several datasets explored, some of the commonly used parameter choices, such as those suggested in the WGCNA tutorial, could have a very detrimental effect on the quality of the resulting co-expression network. In particular, the topological overlap matrix (TOM) transformation that is often used prior to module detection, and the hybrid branch cut method used to partition the hierarchical clustering dendrogram into individual modules, both tended to produce co-expression networks with lower levels of functional enrichment (**figure 9**).

These observations prompted us to take an in-depth look at the role of various parameter choices on the quality of the co-expression network generated, and to devise approaches for choosing between, or combining information across, multiple possible

network parameterizations, in order to arrive at an "optimal" co-expression network that captures as much of the underlying co-expression structure as possible.

A co-expression network scoring method based on functional enrichment

Based on the assumption that co-expression networks are determined, at least in part, by the underlying cellular gene regulatory networks (GRNs), and the fact that we know genes involved in the same regulatory pathways tend to be functionally related (D'haeseleer, Liang, and Somogyi 2000; Michalak 2008), total network functional enrichment was used as a measure of co-expression network quality. Previous research on co-expression network optimization using WGCNA has focused selecting an optimal adjacency power parameter (**table 2**) based on the scale-free network criterion (Langfelder and Horvath 2008). While the degree distributions of many real-world biological networks do indeed follow a power-law (scale-free) distribution, this has not been directly tested in the context of co-expression networks. Moreover, such optimization approaches tend to focus on a single parameter of the network construction, such as the adjacency matrix power transformation exponent, without considering the role that decisions at other steps in the network construction process have on the resulting network.

For each host/parasite dataset, we constructed multiple alternative co-expression networks and network partitionings, using a range of data transformation, network construction, and module detection parameters (log- transformation, batch adjustment, similarity metric, adjacency power, etc.) for both host and parasite (**table 2**). For each network realization, we evaluated its biological meaningfulness by counting the number

of enriched GO terms and KEGG pathways found across all network modules, and assigned a score based on the total number of unique enriched annotations. The utility of this approach is further supported by the observation that, for the human datasets, networks which are strongly enriched for GO and KEGG pathways also tend to be highly enriched for known TF-regulated genes (**figure 10**).

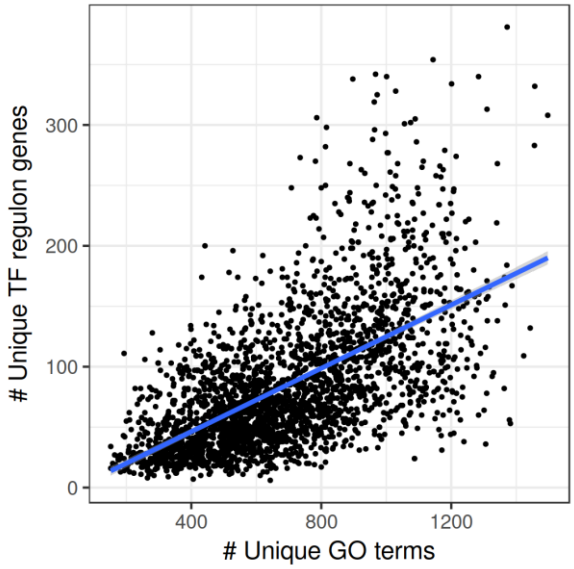

figure 10: Correlation between functional annotation enrichment and TF-regulon enrichment. Number of unique enriched GO terms is plotted against the number of unique Marbach *et al*. (2015) known targets of transcription factors, for 1,344 generated HsLm networks. The moderate correlation (r = 0.62) lends supports to the hypothesis that genes which are functionally related, are often co-regulated.

Parameter	Values considered	Description
Counts-per-million (CPM) transformation	True, False	Size factor adjustment commonly used in differential expression analysis; prevents sample size from influencing the contribution of that sample to the resulting co-expression assessment.
Log2 transformation	True, False	Transforms the overall distribution of the count data; limits the influence of outlier genes with very high expression levels during co-expression determination.

Quantile normalization	True, False	Normalizes distributions across samples; often used in differential expression analysis.
Batch adjustment	Limma, ComBat, None	Attempts to account for and subtract out the effects of experimental batch effects, when present; can lead to more targeted networks (e.g. infection-specific) but results in a loss of information.
Similarity measure	Pearson correlation, Spearman correlation, biweight midcorrelation, cor-dist	Metric used to assess co-expression relationship between genes; metrics vary in their ability to detect nonlinear relationships, avoid the influence of outliers, and account for differences of scale or magnitude.
Adjacency power	1-14	The primary parameter which is optimized in the traditional WGCNA pipeline; determines how high a similarity score has to be to lead to a large edge weight.

table 2 Co-expression network construction parameters tested

For each dataset, co-expression networks were constructed for all possible permutations of the parameter values included in this table. Parameter are listed in the order in which they are applied during the network construction process. The first three options (CPM, Log2, and quantile normalization) are data transformation steps which are applied at the early stages of data preparation. These are commonly used transformations during differential expression and co-expression network analysis of RNA-Seq data. Batch adjustment is a technique for removing variance from the data associated with variables that are not of direct interest, for example, RNA-Seq library preparation date. The methods considered include constructing a linear model using batch alone and then subtracting it out (limma), ComBat batch adjustment, and no batch adjustment. The similarity measures considered include commonly applied correlation measures such as Pearson correlation and Spearman correlation, a robust correlation measure suggested for use in the WGCNA framework (biweight midcorrelation), and a weighted combination of euclidean distance and pearson correlation used in our lab (cor-dist). The adjacency power parameter refers to the exponent of power transformation applied to the similarity matrix before it is converted to a distance matrix and clustering is performed.

When evaluating the effect of parameter choice on co-expression network structure and functional enrichment, it was observed that even for those networks generated using better-performing similarity metrics, network modules would sometimes appear to contain sub-groups of genes with similar overall expression behavior, but operating at very different levels of expression. While using Euclidean distance directly does indeed eliminate this problem, it was found to perform quite poorly overall compared to other similar metrics. In order to address this weakness, a simple similarity

metric was devised by applying a euclidean distance-based penalty to the Pearson correlation metric ("cor-dist"). In testing, cor-dist was found to outperform all other similarity metrics evaluated for nearly half of the datasets analyzed (**figure 11, Table S4,**).

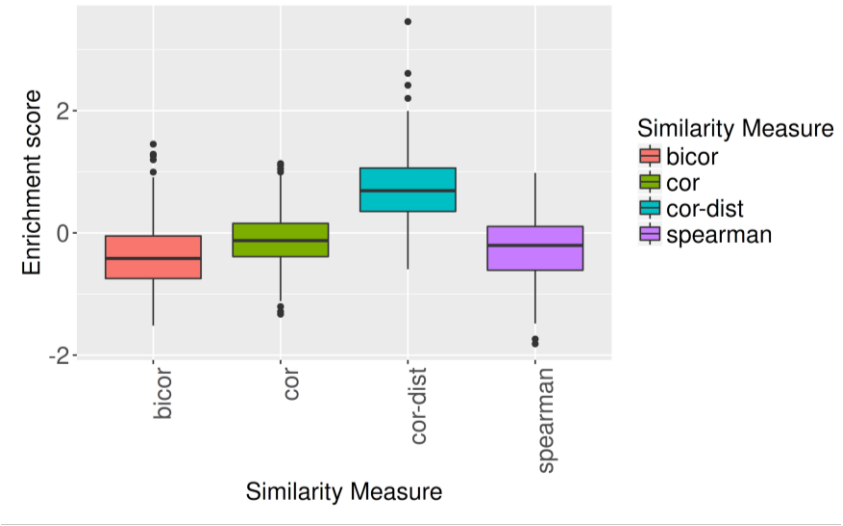

figure 11: Performance of a novel similarity metric (cor-dist).

Of the thirteen different co-expression networks optimized using the functional enrichment approach, six of optimal networks made use of the novel cor-dist similarity measure. One example of a case where the new metric performed well is for the HsLm network. Above, functional enrichment scores are shown for the 1,344 constructed HsLm networks, grouped by the similarity measure used.

Altogether, a total of 13,888 co-expression networks were constructed across seven different experiments, including 445 unique RNA-Seq samples and six different organisms. From these, 13 optimal co-expression networks were inferred using the functional enrichment criterion (**Tables 3-4, Figures S5-6**). Not considering batch adjustment, the three best performing parameter combinations across all 13 datasets were (1) log2-adjustment only, (2) log2- and quantile normalization, and (3) no adjustments, each of which was found to be optimal for three of the thirteen networks. Interestingly, the main parameter which WGCNA attempts to optimize (the adjacency power), was

generally not found to have a significant or reliable influence on the overall quality of the network **(figure 12)**.

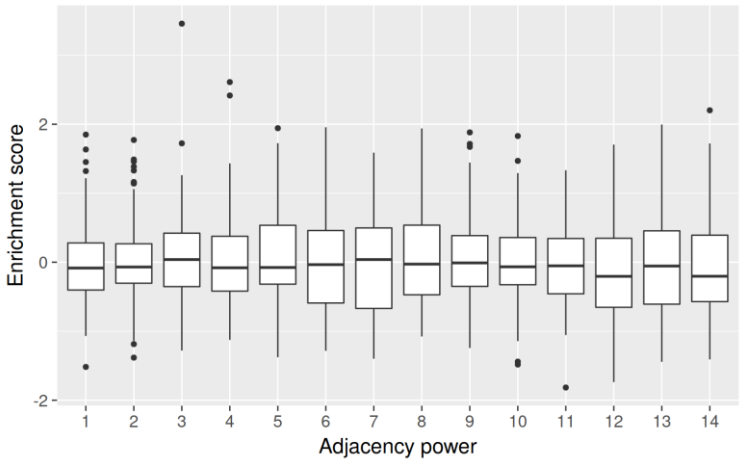

figure 12: Adjacency power not a major determinant of co-expression network quality.

Functional enrichment scores are shown for 1,344 HsLm networks (96 alternative network parameterizations for each specific adjacency power). In general, no consistent influence of adjacency power on the resulting network quality was observed, with high- and low-scoring networks spread across a range of adjacency powers.

Dataset	Network abbreviation	Experiment type
L. major infecting *H. sapiens*	LmHs	Infection dataset (parasite)
L. major infecting *M. musculus*	LmMm	Infection dataset (parasite)
L. major all samples	LmAll	Combined dataset
T. cruzi infecting *H. sapiens*	TcHs	Infection dataset (parasite)
T. cruzi all samples	TcAll	Combined dataset
modENCODE Fly	ModFly	Developmental
modENCODE Worm	ModWorm	Developmental
M. musculus infected with *L. major*	MmLm	Infection dataset (host)
H. sapiens infected with *L. major*	HsLm	Infection dataset (host)
H. sapiens infected with *T. cruzi*	HsTc	Infection dataset (host)
H. sapiens infected with *L. braziliensis*	HsLb	Infection dataset (host)
H. sapiens all samples	HsAll	Combined dataset

Illumina Human BodyMap 2.0	BodyMap	Multi-tissue

table 3: Optimized networks inferred using functional enrichment.

A total of thirteen optimal networks were inferred using the functional enrichment approach. Six of the networks (HsLm, HsTc, MmLm, LmHs, TcHs, and LmMm) were constructed using either host or parasite reads only from dual transcriptomics infection datasets and three of the datasets ("combined") were generated by combining samples from two or more separate experiments. All datasets used are publically available and sources for each are listed in table S3.

One important finding of this study is that, while there were a few specific instances of parameter choices that behaved similarly across the various datasets, in general, each dataset ultimately required different combinations of parameters to arrive at an optimal network (**Figures S2-3**). As such, whenever possible, alternative network instantiations should be constructed and evaluated, using a criterion such as the functional enrichment score described in this work, for each new dataset analyzed. That said, there are two generalizations that did emerge from the comparative analysis. First, of the four similarity metrics evaluated, Pearson correlation and cor-dist performed significantly better than the other metrics for nearly all networks. Spearman correlation typically performed poorly, except when applied to the "Combined" datasets that included samples from multiple experiments. Biweight midcorrelation, the default metric suggested in the WGCNA user's guide, was found to perform well for only one of the 13 networks evaluated. Second, when experimental batch information is available, adjusting for such effects prior to network construction typically results in a higher quality co-expression network. This is consistent with recent research by Leek *et al*. (Parsana et al. 2017) on the influence of batch on co-expression network construction.

A limitation to the above approach for co-expression network optimization is its inherent dependency on the quality of available annotations. For relatively well-annotated

model organisms such as human and mouse, this may be less of an issue, however for the sparsely annotated parasite genomes, the optimized networks may appear to fit the known annotations very well, but in actuality do a poor job capturing the import co-expression relationships in the data. This led us to consider alternative approaches that could be used to construct optimal co-expression networks in a less biased manner. In the next section, we describe such an approach which combines information across multiple alternative co-expression network parameterizations, in order to generate a robust network.

Dataset	Num samples	Species	Num genes	Num modules	GO	KEGG	CPDB	Marbach
LmHs	19	*L. major*	3605	177	318	17	-	-
LmMm	13	*L. major*	3183	167	169	9	-	-
LmAll	73	*L. major*	8744	419	593	68	-	-
TcHs	19	*T. cruzi*	5879	294	225	25	-	-
TcAll	26	*T. cruzi*	7465	369	366	28	-	-
ModFly	132	*D. melanogaster*	12599	611	6001	144	-	-
ModWorm	44	*C. elegans*	13032	654	108	60	-	-
MmLm	26	*M. musculus*	16206	782	1800	158	500	-
HsLm	54	*H. sapiens*	17309	749	2688	129	718	502
HsTc	29	*H. sapiens*	14795	725	1927	119	639	120
HsLb	35	*H. sapiens*	12958	560	4089	177	1162	762
HsAll	129	*H. sapiens*	6099	264	2014	93	503	379
BodyMap	15	*H. sapiens*	12537	636	2552	117	633	276

table 4: Summary of co-expression networks evaluated using functional enrichment.

Thirteen co-expression networks comprising six different species and a total of 445 unique RNA-Seq samples across various experimental conditions including infection studies and multi-tissue experiments were optimized. This table describes the basic properties of the datasets including the number of samples, and number of genes after filtering, as well as the total number of enriched GO, KEGG, etc. functional annotations associated with the optimized networks for each dataset. CPDB annotations are only available for human and mouse, while Marbach TF targets are only included for human datasets.

Robust network generation using consensus approaches

While the above sections focus on the construction of an individual optimized co-expression network, an alternative approach is to instead construct multiple co-expression networks, each using a different set of parameters, and then combine information across those networks in order to arrive at a single "consensus" network. This intuition here is that, while any individual network parameterization is likely to accurately estimate the magnitude of co-expression between some pairs of genes, and misjudge the relationship between others, genes which are in fact strongly co-expressed should have stronger edge weights on average, compared with those genes which are not strongly co-expressed. By summing the edge weights across all possible network instantiations, each individual co-expression network is thus "voting" for the gene-pairs which it believes are actually co-expressed. Such techniques have been used successfully to combine information across multiple alternative clusterings (Yu, Wong, and Wang 2007). Here we apply a similar approach at the level of the network.

Consensus co-expression networks were constructed for each of six datasets: HsLm, HsLmUI, HsTc, HsTcUI, LmHs, and TcHs (**Tables 6-7**), and co-expression modules were detected for each network. In order to evaluate the performance of the consensus network approach, compared with the optimization and selection of a single network parameterization, functional enrichment was assessed for each of the consensus networks. Next, plots were generated showing the distributions of enrichment scores for the individually-constructed networks, and a vertical line was drawn indicating the level of enrichment for the corresponding consensus network. As expected, the level of functional enrichment in the consensus network tended to be much higher than the

average individual network, and in many cases, was higher than the best-performing individual networks (**Figures 14, S5-7; table 5**).

	HsLm	HsLmUI	HsTc	HsTcUI	LmAll	TcAll
GO Terms	0.93	0.89	1	1	0.82	0.97
KEGG Pathways	0.54	0.35	0.67	1	0.84	0.79
CPDB Pathways	0.87	0.94	0.94	1	-	-
Marbach TF Genes	1	1	0.99	0.96	-	-

table 5: Comparison of consensus and individual co-expression network enrichment.

For each unfiltered consensus network, the total functional enrichment score (sum of unique enriched annotations) for each available annotation type is compared to the distribution unfiltered individual network scores. Above, the quantile scores are shown for each network and annotation type indicating the fraction of individual network scores for which the consensus network performed as well or better; A score of 0.9 means that the consensus network had a level of enrichment equal to or above 90% of the individual networks for that annotation type, while a score of 1 indicates that the consensus network performed better than all individual networks. In nearly all cases, the consensus network performed at least as well as the average individual network, and very often had a higher level of functional enrichment than any of the individual networks.

Next, in order to further improve the quality of each of the resulting consensus networks, a filtering scheme was applied in order to remove low-confidence genes and edges. By selectively removing low-confidence genes from the network (those for which a strong co-expression relationship could not be reliably detected with any other genes in the network), we are able to arrive at a robust “core” consensus co-expression network that better captures the true co-expression relationships in the data (**Figures S5-7**). This step is especially beneficial for downstream analyses, such as the differential network analyses described in chapter 4, which build on top of and are highly dependent upon, the quality of the input co-expression networks.

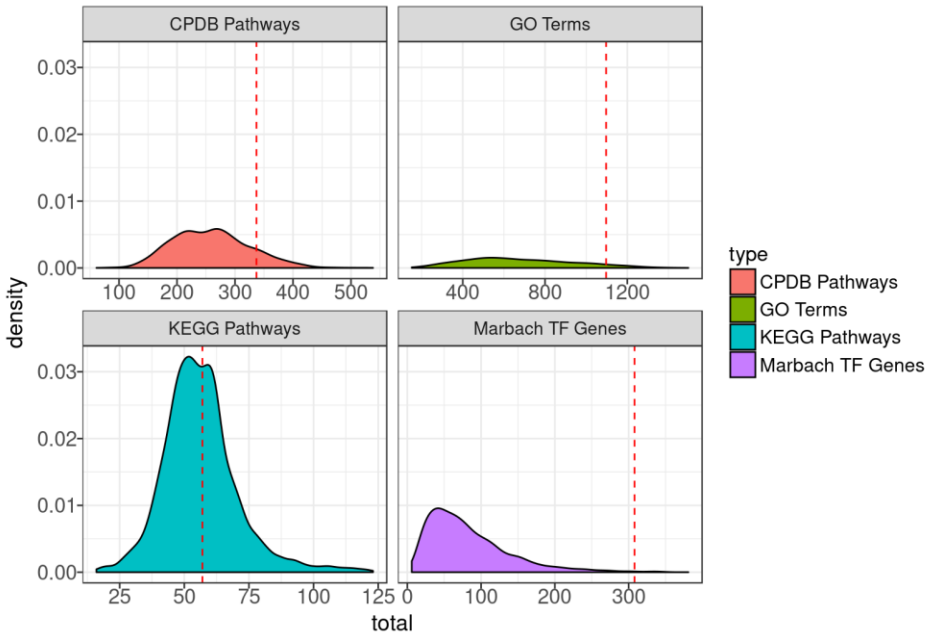

figure 13: Comparison of functional enrichment for consensus and individual co-expression networks.

Kernel density plots depicting the distribution of functional enrichment scores (total number of unique annotations) for each of the 1,344 HsLm co-expression networks are shown, separated by annotation type. For each annotation source, the level of functional enrichment of that same type for the unfiltered consensus co-expression network is shown as a red dashed line. For all annotation types, the unfiltered consensus network had a higher level of enrichment than most of the individual networks. In the case of GO enrichment, the unfiltered consensus network had a higher level enrichment than any of the individual networks (1,320 and 1,669 unique GO terms for the individual networks and consensus network, respectively). Similar plots were constructed for each of the five other consensus networks constructed and the trend was similar.

Host and parasite consensus co-expression networks

Trypanosoma cruzi infecting Human

A *T. cruzi* consensus co-expression network (TcAll) was constructed and optimized as described above. A total of 26 RNA-Seq samples were used to construct the network, including both extracellular stage parasites (epimastigotes and metacyclic trypomastigotes) and intracellular trypomastigotes (4, 6, 12, 20, 24, 48, 72 hpi). A minimum max consensus edge weight cutoff value of 753.9 (22nd quantile) was used to remove non-robust genes and edges, resulting in a final network with 7,666 genes and 143 modules (**Figures 15, S4, table 6**).

Out of the 143 detected co-expression modules, 26.6% (38/143) were found to be enriched for one or more functional annotation. Enriched annotations included metabolic pathways, cell adhesion and metalloendopeptidase activity, and genes encoding putative secreted proteins and GPI-anchored proteins.

Network	Samples	Genes	Modules	GO terms	KEGG pathways	Secreted proteins	Transmembrane proteins (modules)	GPI-anchored proteins (modules)
TcAll	26	7,666	143	230	14	15	1	13
LmAll	73	6,020	115	423	22	2	4	5

table 6: Summary of parasite consensus co-expression networks.

Summary of the two parasite consensus co-expression networks including the number of samples used to construct the network, the number of genes and modules detected in the final network, the number of unique enriched GO and KEGG annotations, and the number of co-expression *modules* enriched for genes encoding putative secreted proteins, transmembrane proteins, or GPI-anchored proteins.

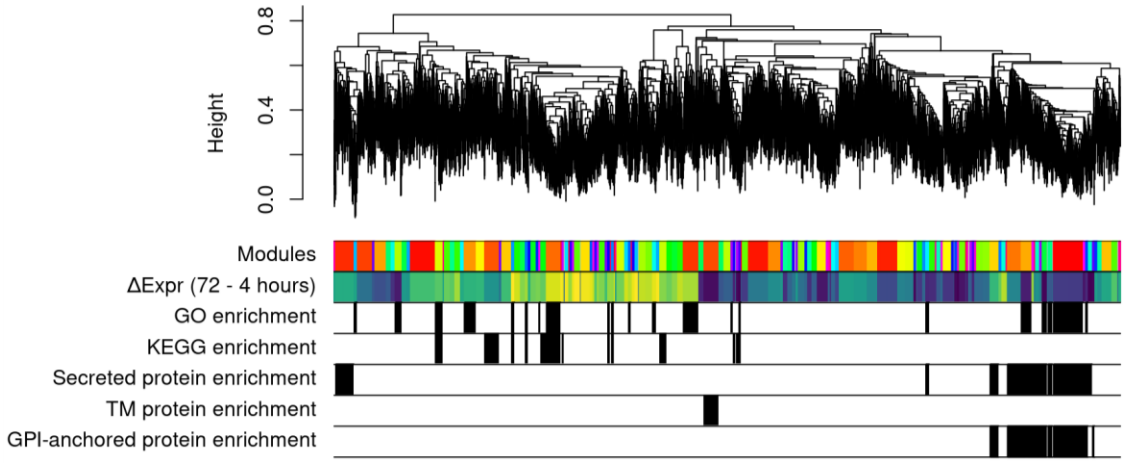

figure 14: *T. cruzi* consensus co-expression network dendrogram with phenotypic correlates.

The *T. cruzi* consensus co-expression network (TcAll) is depicted using a hierarchical clustering dendrogram (**top**). Each leaf in the dendrogram represents a single gene, with gene proximity in the dendrogram indicating similarity in expression profiles. The height of the dendrogram shows the levels of correlation between expression profiles, with genes near the bottom of the dendrogram (those with a lower height) being more highly correlated. **(Bottom)** Gene module assignments and other phenotypic features are depicted as rows of colors. (a) Module assignments, (b) Module net change in average log2-CPM expression between 4 and 72 hours (yellow = up-regulated, purple = down-regulated), and statistical enrichment of: (c) GO terms, (d) KEGG pathways, (e) genes encoding known or predicted secreted proteins, (f) genes encoding known or predicted transmembrane proteins, and (g) genes encoding known or

predicted GPI-anchored proteins. The over-representation of green and yellow bands on the left side of the net expression bar (ΔExpr (72 - 4 hours)) indicates the presence of a large number of co-expression modules which are increasing in expression between 4 and 72 hours. Similarly, the over-representation of purple and blue bands on the right-side of the plot corresponds to modules which are decreasing in expression across time. The presence of many modules encoding secreted and GPI-anchored proteins on the right side suggests a trend towards decreased expression of these protein products as the amastigotes acclimate to the intracellular environment.

Leishmania major infecting Human

An *L. major* consensus co-expression network (LmAll) was constructed and optimized as described above. A total of 73 RNA-Seq samples from three separate experiments including extracellular procyclic promastigotes and metacyclic promastigotes, and intracellular metacyclic promastigotes and amastigote, across three different hosts (human, mouse, and sandfly) was used (Dillon, Okrah, et al. 2015; Fernandes et al. 2016; Inbar et al. 2017). A minimum max consensus edge weight cutoff value of 389 (35th quantile) was used, resulting in a network with 6,020 genes and 115 modules (**Figures S4, S8, table 6**).

Network	Samples	Genes	Modules	GO terms	KEGG pathways	CPDB pathways	TF-regulon genes
HsLm	19	10,385	219	997	45	248	87
HsLmUI	24	7,846	128	1,138	33	319	231
HsTc	19	10,208	157	1,486	83	1,469	230
HsTcUI	10	9,460	160	938	25	927	139

table 7: Summary of host consensus co-expression networks.

Summary of the four host consensus co-expression networks including the number of samples used to construct the network, the number of genes and modules detected in the final network, the number of unique enriched GO, KEGG, and CPDB annotations, and the number of enriched *Marbach et al.* TF regulon genes. The two uninfected human networks (HsLmUI and HsTcUI) are not discussed further in this chapter, but are used in the differential network analyses described in chapter 4.

Human infected with *Leishmania major*

A human consensus co-expression network (HsLm) was constructed using 19 *L. major*-infected macrophage RNA-Seq samples (Fernandes et al. 2016), collected at four

different time points during the infection process: 4 hours, 24 hours, 48 hours, and 72 hours. A minimum max consensus edge weight cutoff value of 739 (40th quantile) was used, resulting in a network with 10,385 genes and 219 modules (**Figure S4, S9**, **table 7**). Of the 219 detected co-expression modules, a total of 52 (23.7%) were found to be functionally enriched for at least one known annotation, spanning a range of functions.

Two of co-expression modules were found specifically to be enriched with the KEGG "Leishmaniasis" pathway (adjusted *P*-values 0.002 and 0.016), and included genes encoding MHC class II complex components and IL-12B. Another co-expression containing 105 genes was found to be enriched for genes involved in the NF-kappa B and IL-12 signaling pathways **(figure 15)**, both of which are known to be targeted by *L. major* during macrophage infection (Nylén and Gautam 2010). IL-12, for example, stimulates the production of IFN-γ, which is required for macrophage activation and parasite clearance.

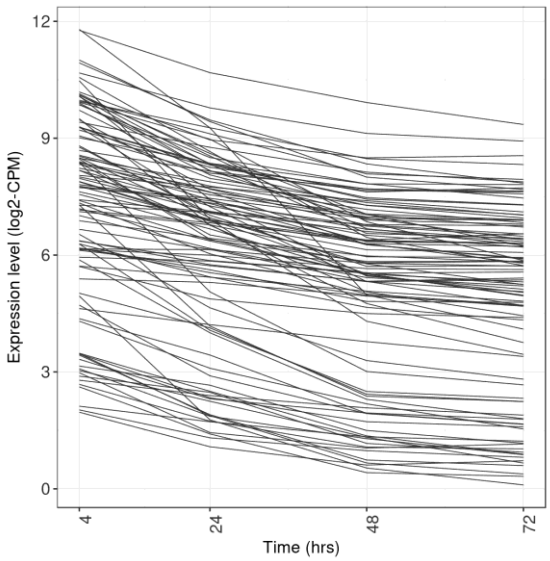

figure 15: NF-kappa B and IL-12 mediated signalling down-regulated during *L. major* infection. Log2-CPM gene expression is shown as a function of time during infection by *L. major* for an HsLm co-expression module found to be enriched for genes associated with the KEGG "NF-kappa B signaling

pathway" (adj. *P*-value ~0.000) and the CPDB pathway "IL12-mediated signaling events" (adj. *P*-value 0.018), two important pathways that *L. major* is known to down-regulated in order to survive in the host macrophage environment. Each time point represents the average expression level across 4-5 replicates.

Human infected with *Trypanosoma cruzi*

A human consensus co-expression network (HsTc) was constructed using 19 *T. cruzi*-infected fibroblast RNA-Seq samples (Y. Li et al. 2016), collected at seven different time points during the infection process: 4, 6, 12, 20, 24, 48, and 72 hours. A minimum max consensus edge weight cutoff value of 751 (31st quantile) was used, resulting in a network with 10,208 genes and 157 modules (**Figure S4, S10**, **table 7**). Of the 157 detected co-expression modules, a total of 55 (35.0%) were found to be functionally enriched for at least one known annotation.

Conclusion

The first part of this chapter describes the development of a basic pipeline for co-expression network construction, module detection, and functional enrichment based on the popular weighted gene co-expression analysis (WGCNA) (Langfelder and Horvath 2008) framework. We explored the impact of various data transformations, similarity measures, and algorithmic parameters on the resulting network and showed that even for a single method such as WGCNA, there are wide range of possible networks that can be inferred. Moreover, we found that many commonly overlooked steps during co-expression network construction, such as early data transformation and normalization decisions, can have a significant impact on the resulting network quality. Meanwhile, other parameters which have received more attention in the literature (such as the similarity matrix power exponent which is the focus of parameter optimization for WGCNA) may in turn have a smaller impact on the inferred network. Much of the later

work described in this thesis built upon the basic methodologies described in the first part of this chapter.

The second part of this chapter detailed two different approaches designed to help construct an optimal co-expression network for a given dataset. In the first approach, a method for scoring co-expression networks using external information in the form of functional annotations was described, enabling us to rank alternative network parameterizations created from the same dataset. The method was used to evaluate the influence of various data transformation and algorithmic parameter choices on the quality of the resulting networks, and demonstrated the highly dataset-dependent nature of various parameters. A key limitation to this approach, however, is its dependency on the quality and availability of annotation sources for the dataset. This is one of the key motivations for exploring alternative network optimization approaches that are independent of functional enrichment.

While the first optimization approach was geared towards selecting a single "best" network parameterization, the second approach instead combined information across multiple network parameterizations in order to construct an ensemble "consensus network". The consensus network construction approach described is entirely independent of any functional annotations, and as such, can be readily applied to any dataset, regardless of the availability of annotations. Next, we discussed an approach for pruning low-confidence genes and edges from a co-expression network, leading to a smaller, more robust core network. While this step could also be performed in an annotation-independent manner (e.g. by arbitrarily selecting a cutoff value to remove some desired number of genes from the network), for the purposes of this thesis, we once

more took advantage of the existing annotation resources in order to guide the selection of a reasonable network filter cutoff value.

Finally, we described four individual consensus co-expression networks (two host networks and two parasite networks) constructed using the co-expression network techniques described in this chapter, and demonstrated their use for elucidating import processes that take place during infection.

In the next chapter, we continue to build on the techniques studied in this chapter, developing a differential co-expression network approach for investigating transcriptional changes that occur during infection. We also compare consensus co-expression networks for human samples infected with two different parasites: *L. major* and *T. cruzi*, and look for similarities and differences in gene co-regulation between the two different infections.

Methods

Data Acquisition

Of the seven datasets investigated in this work (**Table S3**), the host-parasite dual transcriptomics datasets were generated either in our lab, or in collaboration Dr. Mosser's laboratory (Christensen et al. 2016). RNA-Seq data is available for each of these datasets via the Sequence Read Archive (SRA) (Leinonen et al. 2011)). SRA Accession numbers can be found in the corresponding references listed in Table M.1. For the three remaining three data sets (Illumina Human BodyMap 2.0, modENCODE Fly, and modENCODE Worm), pre-generated count tables were downloaded via ReCount online resource of pre-processed RNA-Seq datasets (Frazee, Langmead, and Leek 2011).

Data Preparation

For the host-parasite dual RNA-Seq data generated in our lab, RNA-Seq read quality was assessed using FastQC v0.11.4 (Andrews 2010) and reads were trimmed with Trimmomatic v0.32 (Bolger, Lohse, and Usadel 2014) using the parameters "`PE -phred33 ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:20 TRAILING:20 MINLEN:36`". Paired-end reads found to be missing a mate pair were excluded from the dataset. Reference genomes were downloaded from ENSEMBL (*H. sapiens*, *M. musculus*) and TriTrypDB (*T. cruzi*, *L. major*), and samples were mapped to their corresponding host and parasite reference genomes **(Table S2)**. In cases where samples included mRNA reads from both parasite and host (for example, intracellular *L. major* amastigote samples), reads were mapped to *both* host and parasite reference genomes. Because of the distant evolutionary relationship between the organisms, very few reads mapped ambiguously to both genomes.

Sample reads were then mapped to the human GRCh38.83 reference genome (Schneider et al. 2017) using Tophat 2 v2.1.0 (D. Kim et al. 2013) with the parameters "`--max-multihits 1 --library-type fr-unstranded --GTF Homo_sapiens.GRCh38.83.gtf`". Infection stage samples were additionally mapped to the TriTrypDB version 27 *Leishmania major* reference genome (Aslett et al. 2010) with the parameters "`--max-multihits 1 --library-type fr-unstranded`". HTSeq v0.6.0 (Anders, Pyl, and Huber 2014) was used to generate count tables for each mapped sample. For reads mapped to host, the parameters "`--format bam --stranded no`" were used. For reads mapped to parasite, the parameters "`--format bam --stranded no --type gene --idattr ID`" were used.

Co-expression network construction and module detection

A modified version of the Weighted Gene Co-expression Network Analysis (WGCNA) framework (Langfelder and Horvath 2008) was used to construct co-expression networks and detect network modules. In order to investigate the influence of various data transformations, network construction parameters, and module detection parameters on the resulting co-expression network, numerous alternative network instantiations were created for each dataset analyzed. The influence of network parameter choice on the resulting network quality, and the approaches used to attempt to select an optimal network parameterization, or to combine information across multiple parameterizations to arrive at a robust core co-expression network are described in later sections. Here, the basic approach used to construct a single co-expression network is described.

Starting with the raw count table generated by HTSeq, several possible data transformations were applied: counts-per-million (CPM) normalization, log2 transformation, and quantile normalization (Bolstad et al. 2005). Genes with zero variance in expression were removed, and a possible step of batch adjustment applied. Two batch adjustment approaches were utilized: ComBat batch adjustment (Leek et al. 2012) and subtraction of the component of expression which can be explained using a statistical model containing only batch information (Leek et al. 2010). Next, a similarity matrix was constructed using one of several possible similarity metrics including: Pearson correlation, Spearman correlation, biweight midcorrelation (bicor), or a novel metric developed in this work, referred to as “cor-dist”. Euclidean distance was also considered, but was generally found to produce poorer quality networks, and was discarded.

The cor-dist similarity metric developed in this worked was motivated by an observation regarding the types of genes which get grouped together when using metrics such as Pearson correlation. Pearson correlation does an excellent job detecting linear relationships between variables, but is unaware of "scale". That is, two genes which follow a similar expression pattern across samples, but different in absolute levels of expression by orders of magnitude will still be assigned a similarity score. The result is that some co-expression network modules appear to include two or more sub-modules of genes with similar expression behavior, but expressed at very different levels. In order to address this issue, a simple metric was developed which applies a euclidean distance based penalty to the Pearson correlation score between two genes (Inbar et al. 2017) (**Formula 1**).

$$S = sign(cor(X)) \times \frac{|cor(X)| + (1 - \frac{log(dist(X)+1)}{max(log(dist(X)+1))})}{2}$$

Formula 1: The cor-dist similarity metric.
In order differentiate between genes following similar trajectories of expression across developmental stage, but at very different magnitudes, a simple similarity metric was devised. The new metric (cor-dist) applies euclidean distance-based penalty to the Pearson correlation score between two variables.

Next, the similarity matrix was raised to a power ranging from 1-14 to reduce the number of spurious correlations, and shifted to the range [0, 1] by adding one to the values and dividing by two. An alternative approach is to simply take the absolute value of the similarity scores, however, this results in co-expression modules containing genes whose expression profiles are both positively and negatively correlated with one another, and may lead to sub-optimal networks (Mason et al. 2009). One key difference between the approach used here and that commonly applied with the WGCNA framework is that the topological overlap matrix (TOM) transformation step was not used. The TOM

transformation was initially explored, but was found to result in less functionally-enriched co-expression networks across multiple datasets (**figure 9A**).

Next, the similarity matrix was then converted into a distance matrix by subtracting from one, and hierarchical clustering was applied. Hierarchical clustering was performed using an efficient version of the standard Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical clustering algorithm, implemented in the flashClust package for R. The resulting dendrogram was then partitioned into modules using the `cutreeDynamicTree` method from the `dynamicTreeCut` package for R (Langfelder, Zhang, and Horvath 2008). Rather than use the hybrid branch cut method suggested in the WGCNA vignette, we found that an alternative method developed by the same authors (`cutTreeDynamicTree`) resulted in more functionally enriched networks. A minimum module size (`minModuleSize`) of 10 was used in each case.

Functional annotation of network modules

In each instance where a co-expression network was constructed, functional enrichment of each of the network modules was assessed in order to gain a better understanding of the possible roles of the network modules. Enrichment was measured using the GOSeq package for R (Young et al. 2010), which uses the hypergeometric test after adjusting for a length bias known to be present in RNA-Seq data. One additional benefit to using GOSeq is that it is able to measure enrichment of any arbitrary annotation type, provided you can give it a gene-annotation mapping. For all organisms studied, enrichment of both Gene Ontology (GO) terms (Ashburner et al. 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto 2000) was measured. For human and mouse datasets, we additionally measured

enrichment of ConsensusPathDB (CPDB) pathways (Kamburov et al. 2013), a meta-database containing pathway information from multiple sources. For human datasets alone, we further assessed enrichment of known transcription factor (TF) regulated genes in each network module using a TF-gene mapping created by Marbach *et al. (Marbach et al. 2016a)*. The TF-gene mapping was accessed via a small wrapper package written in R, tftargets (https://github.com/slowkow/tftargets). For *L. major* and *T. cruzi*, enrichment of genes predicted to encode secreted proteins, transmembrane domains, or GPI-anchors was assessed. SignalP 4.1f (Petersen et al. 2011) and TMHMM 2.0c (Sonnhammer, von Heijne, and Krogh 1998) were used to predict genes encoding secreted proteins and transmembrane proteins, respectively, in both *L. major* and *T. cruzi*. To detect genes encoding GPI-anchored proteins, a previously described gene list was used for *L. major* (Myler and Fasel 2008), while PredGPI (Pierleoni, Martelli, and Casadio 2008) was used to predict GPI-anchored proteins in *T. cruzi*. In each case, the default arguments were used. For *L. major*, enrichment of LeishCyc pathways (Doyle et al. 2009) was additionally measured.

Parameter optimization using functional enrichment scores

For a given RNA-Seq dataset, there are many possible approaches to constructing a co-expression network from that dataset. Even when limiting oneself to a specific algorithm or methodology, there are numerous factors including how the data is transformed prior to construction; the similarity metric used to infer co-expression relationships between genes; and various algorithmic parameters relating to the network construction and module detection steps.

In order to assess impact of some of these transformations and parameters on the quality of the resulting co-expression network, a scoring scheme based on functional enrichment was developed. After some initial testing with a wide range of data transformations, filters, and algorithmic parameters, a set of six impactful parameter choices was selected to be thoroughly tested across multiple RNA-Seq datasets (**table 2**). Seven different datasets were selected spanning a range of experiment types (infection, multi-tissue, and developmental transcriptomic datasets) and organisms (*H. sapiens, M. musculus, L. major, T. cruzi, C. elegans, D. melanogaster*) (**table 3**). For each dataset (twice in the case of host-parasite dual transcriptomics datasets) co-expression networks were constructed using all possible combinations of parameter decisions, resulting in 448 - 1,344 individual co-expression networks constructed per dataset/organism, depending on the availability of batch information.

For each co-expression network generated, functional enrichment of the resulting network modules was computed using the GOSeq package for R (Young et al. 2010). Gene Ontology (GO) (Ashburner et al. 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) enrichment were measured for all species investigated. For mouse and human networks, ConsensusPathDB (CPDB) (Kamburov et al. 2013) enrichment was determined, and for human networks alone, transcription factor regulon enrichment was additionally measured using the TF-gene mapping generated by Marbach *et al*. (Marbach et al. 2016a). Each network was then assigned a score based on the weighted combination of *unique* enriched annotations found across all modules in that network (**Formula 2**).

$$\text{score} = \frac{1}{n}(\text{uniq\_go} + \text{uniq\_kegg} + ... + \text{uniq\_annot}_n)$$

Formula 2: Co-expression network scoring metric based on functional enrichment.

After detecting co-expression modules in a network, a score is assigned to the network by taking the weighted average of the scaled unique annotation counts for each available annotation source. By rescaling all individual components of the score, no single annotation type (such as the relatively richly-annotated GO terms) will dominate the score. By only considering *unique* annotations, we avoid double-counting enriched annotations that appear in two separate by highly similar modules.

The number of unique enriched annotations was used in place of total enriched annotations in order to prevent the scoring method from favoring networks with many small but functionally redundant modules. Weighting the contributions of each annotation source equally ensures that no one source with many annotations such as GO dominates the score (**figure 16**).

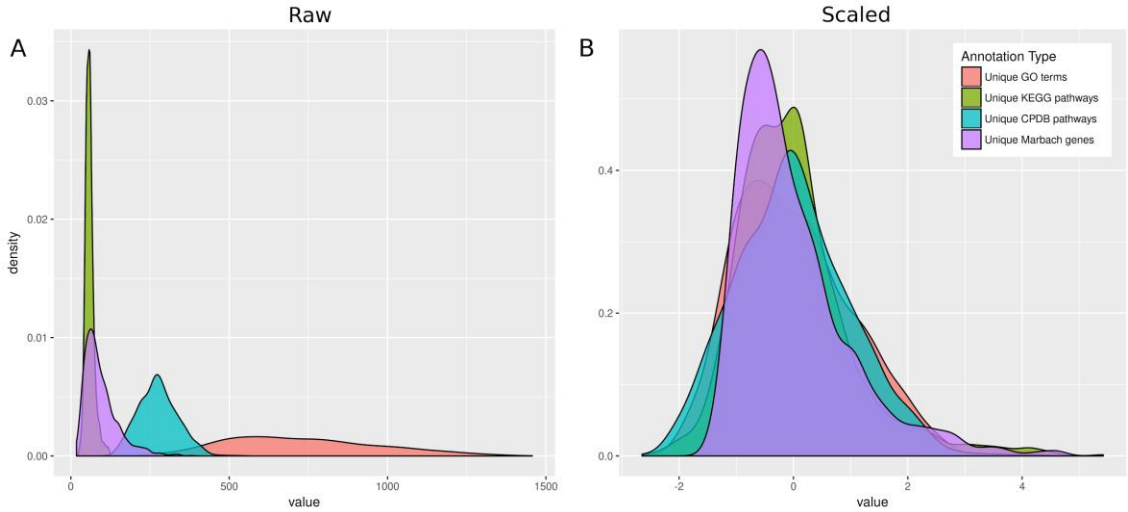

figure 16: Generation of a co-expression network functional enrichment score.

In order to asses the likely quality of a given co-expression network, a network scoring metric based on functional enrichment was developed. Enrichment of functional annotations from several sources (GO, KEGG, CPDB, Known TF regulons) is measured

across all modules in the network and the total number of unique enriched annotations of each type is counted, rescaled, and added together to form the final network score. In the above figure, the count distributions for each annotation type across 1,344 HsLm networks are shown before scaling (A) and after scaling (B). Because of significant differences in the number of annotations assigned to genes from each source, a large amount of variability can be seen in the unscaled count distributions. After rescaling, each annotation source contributes approximately an equal amount to the final network score.

In order to assist in the selection of a *robust* set of optimal parameters network construction, a generalized linear model (GLM) was constructed, attempting to predict network score for a particular dataset, based on the combination of parameters used to construct the network. The model designed to allow for interaction between all network construction parameters except for adjacency power:

```
fit = lm(score ~ cpm * log2 * qnorm * sim_meas + adj_pow, data=networks)
```

Next, plots were generated depicting the predicted network scores for each parameter combination as a function of dataset, for datasets with (**Figure S2**), and without batch information (**Figure S2**). Individual cells in the figure, corresponding to the predicted network score for a given parameter combination and dataset, were colored by performance from green (low network score) to red (high network score), and rows were ordered according to the total predicted score across all datasets.

Robust consensus network generation

An alternative approach to attempting to select a single optimal network parameterization is to combine information across multiple network parameterizations. While any individual network instantiation may not do a good job accurately capturing the underlying co-expression relationships between all genes in the network, as long as

the edge weights for between genes which are co-expressed is higher than what would expect if no signal was present on average, then summing the edge weights across all network instantiations should help to bring out the signal from those edges. This approach serves to both provide robustness against incorrectly inferred edge weights, and to combine different types of information which may be captured by the alternative network instantiations.

To begin with, a consensus adjacency matrix was constructed by summing across all of the individual network adjacency matrices, with each gene-gene pair in the consensus matrix representing the aggregate edge weight for that gene pair across all network instantiations. For the LmAll dataset, networks where no batch adjustment was performed were found to be associated with exceptionally low functional enrichment, and were excluded prior to consensus network construction. Next, in order to eliminate low-confidence genes and edges from the network, the maximum adjacent consensus edge weight was determined for each gene. Genes which do not connect strongly with any other other gene in the network will have relatively low maximum adjacent edge weights, and thus cannot be inferred to be strongly co-expressed with any other genes. Next, a specific cutoff was chosen and all genes with maximum consensus edge weights below that cutoff were removed from the network.

In order to determine an appropriate cutoff value to use for a dataset, multiple cutoff values ranging from the first to the 99th percentile of maximum consensus edge weights were tested. For each filtered network, hierarchical clustering and dynamic branch cutting was performed, and functional enrichment was assessed across all resulting modules. The clustering and enrichment approach used was the same as that

described for the individual co-expression networks above, with one exception: the `cutreeDynamicTree` parameter "`deepSplit`" was set to `FALSE` to allow for a more natural scale of module sizes to be determined.

Plots were generated depicting the number of unique enriched annotations in the network, for each cutoff value tested (**Figure S4**). These plots were then used to choose a specific cutoff value so-as to maximize the overall enrichment across the network, while eliminating as many of the low-confidence genes and edges as possible. In practice, this typically involved looking for inflection points in the graph where the average enrichment per gene starts to increase and/or the total enrichment score across all modules starts to drop off.

figure 17: Effect of low-confidence gene filtering on consensus network functional enrichment.

After a consensus co-expression network has been constructed, low-confidence genes are removed, in order to improve the overall quality and robustness of the network. To filter genes, an edge weight cutoff is selected and all genes which don't have at least one edge weight equal to or greater than the cutoff, and thus cannot be determined to be reliably co-expressed with any other genes, are removed from the network. In order to guide the selection of an appropriate cutoff value to use, edge weight cutoffs corresponding to

the 0.00 - 0.99 percentile of all maximum edge weight values are tested, and the aggregate functional enrichment of the resulting network (sum of -log10 adjusted *P*-values) (**Formula 3**) is plotted as a function of the cutoff used, for several different annotations. The above figure depicts such a plot for the HsLmUI consensus network and the dashed red line indicates the selected cutoff applied for construction of the final consensus network. In this case, the cutoff value associated with the 32nd percentile (x=0.32) was chosen based on its preservation of functional enrichment across several of the annotation sources. Similar plots were generated for each network depicting the number of unique annotations and the ratio of enrichment to number of genes, as a function of edge weight cutoff.

$$\text{score} = \sum -\log_{10}(\text{adj. pval})$$

Formula 3: *P*-value-based scoring metric for network enrichment.

In order to differentiate between annotations which are strongly or weakly enriched, a *P*-value based scoring metric was used, in addition to simply counting the number of enriched annotations in each network. The score is computed as the sum of the negative log10-transformed adjusted *P*-values across the network, for all statistically enriched annotations of a specified type. *P*-values are clipped at 1E-10, so that each enriched annotation can contribute no more than 10 points to the final score.

Chapter 4

Impact of Infection on Host Co-expression Network and Conserved Signatures of Infection

Introduction

During the process of infection, host and parasitic cells interact with and influence one-another in many ways: immune detection and evasion; nutrient acquisition; modulation of host cellular environment, etc. In addition to these types of direct interactions between host and parasite cells, there are also corresponding changes taking place in both the host's and parasite's transcriptome (Jenner and Young 2005). Changes to the host transcriptome can take form in a manner which is beneficial to the host (e.g. successful detection of infection and generation of appropriate immune response), or in a way that is beneficial to the parasite, for example, deregulation of immune response or production of a nutrient or chemical environment beneficial to parasite survival.

So far, most of the research into the transcriptional changes which take place during infection has either focused on specific genes or pathways of interest, or have considered the sets of genes which are up- or down-regulated during infection, as determined by differential expression analyses. For example, Belcher *et al*. (Belcher et al. 2000) used microarray analysis of bronchial epithelial cells infected with *Bordetella pertussis*, the causative agent of whooping cough, to study host immune pathways which are activated early on during infection, and to discover genes necessary for pertussis toxin catalytic activity. In a separate study, Kim *et al*. (S.-K. Kim, Fouts, and Boothroyd 2007)

found that *Toxoplasma gondii* infection of human fibroblast cells results in a dysregulation of IFN-γ-inducible gene expression, benefiting parasite intracellular survival. In trypanosomatids, Maretti-Mira *et al.* (Maretti-Mira et al. 2012) were able to detect transcriptional differences in expression of adaptive immune response genes, chemoattractants to innate cells, and antigen presentation genes in individuals with either Localized Cutaneous Leishmaniasis or Mucosal Leishmaniasis *L. braziliensis* infection.

In addition to the question of how the host transcriptome changes in response to infection, it is also interesting to consider whether there are commonalities in host transcriptional response to infection by different organisms. For example, are there common gene networks which are activated during infection by different intracellular pathogens?

In both of the scenarios above (host transcriptome remodelling in response to infection and detection of conserved transcriptional response to intracellular infection), we are interested in determining similarities and differences in the global host transcriptome, under different environmental conditions. In order to address these questions, we adopt a differential network analysis (DiNA) approach (Mitra et al. 2013; de la Fuente 2010). In recent years, differential network analysis (DiNA) techniques have been used to explore a wide range of biological phenomena, including the detection of causal mutations in transcription factors (Hudson, Reverter, and Dalrymple 2009), the impact of disease on regulatory networks (de la Fuente 2010), the influence of gender on co-expression networks (van Nas et al. 2009), and the prediction of gene targets for effective drug interventions (Zickenrott et al. 2016). Less work has been done, however,

specifically on the network signatures of response to infection (Amit et al. 2009; C. Li et al. 2011; Zhai et al. 2015).

Here, we describe a simple approach which combines differential network analysis with clustering and functional enrichment, in order to detect co-expression sub-networks that are altered across either infection status, or pathogen. This work builds on top of the *L. major* (HsLm and LmHs) and *T. cruzi* (HsTc and TcHs) consensus network analyses described in the previous chapter. Using this approach, we are able to discover widespread changes in the host co-expression network architecture during infection, highlighting specific pathways that become more tightly co-regulated or deregulated during infection, including several relevant immune response pathways not detected by traditional differential expression analysis approaches.

Impact of infection on host co-expression network

While differential expression analysis provides a useful approach for investigating transcriptomics changes that take place in the host during infection (Mandlik et al. 2011; Maertzdorf et al. 2011; Westermann, Gorski, and Vogel 2012), it is often limited to providing a “global” picture of how expression changes during infection, indicating the sets of genes which are up- or down-regulated during infection. It is unable, however, to detect changes in co-expression status between genes or differential rewiring of the host co-expression network. To better understand how the host co-expression network is

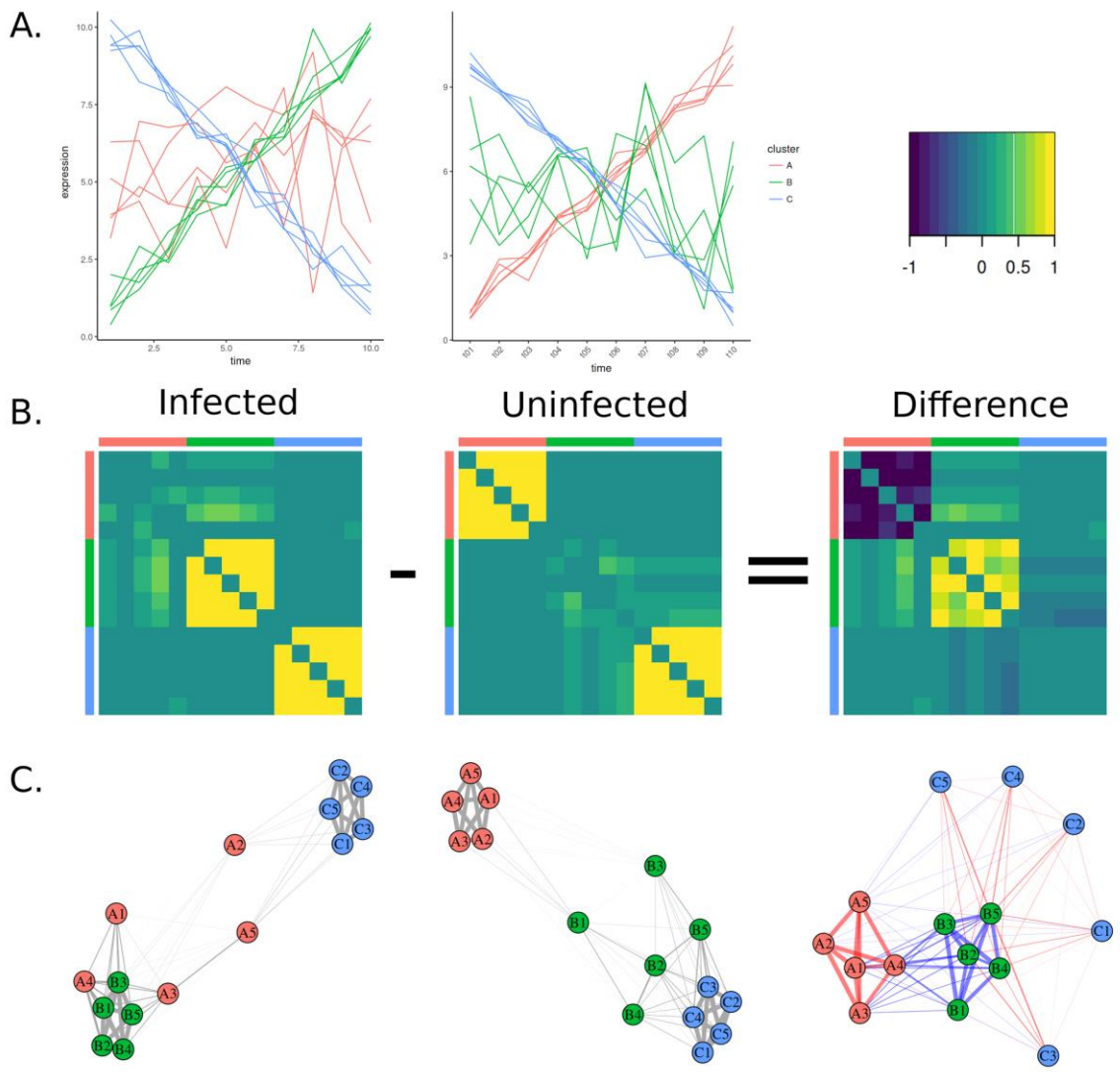

figure 18: Overview of infected vs. uninfected difference network construction.

In the above figure, the basic steps used to construct an in infection difference network are shown. A) Expression profiles are plotted across time for 15 hypothetical genes existing in three separate co-regulatory modules: red , green, and blue. In the uninfected sample (left), the blue and green modules are tightly co-expressed, while and genes in the red module are not co-expression. During infection (middle), the green module genes become deregulated while the red module genes become tightly co-regulated and the blue module genes remain the same. B) Adjacency matrices for the uninfected (left), infected (middle), and difference (right) networks. Values range from -1 (purple) to 0 (green) to +1 (yellow). In the infected and uninfected matrices, all edge weights are positive, while the difference network contains modules with both positive edges (genes which are becoming more tightly co-expressed during infection) and negative edges (genes which are become deregulated during infection). The Blue module of genes whose co-expression status does not change does not form a module in the difference network. C) Network representations of each of the three adjacency matrices. For each, edge thickness corresponds to edge weight. For the difference network, the edges are colored based on the direction of co-expression change: red edges indicate a decrease in co-expression during infection while blue edges indicate an increase in co-expression during infection.

changed during the process of infection, we generated weighted difference co-expression networks (**Figure 19**), capturing the changes in gene co-expression status between infected and uninfected samples, for the HsLm and HsTc datasets.

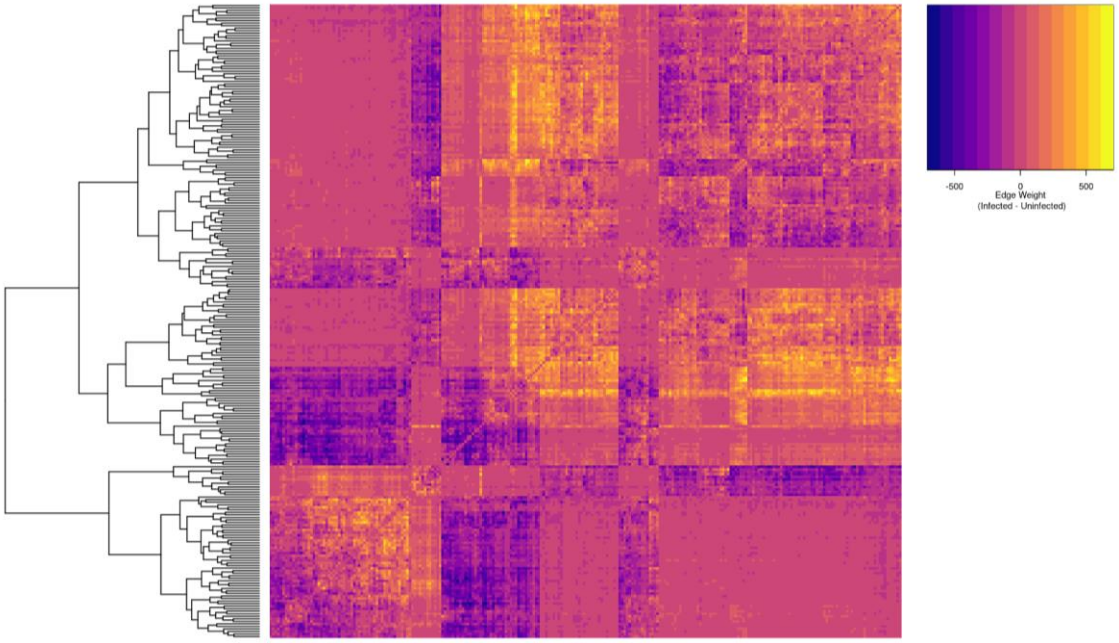

figure 19: Human infected with *L. major* infected vs. uninfected difference adjacency matrix.

A difference co-expression network representing the changes that occur during infection by *L. major* was generated by subtracted the normalized uninfected consensus adjacency matrix (HsLmUI) from the normalized infection consensus adjacency matrix (HsLm). Above, a biclustering heatmap is depicted showing the overall architecture and distribution of edge weights in the resulting network. Edge weights range from -1,179.4 to 1,183.9, with large negative values (purple) indicating gene pairs that become deregulated during infection, and large positive values (yellow) gene pairs that become more tightly co-expressed during infection. The presence of large purple and yellow blocks along the diagonal indicate the presence of sub-networks of genes which undergo similar changes in co-expression architecture during infection. A similar heatmap was generated for the HsTc - HsTcUI difference co-expression network, and the overall patterns and features were found to be similar. For the purposes of visualization, a random sample of 250/5,539 was selected for display above. As such, this plot only shows the large-scale features of the difference network.

H. sapiens infected with *L. major* difference co-expression network

A difference co-expression network was generated by subtracting the normalized uninfected consensus network (HsLmUI) from the normalized infected consensus

network (HsLm), for the set of intersecting genes found in both networks (**figure 19**). The resulting network included 5,539 genes and contained edges ranging from -1,179.4 to 1,183.9 (the theoretical limited of +/- 1,344 determined by the number of individual networks that were used to derive the input consensus networks). Quantile normalization was applied in the consensus network edge weight distributions, prior to difference network construction, in order to account for differences in consensus edge weight distributions due to varying number of RNA-Seq samples used for network construction (**figure 22**).

Next, since it is reasonable to expect that there may be whole pathways or sub-networks or genes that change in co-expression status together (either becoming deregulated or more tightly co-expressed), we performed hierarchical clustering in order to detect modules of genes with similar patterns of co-expression change. A total of 129 modules (median size = 33 genes) were detected, and functional enrichment and multiple testing correction was performed as described in chapter 3. Of the 129 detected modules, 13 were found to be functionally enriched (**table 8).**

Several of the modules containing genes that were deregulated upon infection by *L. major* were enriched for immune response functions and transcription factor regulon genes. Among these, one of the modules (#00FF77, n=36 genes) was enriched for a large number of innate and adaptive immune response related functions including GO terms associated with "regulation of T cell mediated immunity" (adj. *P*-value = 0.0005), "response to interferon-gamma" (adj. *P*-value = 0.026), "reactive nitrogen species metabolic process" (adj. *P*-value = 0.22), and "reactive oxygen species biosynthetic process" (adj. *P*-value = 0.027) (**figure 20**). T cell dependent IFNγ production, which

leads to macrophage activation and production of reactive oxygen and nitrogen species, is required for successful clearance of *L. major* infections, and is a known target of *L. major* inhibition (Nylén and Gautam 2010). Interestingly, none of the aforementioned functions were detected by differential expression analyses previously performed on the same dataset (Fernandes et al. 2016), highlighting the potential power of differential network techniques for host-pathogen transcriptomics analysis.

***H. sapiens* infected with *T. cruzi* difference co-expression network**

A *H. sapiens* / *T. cruzi* infection difference co-expression network was generated from the quantile-normalized HsTc and HsTcUI consensus networks, as described in the previous section and methods. The resulting network included 4,148 genes and contained edges ranging from -1,179.4 to 1,183.9. A total of 95 modules (median size = 32 genes) were detected, and functional enrichment and multiple testing correction was performed as described in chapter 3. Of the 95 detected modules, 14 were found to be functionally enriched (**table 8).**

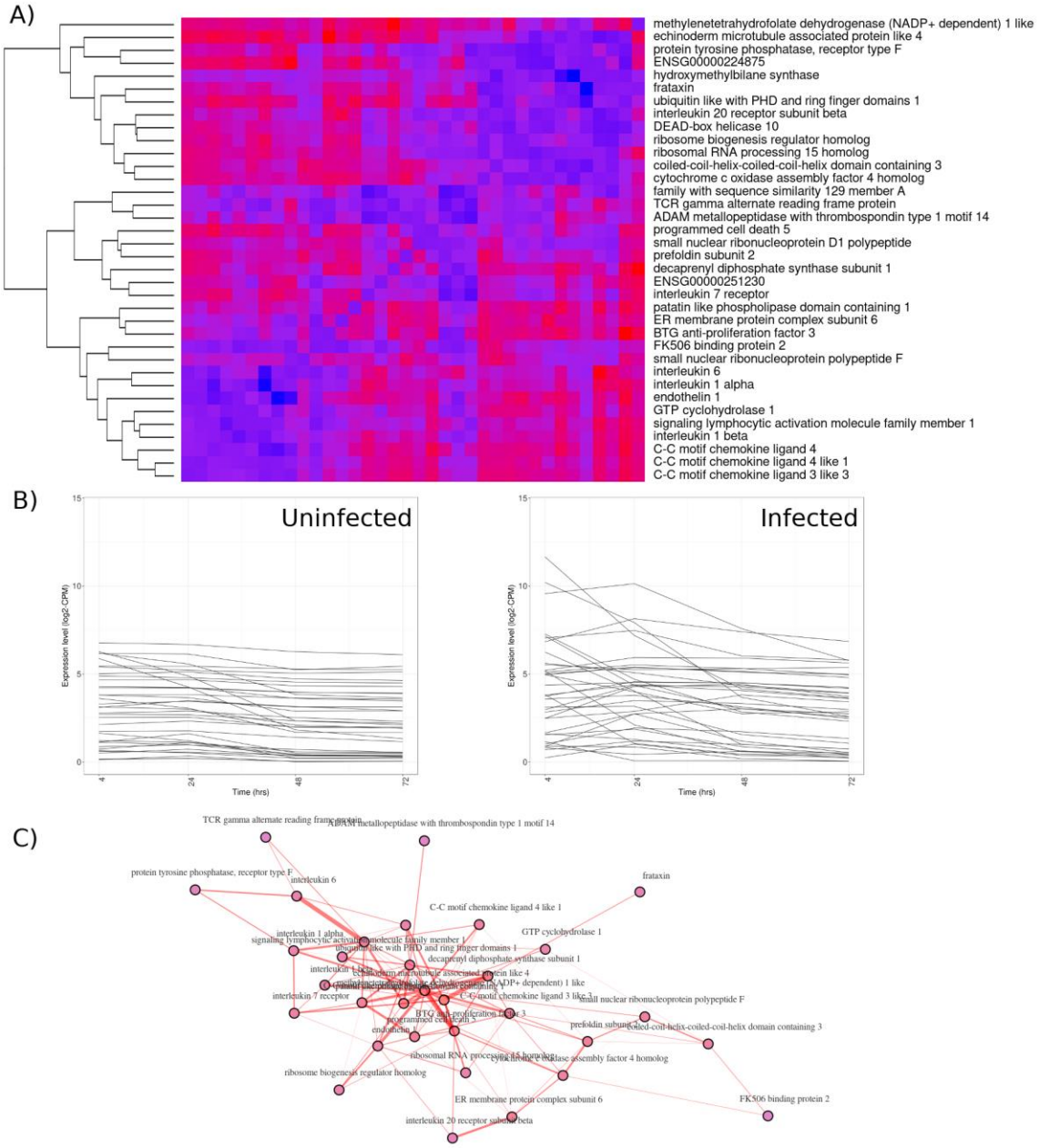

figure 20: Deregulation of key host innate and adaptive immune response pathways upon infection by *L. major*.

Differential network analysis revealed a strong decrease in co-expression of several key pathways known to be involved in host clearance of *L. major* infection, including T-cell activation, IFNγ production, and reactive oxygen and nitrogen species generation. (A) A biclustering heatmap of the 36 difference module genes is shown, with edge weights ranging from -755.3 (red) to 390.3 (blue). Negative edge weights correspond to deregulation during infection, while positive edge weights indicate increases in co-regulation during infection. (B) Log2-CPM expression profiles, averaged across replicates, are shown for each of the 34 genes between 4 and 72 hours post-infection. The increased variance present in the infected expression profiles is reflective of the deregulation which takes place within the module during infection. (C) Network plot of the difference module genes with edge thickness corresponding to the magnitude of change between the infected and uninfected co-expression networks.

Network	Genes	Modules	GO terms	KEGG pathways	CPDB pathways	TF-regulon genes
HsLm (inf vs. uninf)	5,539	129	518	9	520	10
HsTc (inf vs. uninf)	4,148	95	170	0	162	27
HsTc vs. HsLm	1,367	33	121	0	1	8

table 8: Summary of difference network functional enrichment.
Summary of difference network functional enrichment results for the three networks described in this chapter, including number of genes, number of network modules, and the number of unique GO terms, KEGG pathways, CPDB pathways, and TF-regulon genes.

Unlike the HsLm - HsLmUI difference network, the modules detected here were enriched primarily for non-immune response related functions (cell cycle, translation, metabolic processes, etc.) We suspect that the observed lower levels of enrichment may be due, at least in part, to the more limited number of samples used to construct the HsTc and HsTcUI consensus networks (19 and 9, respectively). Despite this, a number of network modules were detected whose co-expression structure changed significantly during infection, and may be interesting to explore for previously unappreciated processes related to pathogenesis and immune response.

Conserved signatures of infection

In addition to considering how a host cell's co-expression network changes during the process of infection by an intracellular parasite, it is also interesting to ask whether any of the changes are conserved across infection by other intracellular parasites. Using the differential network techniques developed in the preceding sections, we next sought to determine whether any such common host transcriptional signatures could be found for

human cells infected with either *T. cruzi* or *L. major* parasites. Starting with the individual infected and uninfected networks discussed above (HsLm, HsLmUI, HsTc, HsTcUI), a scoring scheme was devised, as described in the methods section, and used to generate a conservation score adjacency matrix (**figure 21**). The scoring metric was designed such that gene pairs which were not tightly co-regulated in uninfected cells, but became tightly co-regulated during infection would be assigned positive values near 1, while pairs of genes which were tightly co-expressed in uninfected cells, but became deregulated during infection would be assigned negative score close to -1.

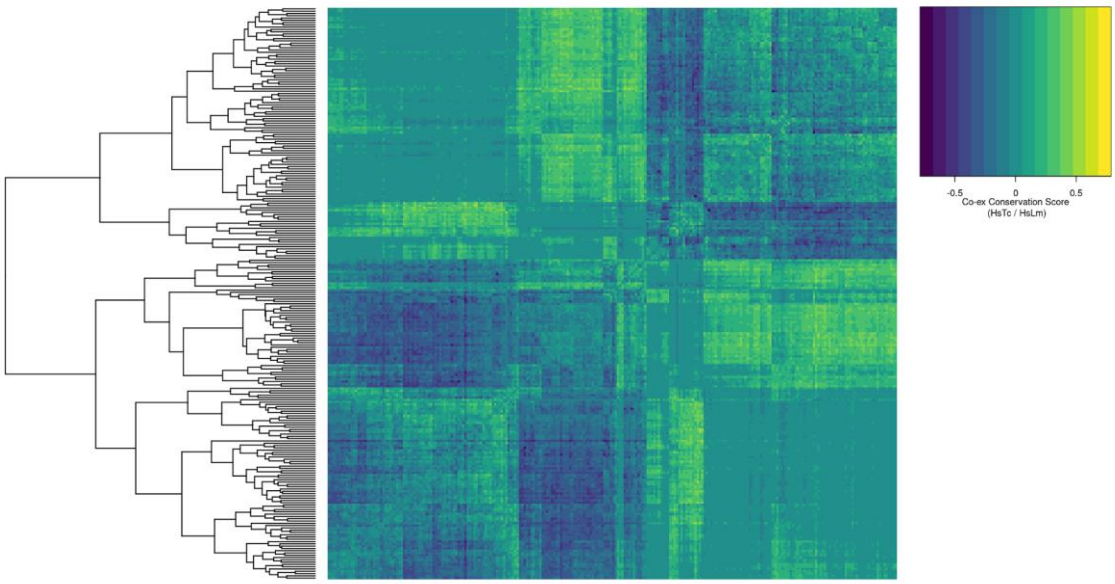

figure 21: Human infected with *L. major* vs. *T. cruzi* difference adjacency matrix.

A co-expression conservation network was constructed from the four consensus co-expression network described above, with edge weights correlated to the level of conservation of co-expression status between genes within infection status. Above, a biclustering heatmap is depicted showing the overall architecture and distribution of edge weights in the resulting network. Edge weights range from -1.0 to 0.75. Large negative values (blue) indicate gene pairs that are highly co-expressed in both uninfected networks, but become deregulated during infection, for both parasites. Large positive values (yellow) indicate host genes pairs which are not co-expressed in either uninfected network, but become tightly co-expressed in both infected networks. The presence of large blue blocks along the diagonal indicate the presence of sets of genes who undergo similar patterns of deregulation during infection in both host-parasite systems. The presence of large purple and yellow blocks along the diagonal indicate the presence of sub-networks of

genes which undergo similar changes in co-expression architecture during infection. A similar heatmap was generated for the HsTc - HsTcUI difference co-expression network, and the overall patterns and features were found to be similar. For the purposes of visualization, a random sample of 250/5,539 was selected for display above. As such, this plot only shows the large-scale features of the difference network.

Conclusion

In the previous chapter, we described tools and techniques for constructing, optimizing, and analyzing individual host and parasite co-expression networks, with a particular emphasis on infection-related networks. Here, we expanded on these approaches, studying methods for comparing changes that occur *across* co-expression networks during infection.

In the first part of the chapter, we developed a simple approach for constructing weighted differential co-expression networks and used that approach to examine the network-level changes that take place during human infection by *L. major* or *T. cruzi*. In the case of the *L. major* system, we demonstrated the power of the differential network approach to detect changes in co-regulation status that are relevant to infection, and undetectable by traditional differential expression analysis methods. This ability to detect changes in co-regulation status of genes, even where there are no significant changes in expression levels, is one of the major advantages of the differential network approaches described here.

Finally, in order to look for a conserved transcriptional signature of intracellular infection in human cells, we extended the basic differential network approach above to rank each gene pair based on their common increased or decreased level of co-regulation during infection. While we were able to observe widespread changes in co-expression status shared by both infection systems, no compelling evidence was found to support the

existence of such a shared transcriptional signature of infection. There are several weaknesses of the current study which may be limiting our ability to detect such signatures. For one, the conservation scoring method requires that each edge evaluated exist in all four input networks (HsLm, HsLmUI, HsTc, and HsTcUI). Due to earlier filtering applied to each of these input networks, the intersection of genes available to study was relatively small (~1,000). One possible way alleviate this issue would be to postpone earlier filtering steps until a later point in the analysis, or simply relax the filtering stringency altogether. Considering the performance of the unfiltered consensus co-expression networks compared to the individual networks (**table 5**), this might not be as detrimental as one would expect, and may deepen the chances of detecting such shared signatures of infection. Finally, in the present study, transcriptional signatures of macrophages and non-myeloid cells were compared. In the future, it may be interesting to look for conserved signatures for the same cell type infected by two different parasites.

Methods

Host infected vs. uninfected differential network analysis

Robust consensus networks were constructed using either infected sample (HsTc, HsLm), or matched uninfected samples (HsTcUI, HsLmUI) as described above. Next, for each host-parasite system, the intersection of genes present in the infected and uninfected networks was determined, and genes not shared by the two networks were removed. Due to differences in the number of samples used to construct the infected and uninfected consensus networks, the distribution of consensus edge weights differed across networks. This is to be expected as networks constructed from few samples will contain a greater number of spurious correlations. To address this issue, the edge weight distributions of

the infected and uninfected consensus networks were quantile normalized (Bolstad et al. 2005) so that each network contains the same distribution of edge weight values. (**figure 22**)

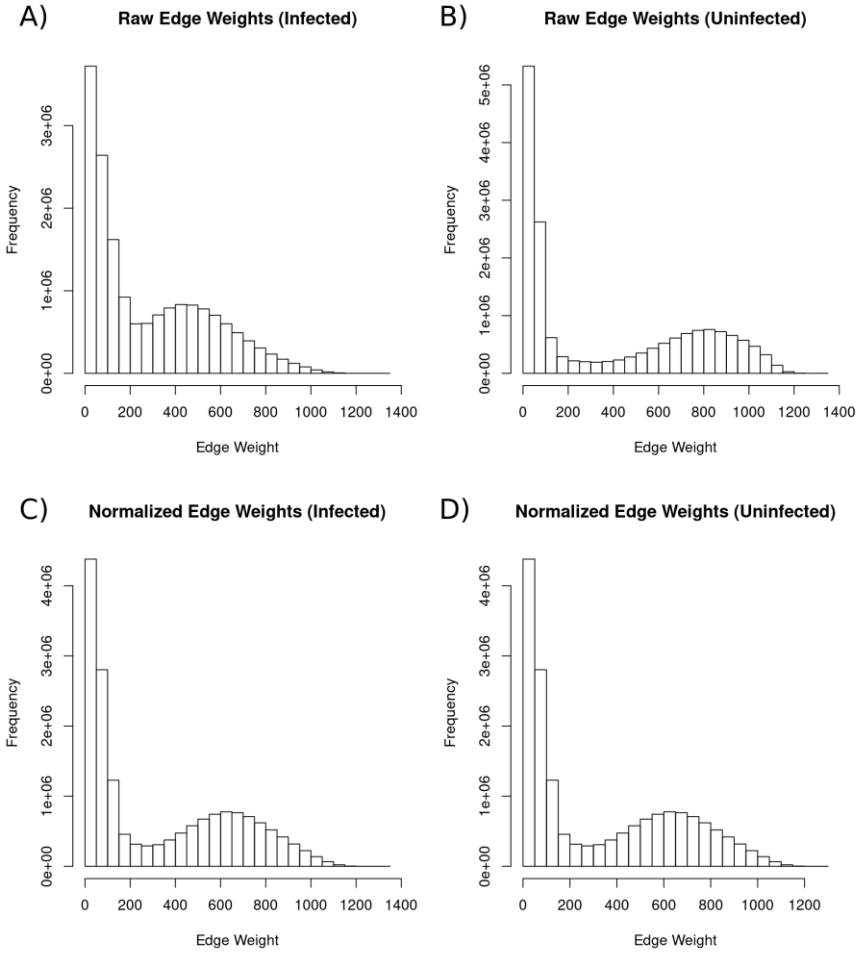

figure 22: Quantile normalization of difference network edge weight distributions.

In order to account for differences in consensus co-expression network edge weight distributions, due to factors such as vary numbers of input samples use for network construction, the infected and uninfected edge weight distributions are quantile normalized, prior to computing the difference network. Shown above are the HsTc (A) and HsTcUI (B) edge weight distributions, prior to normalization. The smaller number of RNA-Seq samples (n=9) used to construct the HsTcUI consensus network, biases the distributions of edge weights towards largely values. After normalization is performed, both the infected (C) and uninfected (D) edge weights have the same distribution.

Next, after reordering the normalized infected and uninfected consensus adjacency matrices, a difference consensus network was constructed by subtracting the

uninfected network adjacency matrix from the infected network adjacency matrix. Each edge in the resulting difference network represents the change in co-expression magnitude between infected and uninfected cells: large positive edge weights represent an increase in co-expression between two genes during infection, while large negative edge weights represent a deregulation during infection. A distance matrix was generated by subtracting the absolute value of the difference matrix from one (ensuring that groups of genes which are changing in co-expression states together will remain close to one another in the distance matrix), and hierarchical clustering and dynamic branch cutting was performed, as described for the regular consensus co-expression networks. Finally, functional enrichment of of GO terms, KEGG pathways, CPDB pathways, and TF regulon genes was assessed for each difference module, using the approach described for the regular consensus networks, and visualizations of specific sub-networks of interest were generated using the igraph package for R (Csardi and Nepusz 2006).

Detection of common transcriptional response to infection by intracellular parasites

For each of the host-parasite systems studied (*H. sapiens* / *T. cruzi* & *H. sapiens* / *L. major*), consensus networks were constructed for the infected samples (HsLm, HsTc) and uninfected samples (HsLmUI, HsTcUI), as described in chapter 3, and edge weights were quantile normalized, as described in the previous section. Next, each matrix was max-scaled (divided by the maximum value in the matrix), in order to shift the values to the range [0, 1], and a co-expression "conservation score" matrix was generated as follows (**Formula 4**):

$$\text{score} = \frac{1}{2}(\text{HsLm} + \text{HsTc}) - \frac{1}{2}(\text{HsLmUI} + \text{HsTcUI})$$

Formula 4: **Scoring metric for conservation of transcriptional signature across infections.** Conservation score metric indicating level of infection-specific co-expression conservation for two different parasites infecting the same host.

The score is designed such that gene pairs which are not tightly coregulated in either uninfected network but become highly co-regulated during infection will have a large positive value close to 1, while pairs of genes which are co-regulated under normal circumstances but become deregulated during infection will have a negative value close to -1. Finally, the conservation matrix is convert to a distance matrix and clustering is performed, as described in the previous section.

Chapter 5

Future Directions

The research described in this thesis has resulted in the production of a large number of resources relating to trypanosomatid gene structure, infection-related host and parasite co-expression networks, and functionally annotated host and parasite co-expression modules. Further, comparative research on the construction and optimization of co-expression networks from RNA-Seq data across a variety of experimental designs and organisms provides useful guidance for future efforts to construct co-expression networks. While each of the analyses discussed has already led to a number of interesting findings, significant opportunities remain to build on the work described here.

Combine detailed trypanosomatid gene structure information with detected co-expression modules to predict key regulatory elements.

In this thesis, generalized methods were developed both for the prediction of trypanosomatid gene structure information (UTR boundaries, alternative trans-splicing and polyadenylation sites, etc.) and for the generation of robust parasite co-expression networks and modules. A key challenge in the field of trypanosomatid research is the elucidation of the precise mechanisms of gene regulation used by parasites to control expression at various life cycle stages. While some example of individual RNA binding proteins or 3’ UTR motifs related to the regulation of one or several genes have been described in various trypanosomatid species, a cohesive understand of how regulation occurs at the genome level has remained elusive.

The results of this work provide a powerful resource for the large-scale bioinformatic detection of important parasite regulatory elements. By combining the UTR boundary information described here for *T. cruzi* (and in previous research I contributed to for *L. major*), along with the detected co-expression modules for each of these parasites, one can start to look for overrepresented primary or secondary motifs in the UTRs of genes in each module. Indeed, a pipeline to tackle this issue has already been developed (https://github.com/elsayed-lab/tryp-reg-predict), and is currently being used to analyze the outputs from this research. Future work is required, however, to optimize the pipeline detection parameters and supervised learning approach, as well as to analyze the outputs of this effort. Finally, once specific predictions have been made regarding likely regulatory elements, experimental validation will be required, for example using the CRISPR/Cas9 protocols currently being developed for some trypanosomatid species to insert predicted regulatory elements into genes where they are not usually found, to ensure that the predicted elements behave as expected.

Investigate the relationship between trypanosomatid UTR length and expression

An interesting observation to come out of this work is the unexpected appearance of a relationship between 5' and 3' UTR length in both *L. major* and *T. cruzi*. In both parasites, A significant different in the average UTR lengths could be seen between sets of genes which were increasing or decreasing in expression across the duration of infection. While some research exists to suggest that UTR length can be detected directly, and involved in the modulation of mRNA stability and decay in other organisms (Hogg and Goff 2010; Mishima and Tomari 2016), no such machinery has been described to

date in trypanosomatids. In this work, the relationship between UTR length and expression was explored for genes with UTR lengths at the extreme ends, and for genes which were differentially expressed. When UTR length was plotted in relation to co-expression status in co-expression network dendrograms for *L. major* and *T. cruzi*, a global trend also appeared to be present.

As a first step, one could explore the role of UTR length at the co-expression modules level. For example - do some co-expression modules have significantly shorter or longer UTRs than expected? Do modules containing genes with similar UTR lengths tend to follow similar expression patterns? And so on. Additionally, since it has been shown here and elsewhere that alternative trans-splicing and polyadenylation is pervasive in these parasites, it may be interesting to monitor UTR length across developmental stage for some of the co-expressed genes with atypically long or short UTRs. Finally, since it is possible that UTR length may *indirectly* play a role in expression, through the variable inclusion or exclusion of regulatory elements, it may be interesting to look for specific motifs which are present in the sets of co-expressed genes with long UTRs, especially in cases where the same genes were found to have alternative TS / Poly(A) sites that would lead to much shorter UTRs.

Extend evaluation of co-expression network techniques to additional methods and broaden validation approach

In this thesis, we investigated the influence of various steps in network construction and module detection on the quality of the resulting network, across a range of different datasets, at a much wider scope that has been described previously. While

work has been done to describe the “best practices” for analyzing RNA-Seq data using differential expression and functional enrichment approaches (Conesa et al. 2016), this research offers some of the first guidance for determining how to address such basic questions as how RNA-Seq data should be transformed prior to co-expression network construction. That said, while we attempted to consider as many important decision points as possible during network construction (and have previously tested and ruled out a number of others), there are many other possible transformations and parameter changes that we did not explore.

Moreover, in this thesis, we primarily focused on a single widely-used co-expression network construction method (WGCNA). While the WGCNA pipeline provides a lot of flexibility for incorporating different similarity metrics and so on, the basic approach is still the same for all networks constructed in this manner. Expanding this work to alternative co-expression network inference methods, in particular those which attempt to separate the direct and indirect contributions of genes to co-expression relationships, such as ARACNE (Margolin et al. 2006), and other methods based on partial correlation (Allen et al. 2012), would provide a useful benchmark for future researchers attempting to choose between alternative methods.

Finally, much of the work in this thesis relating to the influence of parameter choice on network quality, as well as the optimization and interpretation of specific co-expression networks, depends crucially on our ability to accurately judge network quality. While the incorporation of a recent high-quality TF-gene map (Marbach et al. 2016b) provides a useful first step in the direction of external co-expression network

validation, inclusion of additional sources of high-quality experimentally validated regulatory networks would enhance nearly all of the network-based analyses described in this thesis.

Appendices

Appendix 1: Supplemental figures

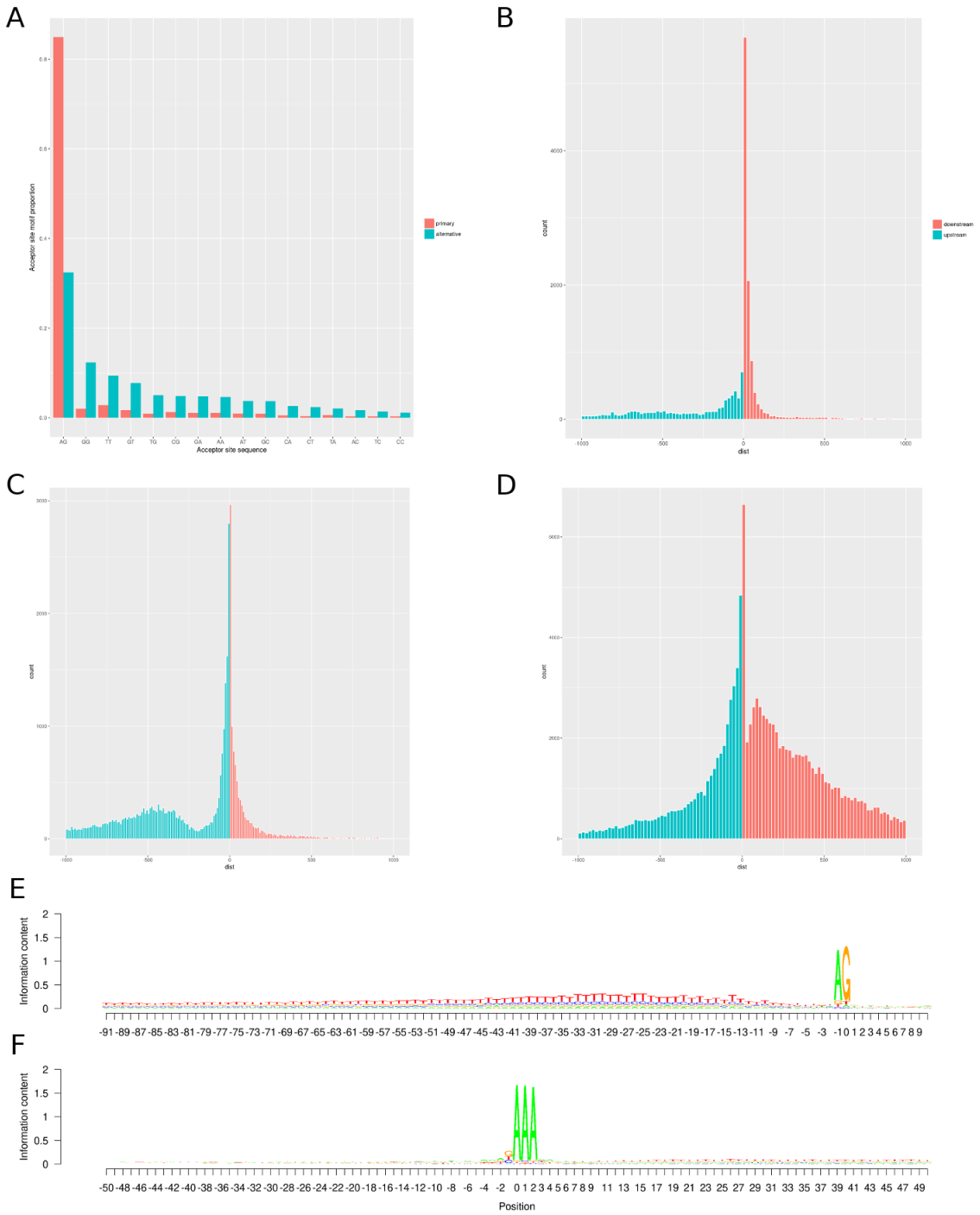

Figure S1: Sequence composition at RNA processing sites and distance between primary and alternative RNA processing sites.

Spliced leader acceptor sites were retrieved by extracting the dinucleotide sequence upstream of each detected *trans*-splicing site. **(A)** The frequency of dinucleotide usage was determined for the primary and alternative acceptor sites. The distribution of distances between primary and alternative SL *trans*-splicing sites was determined for splicing sites using the canonical AG (**B)** and non-AG sequences **(C)**. Negative lengths correspond to alternative acceptor sites located upstream of their cognate primary acceptor sites while positive values indicate alternative acceptor sites found downstream. **(D)** The distribution of the distances between primary and alternative polyadenylation sites was calculated. Negative lengths correspond to alternative polyadenylation sites located upstream of their cognate primary acceptor sites while positive values indicate alternative polyadenylation sites found downstream. A similar analysis was performed with transcripts derived from trypomastigotes, epimastigotes, and amastigotes (**Fig. S6**). The sequence composition of the region encompassing 90 nt upstream and 10 nt downstream of the *trans*-splicing sites was plotted for the primary **(E)** and alternative (**Fig. S4)** splice sites using seqLogo. The sequence composition of the region encompassing 50 nt upstream and 50 nt downstream of the primary poly(A) addition site was plotted for the primary **(F)** and alternative (**Fig. S5**) poly(A) addition sites using seqLogo.

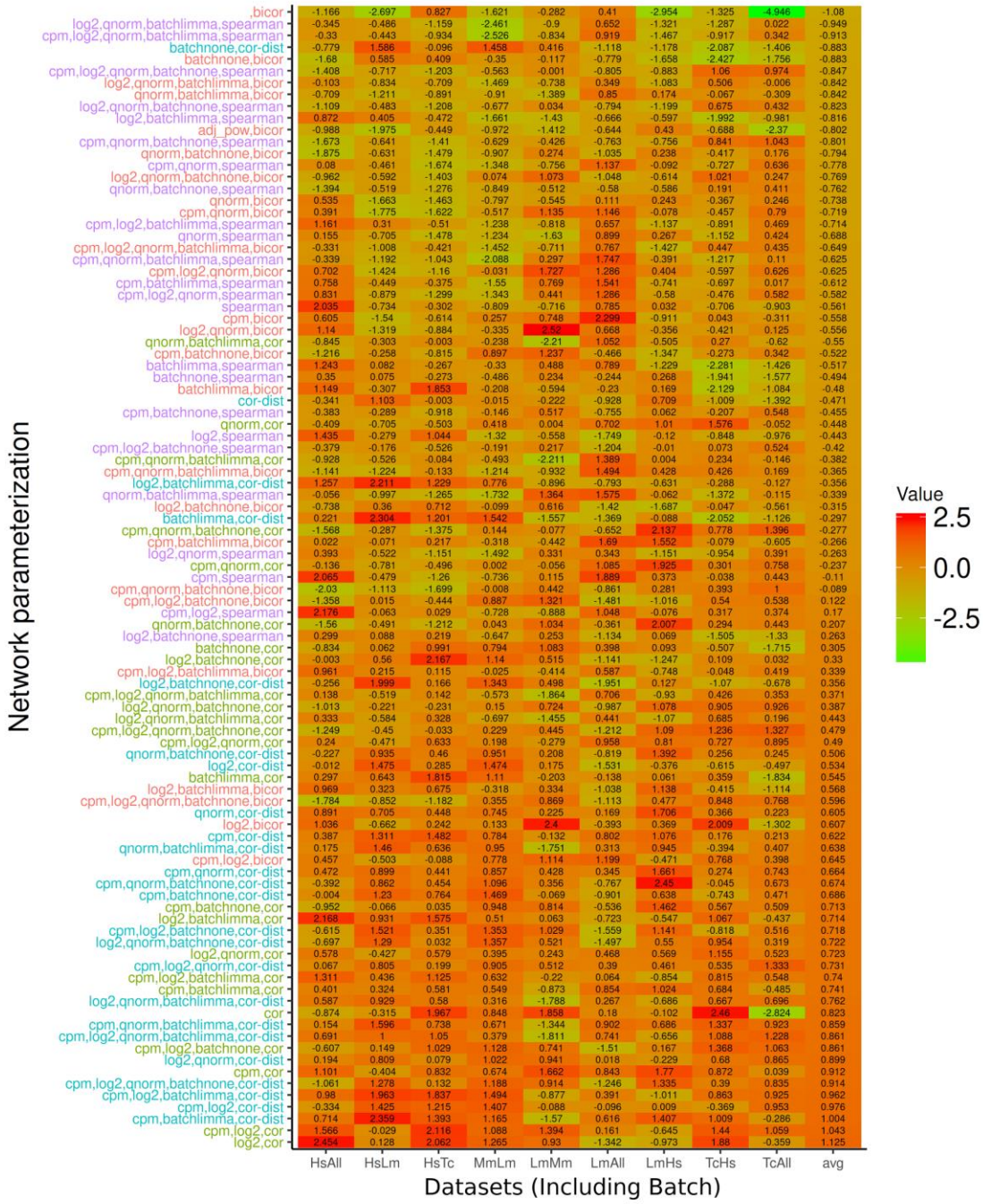

Figure S2: Summary of parameter choice influence on network functional enrichment across datasets with known batch information.

For each of the nine dataset/organism pairs analyzed, 1,344 possible alternative network parameterizations were constructing using each possible combination of alternative parameter values. For each network, functional enrichment was measured and a score assigned. Excluding adjacency power, which did not appear to have any consistent impact on the enrichment scores, each possible parameter combination is shown across a single row in the plot. The average enrichment score across all datasets for each parameter combination was computed (rightmost column) and the rows were reordered such that the parameter combinations that produced the most functionally enriched networks are displayed at the bottom, while

poorer performing parameter combinations are display at the top. The specific parameter combinations are listed in the labels along the y-axis, and are colored according to the similarity metric they use. The fact that bright red cells (those associated with highly enriched networks) are not limited entirely to the bottom of the plot demonstrates that different parameter combinations work well for different datasets. On the other hand, the strong over-representation of Pearson correlation and cor-dist among high-ranking parameter combinations shows that there are some general trends which persist across many different datasets.

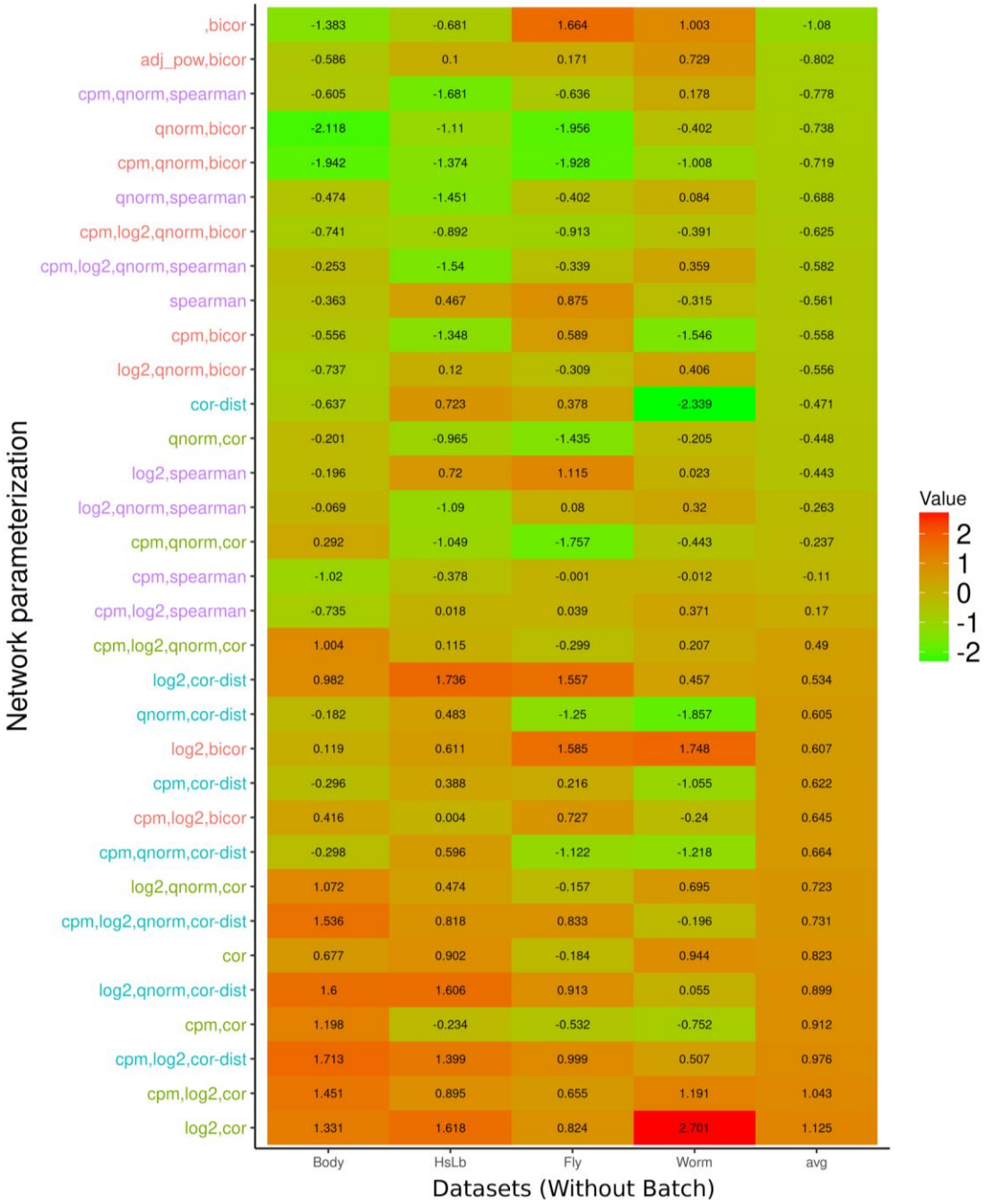

Figure S3 Summary of parameter choice influence on network functional enrichment across datasets without known batch information.

Similar to Figure 3.6, except that only datasets for which no batch information was available (and thus batch adjustment methods could not be evaluated) are shown. A similar presence of both dataset-specific variability and general trends with respect to specific similarity metrics can be observed.

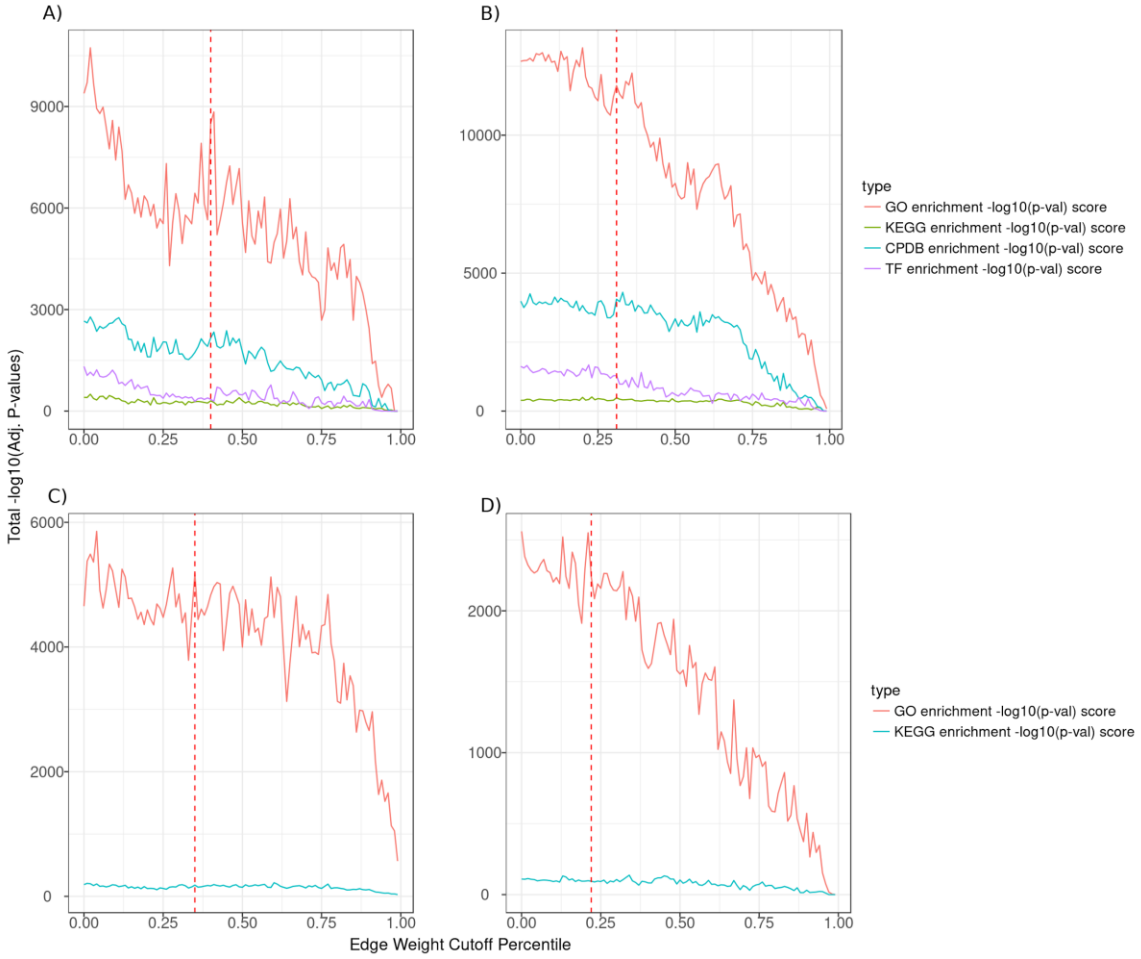

Figure S4: Impact of filtering on consensus network functional enrichment.
In order to construct remove low-confidence genes and create a more robust "core" consensus co-expression network for each dataset, genes with low maximal edge weights were progressively removed from the network, and functional enrichment (the sum of -log10 adjusted *P*-values for all significantly enriched annotations; (**formula 3.3**)) was computed for each of the resulting filtered networks. Above, enrichment scores are plotted as a function of the cutoff used for (A) HsLm, (B) HsTc, (C) LmAll, and (D) TcAll.

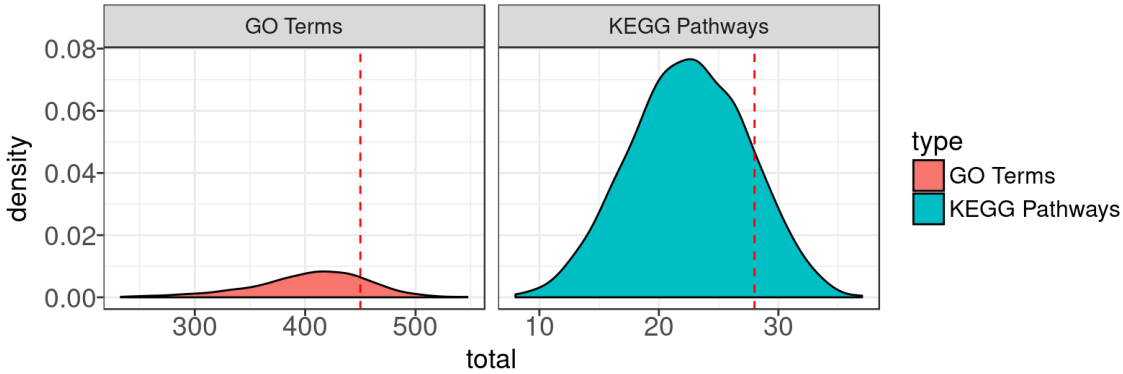

Figure S5: Comparison of functional enrichment for consensus and individual co-expression networks (LmAll).

Kernel density plots depicting the distribution of functional enrichment scores (total number of unique annotations) for each of the 896 LmAll co-expression networks are shown, separated by annotation type. For each annotation source, the level of functional enrichment of that same type for the unfiltered consensus co-expression network is shown as a red dashed line.

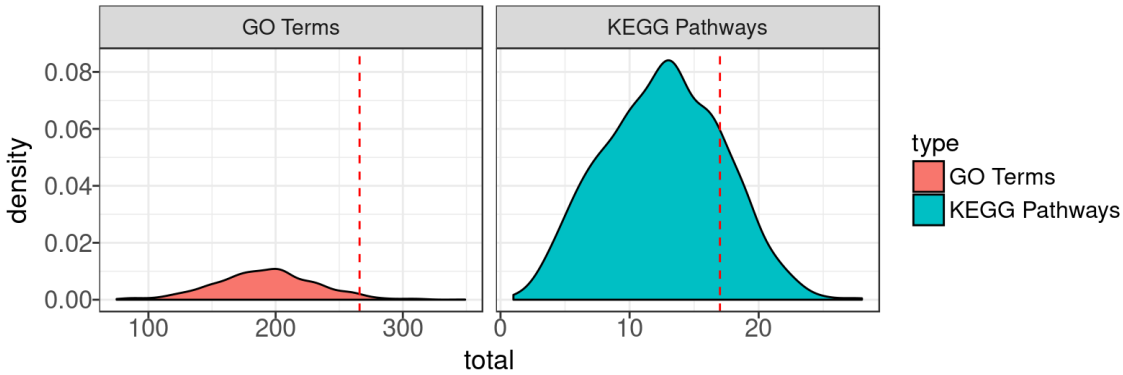

Figure S6: Comparison of functional enrichment for consensus and individual co-expression networks (TcAll).

Kernel density plots depicting the distribution of functional enrichment scores (total number of unique annotations) for each of the 1,344 TcAll co-expression networks are shown, separated by annotation type. For each annotation source, the level of functional enrichment of that same type for the unfiltered consensus co-expression network is shown as a red dashed line.

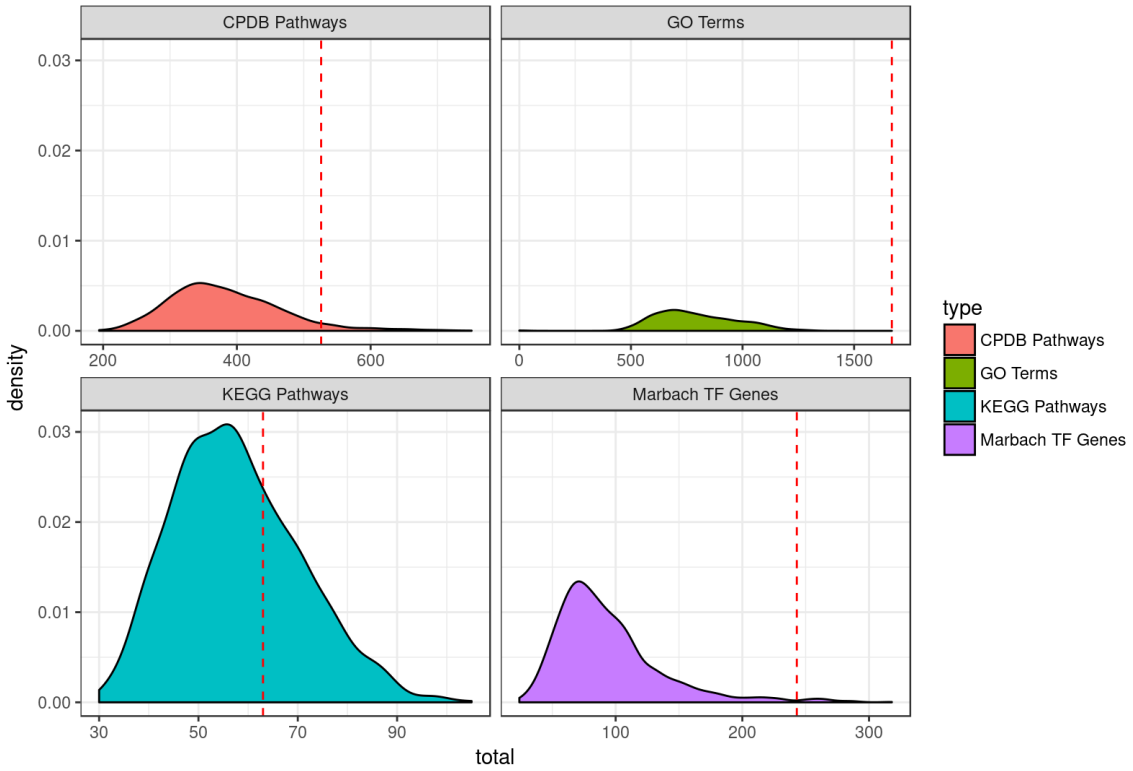

Figure S7: Comparison of functional enrichment for consensus and individual co-expression networks (HsTc).

Kernel density plots depicting the distribution of functional enrichment scores (total number of unique annotations) for each of the 1,344 HsTc co-expression networks are shown, separated by annotation type. For each annotation source, the level of functional enrichment of that same type for the unfiltered consensus co-expression network is shown as a red dashed line.

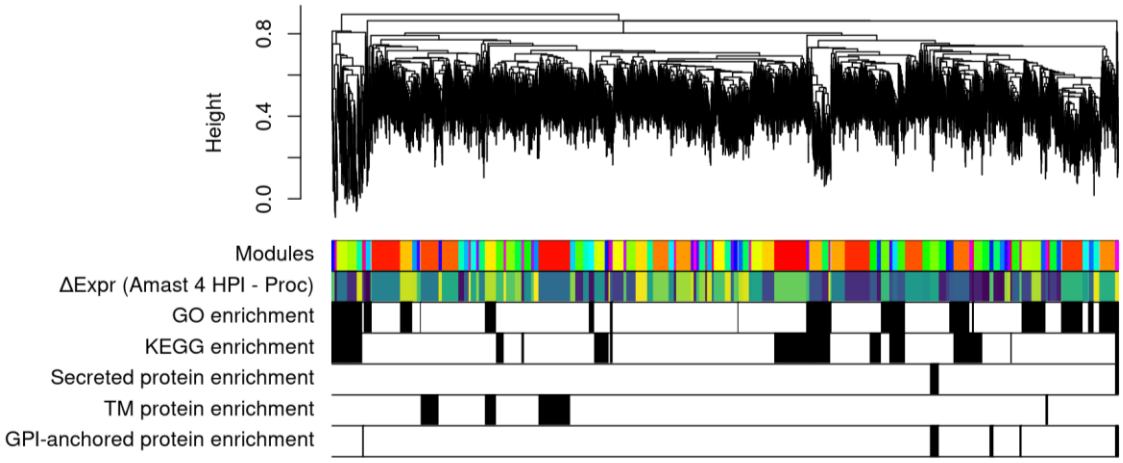

Figure S8: *L. major* consensus co-expression network dendrogram with phenotypic correlates.

The *L. major* consensus co-expression network (LmAll) is depicted using a hierarchical clustering dendrogram (**top**). Each leaf in the dendrogram represents a single gene, with gene proximity in the dendrogram indicating similarity in expression profiles. The height of the dendrogram shows the levels of correlation between expression profiles, with genes near the bottom of the dendrogram (those with a lower height) being more highly correlated. **(Bottom)** Gene module assignments and other phenotypic features are depicted as rows of colors. (a) Module assignments, (b) Module net change in average log2-CPM expression between procyclic promastigotes and amastigotes (4 hpi) hours (yellow = up-regulated, purple = down-regulated), and statistical enrichment of: (c) GO terms, (d) KEGG pathways, (e) genes encoding known or predicted secreted proteins, (f) genes encoding known or predicted transmembrane proteins, and (g) genes encoding known or predicted GPI-anchored proteins.

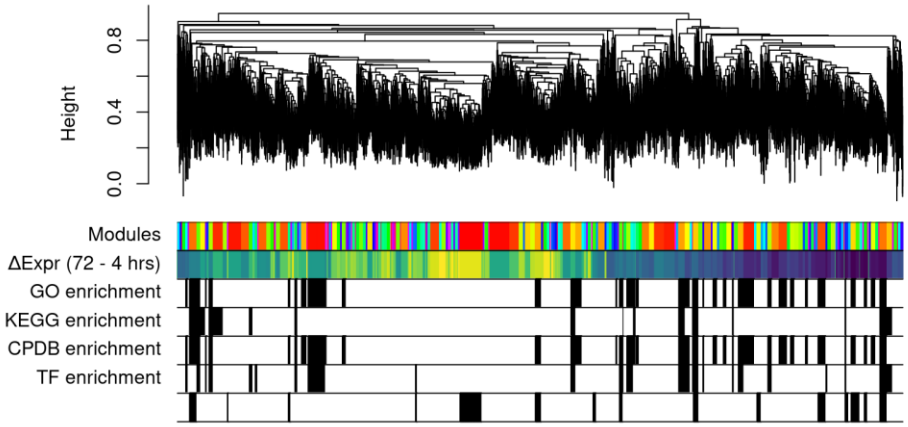

Figure S9: Human infected with *L. major* consensus co-expression network dendrogram with phenotypic correlates.

The *Human infected with L. major* consensus co-expression network (HsLm) is depicted using a hierarchical clustering dendrogram (**top**). Each leaf in the dendrogram represents a single gene, with gene proximity in the dendrogram indicating similarity in expression profiles. The height of the dendrogram shows the levels of correlation between expression profiles, with genes near the bottom of the dendrogram (those with a lower height) being more highly correlated. **(Bottom)** Gene module assignments and other phenotypic features are depicted as rows of colors. (a) Module assignments, (b) Module net change in average log2-CPM expression between 4 and 72 hours post-infection (yellow = up-regulated, purple = down-regulated), and statistical enrichment of: (c) GO terms, (d) KEGG pathways, (e) CPDB pathways, and (f) known TF regulon genes (Marbach et al. 2016b).

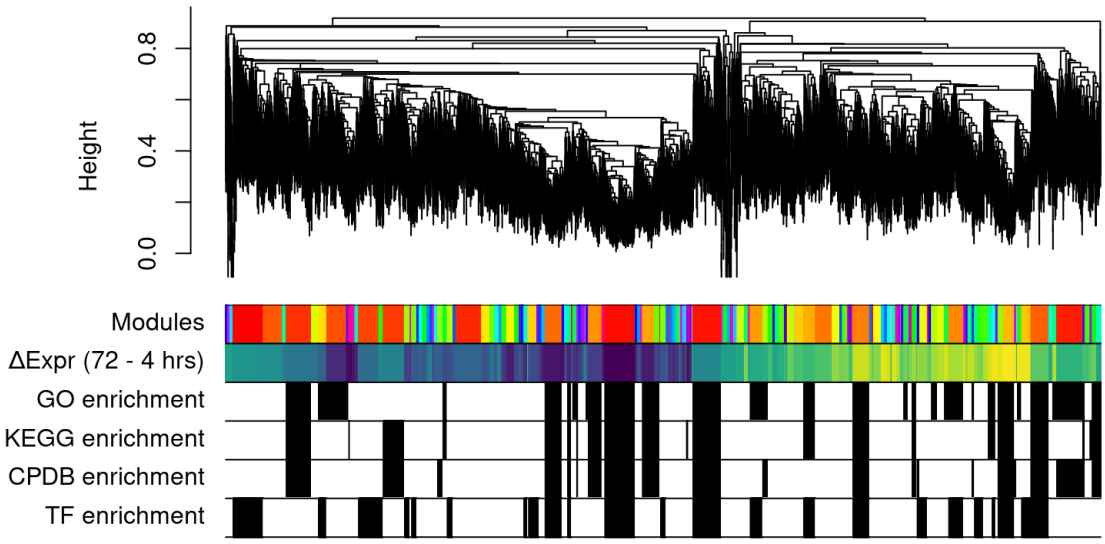

Figure S10: Human infected with *T. cruzi* consensus co-expression network dendrogram with phenotypic correlates.

The *Human infected with T. cruzi* consensus co-expression network (HsTc) is depicted using a hierarchical clustering dendrogram (**top**). Each leaf in the dendrogram represents a single gene, with gene proximity in the dendrogram indicating similarity in expression profiles. The height of the dendrogram shows the levels of correlation between expression profiles, with genes near the bottom of the dendrogram (those with a lower height) being more highly correlated. **(Bottom)** Gene module assignments and other phenotypic features are depicted as rows of colors. (a) Module assignments, (b) Module net change in average log2-CPM expression between 4 and 72 hours post-infection (yellow = up-regulated, purple = down-regulated), and statistical enrichment of: (c) GO terms, (d) KEGG pathways, (e) CPDB pathways, and (f) known TF regulon genes (Marbach et al. 2016b).

Appendix 2: Supplemental tables

Parasite Stage	# reads mapped to host	# reads mapped to parasite	# SL-containing reads	% SL-containing reads	# Poly(A)-containing reads	% Poly(A) -containing reads
Amastigote 04 hpi	192,152,264	3,912,746	22,986	0.59	47,183	1.22
Amastigote 06 hpi	292,863,512	5,488,512	36,733	0.67	69,423	1.25
Amastigote 12 hpi	146,330,584	2,652,012	21,540	0.81	34,746	1.32
Amastigote 24 hpi	187,305,678	9,236,840	97,795	1.05	121,513	1.32
Amastigote 48 hpi	175,666,258	34,415,822	458,606	1.27	344,333	0.97
Amastigote 72 hpi	122,860,082	33,368,902	424,001	1.24	299,727	0.91
Epimastigote	0	72,725,088	773,333	1.09	521,194	0.72
Trypomastigote	31,014	122,636,020	1,044,193	0.88	1,295,188	1.01

Table S1: Summary of RNA-Seq read support for RNA processing events.

Number of reads mapped to host and parasite for each developmental stage, including the amount of reads for which a *trans*-splicing or polyadenylation event was detected.

Organism	Reference Genome	Source
H. sapiens	GRCh38.83/hg38	UCSC
M. musculus	GRCm38/mm10	UCSC
L. major strain Friedlin	TriTrypDB-27_LmajorFriedlin	TriTrypDB
T. cruzi strain Y	TriTrypDB-27_TcruziCLBrenerEsmeraldo-like	TriTrypDB

Table S2: Host and parasite reference genomes.

For RNA-Seq samples generated in our lab, reads were mapped to available reference genomes accessed from either ENSEMBL (host) or TriTrypDB (parasite). This table lists the specific reference genome versions used for each dataset. For the modENCODE Fly and Worm datasets used in this work, pre-generated count tables based on ENSEMBL 61 were downloaded via Recount (Frazee, Langmead, and Leek 2011).

Dataset	Organism	Experiment type	Samples	Reference
L. major infecting *H. sapiens* (LmHs)	*Leishmania major*	Host-pathogen	19	(Fernandes et al. 2016)
H. sapiens infected with *L. major* (HsLm)	*Homo sapiens*	Host-pathogen	54	(Fernandes et al. 2016)
L. major infecting *M. musculus* (LmMm)	*Leishmania major*	Host-pathogen	13	(Dillon et al. 2015)
M. musculus infected with *L. major* (MmLm)	*Mus musculus*	Host-pathogen	27	(Dillon et al. 2015)
L. major multi-experiment (LmAll)	*Leishmania major*	Multi-experiment	73	(Dillon et al. 2015; Fernandes et al. 2016; Inbar et al. 2017)
T. cruzi infecting *H. sapiens* (TcHs)	*Trypanosoma cruzi*	Host-pathogen	19	(Y. Li et al. 2016)
H. sapiens infected with *T. cruzi* (HsTc)	*Homo sapiens*	Host-pathogen	29	(Y. Li et al. 2016)
T. cruzi all stages (TcAll)	*Trypanosoma cruzi*	Host-pathogen (including extracellular stages)	26	(Li et al. 2016)
H. sapiens infected with *L. braziliensis* (HsLb)	*Homo sapiens*	Host-pathogen	35	(Christensen et al. 2016)
H. sapiens multi-experiment (HsAll)	*Homo sapiens*	Multi-experiment	129	(Fernandes et al. 2016; Li et al. 2016; Christensen et al. 2016)
Illumina Human BodyMap 2.0 (Body)	*Homo sapiens*	Multi-tissue	19	(*)
modENCODE Fly (Fly)	*Drosophila Melanogaster*	Developmental transcriptome	30	(Graveley et al. 2011)
modENCODE Worm (Worm)	*Caenorhabditis elegans*	Developmental transcriptome	46	(Hillier et al. 2009)

Table S3: Summary of datasets used for co-expression network analysis.
The abbreviation used for each dataset throughout this thesis shown in parentheses. The samples column indicates the total number of RNA-Seq samples used for network construction. Note that some of the original experiments included additional samples, such as extracellular parasite stages, which are not included in the sample counts above. (*****Illumina Human BodyMap 2.0 dataset is unpublished, but can be accessed at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/)

Dataset	Log2	CPM	Quantile normalization	Batch	Similarity measure	Adjacency power
LmHs		X	X	combat	cor	9
LmMm	X		X	combat	bicor	3
LmAll		X		combat	spearman	1
TcHs				combat	cor	7
TcAll		X	X	none	cor	8
ModFly	X			N/A	cor-dist	12
ModWorm	X			N/A	cor	1
MmLm				limma	cor-dist	10
HsLm				limma	cor-dist	3
HsTc		X		limma	cor-dist	3
HsLb	X		X	N/A	cor-dist	4
HsAll	X			combat	cor	10
BodyMap	X		X	N/A	cor-dist	5

Table S4: Summary of parameter choices for optimized co-expression networks.

The specific optimized parameter combinations for each of the thirteen co-expression networks constructed are shown. Batch adjustment method aside, the three best performing parameter combinations were (1) log2-adjustment only, (2) log2- and quantile normalization, and (3) no adjustments, each of which was found to be optimal for three of the thirteen networks. The best-performing batch adjustment method is ComBat (5/9 networks with batch information), while the best performing similarity metric across all datasets is cor-dist (6/13 datasets). No strong preference for any particular adjacency power was observed among the optimized networks.

References

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

Aghdam, R., Ganjali, M., Zhang, X., & Eslahchi, C. (2014). CN: A Consensus Algorithm for Inferring Gene Regulatory Networks Using SORDER Algorithm and Conditional Mutual Information Test. *Molecular bioSystems*. https://doi.org/10.1039/C4MB00413B

Allen, J. D., Xie, Y., Chen, M., Girard, L., & Xiao, G. (2012). Comparing statistical methods for constructing large scale gene networks. *PloS One*, *7*(1), e29348.

Allocco, D. J., Kohane, I. S., & Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, *5*, 18.

Alvar, J., Vélez, I. D., Bern, C., Herrero, M., Desjeux, P., Cano, J., … den Boer, M. (2012). Leishmaniasis worldwide and global estimates of its incidence. *PloS One*, *7*(5), e35671.

Amit, I., Garber, M., Chevrier, N., Leite, A. P., Donner, Y., Eisenhaure, T., … Regev, A. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* , *326*(5950), 257–263.

Andersson, B. (2010). The promise of T. cruzi genomics. *Nature*, (June).

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* , 1–4.

Andrade, L. O., & Andrews, N. W. (2005). The Trypanosoma cruzi-host-cell interplay: location, invasion, retention. *Nature Reviews. Microbiology*, *3*(10), 819–823.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., … Wang, H. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research*, *38*(Database issue), D457–62.

Babtie, A. C., Kirk, P., & Stumpf, M. P. H. (2014). Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences*, *2014*(14), 201414026.

Belcher, C. E., Drenkow, J., Kehoe, B., Gingeras, T. R., McNamara, N., Lemjabbar, H., … Relman, D. A. (2000). The transcriptional responses of respiratory epithelial cells to Bordetella pertussis reveal host defensive and pathogen counter-defensive strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(25), 13847–13852.

Bembom, O. (2017). seqLogo: Sequence logos for DNA sequence alignments.

Benabdellah, K., González-Rey, E., & González, A. (2007). Alternative trans-splicing of the Trypanosoma cruzi LYT1 gene transcript results in compartmental and functional switch for the encoded protein. *Molecular Microbiology*, *65*(6), 1559–1567.

Benz, C., Nilsson, D., Andersson, B., Clayton, C., & Guilbride, D. L. (2005). Messenger RNA processing sites in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, *143*(2), 125–134.

Bern, C. (2015). Chagas' Disease. *The New England Journal of Medicine*, *373*(19), 1882.

Bilbe, G. (2015). Infectious diseases. Overcoming neglect of kinetoplastid diseases. *Science*, *348*(6238), 974–976.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., … Ussery, D. W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & Integrative Genomics*, *6*(3), 165–185.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina

sequence data. *Bioinformatics* , *30*(15), 2114–2120.
Bolstad, B. M., Irizarry, R. A., Gautier, L., & Wu, Z. (2005). Preprocessing High-density Oligonucleotide Arrays. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 13–32). Springer New York.
Brehm, K., Carlton, J. M., & Hoffmann, K. F. (2012). Parasite genomics and post-genomic activities: 21st century resources for the parasite immunologist. *Parasite Immunology*, *34*(2-3), 47–49.
Brener, Z. (1971). Life cycle of Trypanosoma cruzi. *Revista Do Instituto de Medicina Tropical de Sao Paulo*, *13*(3), 171–178.
Butte, a. J., & Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 418–429.
Carlson, M. R. J., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, *7*, 40.
Choi, J., & El-Sayed, N. M. (2012). Functional genomics of trypanosomatids. *Parasite Immunology*, *34*(November 2011), 72–79.
Christensen, S. M., Dillon, L. A. L., Carvalho, L. P., Passos, S., Novais, F. O., Hughitt, V. K., … Mosser, D. M. (2016). Meta-transcriptome Profiling of the Human-Leishmania braziliensis Cutaneous Lesion. *PLoS Neglected Tropical Diseases*, *10*(9), e0004992.
Chuang, H.-Y., Hofree, M., & Ideker, T. (2010). A decade of systems biology. *Annual Review of Cell and Developmental Biology*, *26*, 721–744.
Clayton, C. (2013). The Regulation of Trypanosome Gene Expression by RNA-Binding Proteins. *PLoS Pathogens*, *9*(11), 9–12.
Clayton, C. E. (2014). Networks of gene expression regulation in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, *195*(2), 96–106.
Clayton, C., & Shapira, M. (2007). Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Molecular and Biochemical Parasitology*, *156*(2), 93–101.
Clayton, J. (2010). Chagas disease: pushing through the pipeline. *Nature*, *465*(7301), S12–5.
Cock, P. J. a., Antao, T., Chang, J. T., Chapman, B. a., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* , *25*(11), 1422–1423.
Committee, W. H. O. E. (2002). Control of Chagas disease. *World Health Organization Technical Report Series*, *905*, i–vi, 1–109, back cover.
Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13.
Coura, J. R., & Dias, J. C. P. (2009). Epidemiology, control and surveillance of Chagas disease: 100 years after its discovery. *Memorias Do Instituto Oswaldo Cruz*, *104 Suppl 1*, 31–40.
Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*. Retrieved from http://igraph.org
Daniels, J.-P., Gull, K., & Wickstead, B. (2010). Cell biology of the trypanosome genome. *Microbiology and Molecular Biology Reviews: MMBR*, *74*(4), 552–569.
De Gaudenzi, J. G., Carmona, S. J., Agüero, F., & Frasch, A. C. (2013). Genome-wide analysis of 3'-untranslated regions supports the existence of post-transcriptional regulons controlling gene expression in trypanosomes. *PeerJ*, *1*, e118.
de la Fuente, A. (2010). From "differential expression" to "differential networking" – identification of dysfunctional regulatory networks in diseases. *Trends in Genetics: TIG*, *26*(7), 326–333.
der Walt, S. van, Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for

Efficient Numerical Computation. *Computing in Science Engineering*, *13*(2), 22–30.
De Smet, R., & Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nature Reviews. Microbiology*, *8*(10), 717–729.
de Souza, W., de Carvalho, T. M. U., & Barrias, E. S. (2010). Review on Trypanosoma cruzi: Host Cell Interaction. *International Journal of Cell Biology*, *2010*. https://doi.org/10.1155/2010/295394
D'haeseleer, P., Liang, S., & Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* , *16*(8), 707–726.
Diehn, M., & Relman, D. a. (2001). Comparing functional genomic datasets: lessons from DNA microarray analyses of host–pathogen interactions. *Current Opinion in Microbiology*, *4*(1), 95–101.
Dillon, L. A. L., Okrah, K., Hughitt, V. K., Suresh, R., Li, Y., Fernandes, M. C., … El-Sayed, N. M. (2015). Transcriptomic profiling of gene expression and RNA processing during Leishmania major differentiation. *Nucleic Acids Research*, *43*(14), 6799–6813.
Dillon, L. A. L., Suresh, R., Okrah, K., Corrada Bravo, H., Mosser, D. M., & El-Sayed, N. M. (2015). Simultaneous transcriptional profiling of Leishmania major and its murine macrophage host cell reveals insights into host-pathogen interactions. *BMC Genomics*, *16*, 1108.
Di Noia, J. M., D'Orso, I., Sánchez, D. O., & Frasch, A. C. C. (2000). AU-rich Elements in the 3′-Untranslated Region of a New Mucin-type Gene Family of Trypanosoma cruzi Confers mRNA Instability and Modulates Translation Efficiency. *The Journal of Biological Chemistry*, *275*(14), 10218–10227.
D'Orso, I., De Gaudenzi, J. G., & Frasch, A. C. C. (2003). RNA-binding proteins and mRNA turnover in trypanosomes. *Trends in Parasitology*, *19*(4), 151–155.
Doyle, M. a., MacRae, J. I., De Souza, D. P., Saunders, E. C., McConville, M. J., & Likić, V. a. (2009). LeishCyc: a biochemical pathways database for Leishmania major. *BMC Systems Biology*, *3*, 57.
Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews. Genetics*, *14*(7), 496–506.
El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A.-N., … Andersson, B. (2005). *The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease* (Vol. 309, pp. 409–415).
El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., … Hall, N. (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science*, *309*(5733), 404–409.
Fadda, A., Ryten, M., Droll, D., Rojas, F., Färber, V., Haanstra, J. R., … Clayton, C. (2014). Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. *Molecular Microbiology*, *94*(2), 307–326.
Fernandes, M. C., Dillon, L. A. L., Belew, A. T., Bravo, H. C., Mosser, D. M., & El-Sayed, N. M. (2016). Dual Transcriptome Profiling of Leishmania-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. *mBio*, *7*(3). https://doi.org/10.1128/mBio.00027-16
Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2014). Stability indicators in network reconstruction. *PloS One*, *9*(2), e89815.
Forrester, S. J., & Hall, N. (2014). The revolution of whole genome sequencing to study parasites. *Molecular and Biochemical Parasitology*, *195*(2), 77–81.
Frazee, A. C., Langmead, B., & Leek, J. T. (2011). ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, *12*(1), 449.
Furger, A., Schürch, N., Kurath, U., & Roditi, I. (1997). Elements in the 3' untranslated region of procyclin mRNA regulate expression in insect forms of Trypanosoma brucei by modulating RNA stability and translation. *Molecular and Cellular Biology*, *17*(8), 4372–4380.

Gazestani, V. H., Lu, Z., & Salavati, R. (2014). Deciphering RNA regulatory elements in trypanosomatids: one piece at a time or genome-wide? *Trends in Parasitology*, 1–7.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y., & Kitano, H. (2011). Software for systems biology: from tools to integrated platforms. *Nature Reviews. Genetics*, *12*(12), 821–832.

Gibson, S. M., Ficklin, S. P., Isaacson, S., Luo, F., Feltus, F. A., & Smith, M. C. (2013). Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. *PloS One*, *8*(2). https://doi.org/10.1371/journal.pone.0055871

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., … Celniker, S. E. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature*, *471*(7339), 473–479.

Greif, G., Ponce de Leon, M., Lamolle, G., Rodriguez, M., Piñeyro, D., Tavares-Marques, L. M., … Alvarez-Valin, F. (2013). Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax. *BMC Genomics*, *14*, 149.

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

Hillier, L. W., Reinke, V., Green, P., Hirst, M., Marra, M. A., & Waterston, R. H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Research*, *19*(4), 657–666.

Hogg, J. R., & Goff, S. P. (2010). Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell*, *143*(3), 379–389.

Holland, B., & Moulton, V. (2003). Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees. In *Algorithms in Bioinformatics* (pp. 165–176). Springer, Berlin, Heidelberg.

Houston-Ludlam, G. A., Belew, A. T., & El-Sayed, N. M. (2016). Comparative Transcriptome Profiling of Human Foreskin Fibroblasts Infected with the Sylvio and Y Strains of Trypanosoma cruzi. *PloS One*, *11*(8), e0159197.

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

Hudson, N. J., Reverter, A., & Dalrymple, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Computational Biology*, *5*(5), e1000382.

Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z., & DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, *13*(3), 281–291.

Inbar, E., Hughitt, V. K., Dillon, L. A. L., Ghosh, K., El-Sayed, N. M., & Sacks, D. L. (2017). The Transcriptome of Leishmania major Developmental Stages in Their Natural Sand Fly Vector. *mBio*, *8*(2). https://doi.org/10.1128/mBio.00029-17

Jansen, R., Greenbaum, D., & Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Research*, *12*(1), 37–46.

Jenner, R. G., & Young, R. A. (2005). Insights into host responses against pathogens from transcriptional profiling. *Nature Reviews. Microbiology*, *3*(4), 281–294.

Ji, G., Guan, J., Zeng, Y., Li, Q. Q., & Wu, X. (2014). Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbu011

Kamburov, A., Stelzl, U., Lehrach, H., & Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Research*, *41*(D1), 793–800.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Khare, S., Nagle, A. S., Biggart, A., Lai, Y. H., Liang, F., Davis, L. C., … Supek, F. (2016). Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness. *Nature*, *537*(7619), 229–233.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36.

Kim, S.-K., Fouts, A. E., & Boothroyd, J. C. (2007). Toxoplasma gondii dysregulates IFN-gamma-inducible gene expression in human fibroblasts: insights from a genome-wide transcriptional profiling. *Journal of Immunology* , *178*(8), 5154–5165.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38.

Kolev, N. G., Franklin, J. B., Carmi, S., Shi, H., Michaeli, S., & Tschudi, C. (2010). The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. *PLoS Pathogens*, *6*(9), e1001090.

Kolev, N. G., Ullu, E., & Tschudi, C. (2014). The emerging role of RNA-binding proteins in the life cycle of Trypanosoma brucei. *Cellular Microbiology*, *16*(4), 482–489.

König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Irelan, J. T., … Chanda, S. K. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, *135*(1), 49–60.

Kramer, S. (2012). Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. *Molecular and Biochemical Parasitology*, *181*(2), 61–72.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

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

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* , *24*(5), 719–720.

LeBowitz, J. H., Smith, H. Q., Rusche, L., & Beverley, S. M. (1993). Coupling of poly(A) site selection and trans-splicing in Leishmania. *Genes & Development*, *7*(6), 996–1007.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* , *28*(6), 882–883.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., … Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, *11*(10), 733–739.

Legros, D., Ollivier, G., Gastellu-Etchegorry, M., Paquet, C., Burri, C., Jannin, J., & Büscher, P. (2002). Treatment of human African trypanosomiasis—present situation and needs for research and development. *The Lancet Infectious Diseases*, *2*(7), 437–440.

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–21.

Liang, Y.-H., Cai, B., Chen, F., Wang, G., Wang, M., Zhong, Y., & Cheng, Z.-M. (max). (2014). Construction and validation of a gene co-expression network in grapevine (Vitis vinifera. L.). *Horticulture Research*, *1*(April), 14040.

Li, C., Bankhead, A., Eisfeld, A. J., Hatta, Y., Jeng, S., Chang, J. H., … Kawaoka, Y. (2011). Host regulatory network response to infection with highly pathogenic H5N1 avian influenza virus. *Journal of Virology*, *85*(21), 10955–10967.

Li, W., Freudenberg, J., & Oswald, M. (2015). Principles for the organization of gene-sets. *Computational Biology and Chemistry*, *59 Pt B*, 139–149.

Li, Y., Shah-Simpson, S., Okrah, K., Belew, A. T., Choi, J., Caradonna, K. L., … Burleigh, B. A.

(2016). Transcriptome Remodeling in Trypanosoma cruzi and Human Cells during Intracellular Infection. *PLoS Pathogens*, *12*(4), e1005511.

López-Kleine, L., Leal, L., & López, C. (2013). Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. *Briefings in Functional Genomics*, *12*(5), 457–467.

Maertzdorf, J., Repsilber, D., Parida, S. K., Stanley, K., Roberts, T., Black, G., … Kaufmann, S. H. E. (2011). Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes and Immunity*, *12*(1), 15–22.

Mandlik, A., Livny, J., Robins, W. P., Ritchie, J. M., Mekalanos, J. J., & Waldor, M. K. (2011). RNA-Seq-based monitoring of infection-linked changes in Vibrio cholerae gene expression. *Cell Host & Microbe*, *10*(2), 165–174.

Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S. D., … Kim, J. K. (2010). The landscape of C. elegans 3'UTRs. *Science* , *329*(5990), 432–435.

Manning-Cela, R., González, A., & Swindle, J. (2002). Alternative splicing of LYT1 transcripts in Trypanosoma cruzi. *Infection and Immunity*, *70*(8), 4726–4728.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016a). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, *13*(4), 366–370.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016b). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods*, *13*(4), 366–370.

Maretti-Mira, A. C., Bittner, J., Oliveira-Neto, M. P., Liu, M., Kang, D., Li, H., … Craft, N. (2012). Transcriptome patterns from primary cutaneous Leishmania braziliensis infections associate with eventual development of mucosal disease in humans. *PLoS Neglected Tropical Diseases*, *6*(9), e1816.

Margolin, A. a., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7 Suppl 1*, S7.

Martínez-Calvillo, S., Vizuet-de-Rueda, J. C., Florencio-Martínez, L. E., Manning-Cela, R. G., & Figueroa-Angulo, E. E. (2010). Gene expression in trypanosomatid parasites. *Journal of Biomedicine & Biotechnology*, *2010*, 1–15.

Mason, M. J., Fan, G., Plath, K., Zhou, Q., & Horvath, S. (2009). Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, *10*, 327.

Matthews, K. R., Tschudi, C., & Ullu, E. (1994). A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. *Genes & Development*, *8*(4), 491–501.

McCarthy-Burke, C., Taylor, Z. a., & Buck, G. a. (1989). Characterization of the spliced leader genes and transcripts in Trypanosoma cruzi. *Gene*, *82*(1), 177–189.

McGettigan, P. a. (2013). Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology*, *17*(1), 4–11.

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

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, *91*(3), 243–248.

Miles, M. A., Feliciangeli, M. D., & de Arias, A. R. (2003). American trypanosomiasis (Chagas' disease) and the role of molecular epidemiology in guiding control strategies. *BMJ* , *326*(7404), 1444–1448.

Minning, T. a., Weatherly, D. B., Atwood, J., Orlando, R., & Tarleton, R. L. (2009). The steady-state transcriptome of the four major life-cycle stages of Trypanosoma cruzi. *BMC Genomics*, *10*, 370.

Mishima, Y., & Tomari, Y. (2016). Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell*, *61*(6), 874–885.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews. Genetics*, *14*(10), 719–732.

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

Myler, P. J., & Fasel, N. (2008). *Leishmania: after the genome*. Horizon Scientific Press.

Ng, A., Bursteinas, B., Gao, Q., Mollison, E., & Zvelebil, M. (2006). Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Briefings in Bioinformatics*, *7*(4), 318–330.

Nozaki, T., & Cross, G. a. (1995). Effects of 3' untranslated and intergenic regions on gene expression in Trypanosoma cruzi. *Molecular and Biochemical Parasitology*, *75*(1), 55–67.

Nylén, S., & Gautam, S. (2010). Immunological perspectives of leishmaniasis. *Journal of Global Infectious Diseases*, *2*(2), 135–146.

Palenchar, J. B., & Bellofatto, V. (2006). Gene transcription in trypanosomes. *Molecular and Biochemical Parasitology*, *146*(2), 135–141.

Parsana, P., Ruberman, C., Jaffe, A. E., Schatz, M. C., Battle, A., & Leek, J. T. (2017, October 13). *Addressing confounding artifacts in reconstruction of gene co-expression networks*. *bioRxiv*. https://doi.org/10.1101/202903

Peña, I., Pilar Manzano, M., Cantizani, J., Kessler, A., Alonso-Padilla, J., Bardera, A. I., … Julio Martin, J. (2015). New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource. *Scientific Reports*, *5*, 8771.

Petersen, T. N., Brunak, S. R., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786.

Pierleoni, A., Martelli, P., & Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*, *9*(1), 392.

Quigley, J., Hughitt, V. K., Velikovsky, C. A., Mariuzza, R. A., El-Sayed, N. M., & Briken, V. (2017). The Cell Wall Lipid PDIM Contributes to Phagosomal Escape and Host Cell Exit of Mycobacterium tuberculosis. *mBio*, *8*(2). https://doi.org/10.1128/mBio.00148-17

Rassi, A., & Marin-Neto, J. A. (2010). Chagas disease. *The Lancet*, *375*(9723), 1388–1402.

Rodrigues, D. C., Silva, R., Rondinelli, E., & Urményi, T. P. (2010). Trypanosoma cruzi: modulation of HSP70 mRNA stability by untranslated regions during heat shock. *Experimental Parasitology*, *126*(2), 245–253.

Romaniuk, M. A. (2016). Regulation of RNA binding proteins in trypanosomatid protozoan parasites. *World Journal of Biological Chemistry*, *7*(1), 146.

Saha, A., Kim, Y., Gewirtz, A. D. H., Jo, B., Gao, C., McDowell, I. C., … Battle. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, *27*(11), 1843–1858.

Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, *11*(1), 22–24.

Santos, D. O., Coutinho, C. E. R., Madeira, M. F., Bottino, C. G., Vieira, R. T., Nascimento, S. B., … Castro, H. C. (2008). Leishmaniasis treatment--a challenge that remains: a review. *Parasitology Research*, *103*(1), 1–10.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864.

Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M., & Ligterink, W. (2016). Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science*, *7*, 444.

Siegel, T. N., Gunasekera, K., Cross, G. a. M., & Ochsenreiter, T. (2011). Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing. *Trends in Parasitology*, *27*(10), 434–441.

Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S., & Cross, G. a. M. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. *Nucleic Acids Research*, *38*(15), 4946–4957.

Siegel, T. N., Tan, K. S. W., & Cross, G. A. M. (2005). Systematic study of sequence motifs for RNA trans splicing in Trypanosoma brucei. *Molecular and Cellular Biology*, *25*(21), 9586–9594.

Simpson, A. G. B., Stevens, J. R., & Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends in Parasitology*, *22*(4), 168–174.

Smith, C. W., Porro, E. B., Patton, J. G., & Nadal-Ginard, B. (1989). Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. *Nature*, *342*(6247), 243–247.

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*(1), 91.

Song, L., Langfelder, P., & Horvath, S. (2012). *Comparison of co-expression measures: mutual information, correlation, and model based indices* (Vol. 13, p. 328).

Sonnhammer, E. L., von Heijne, G., & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *6*, 175–182.

Stich, A., Barrett, M. P., & Krishna, S. (2003). Waking up to sleeping sickness. *Trends in Parasitology*, *19*(5), 195–197.

Stich, A., Ponte-Sucre, A., & Holzgrabe, U. (2013). Do we need new drugs against human African trypanosomiasis? *The Lancet Infectious Diseases*, *13*(9), 733–734.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, *302*(5643), 249–255.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* , *25*(9), 1105–1111.

van Nas, A., Guhathakurta, D., Wang, S. S., Yehya, N., Horvath, S., Zhang, B., … Lusis, A. J. (2009). Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks. *Endocrinology*, *150*(3), 1235–1249.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Weatherly, D. B., Boehlke, C., & Tarleton, R. L. (2009). Chromosome level assembly of the hybrid Trypanosoma cruzi genome. *BMC Genomics*, *10*, 255.

Westermann, A. J., Gorski, S. a., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews. Microbiology*, *10*(9), 618–630.

WHO. (2017, March 21). Trypanosomiasis, human African (sleeping sickness). Retrieved November 20, 2017, from http://www.who.int/mediacentre/factsheets/fs259/en/

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, *11*(2), R14.

Yu, Z., Wong, H.-S., & Wang, H. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics* , *23*(21), 2888–2896.

Zhai, Y., Franco, L. M., Atmar, R. L., Quarles, J. M., Arden, N., Bucasas, K. L., … Couch, R. B. (2015). Host Transcriptional Response to Influenza and Other Acute Respiratory Viral Infections. A Prospective Cohort Study. *PLoS Pathogens*, *11*(6), 1–29.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), Article17.

Zickenrott, S., Angarica, V. E., Upadhyaya, B. B., & del Sol, A. (2016). Prediction of disease-gene-drug relationships following a differential network analysis. *Cell Death & Disease*, *7*, e2040.