MINING THE TOXOPLASMA GONDII GENOME FOR MICRORNA REGULATORY PATTERNS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Biotechnology

by
İlhan Erkin ACAR

June 2017
İZMİR

We approve the thesis of **İlhan Erkin ACAR**

Examining Committee Members:

Assoc. Prof. Dr. Jens ALLMER
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Prof. Dr. Anne Frary
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Prof. Dr. Bahattin TANYOLAÇ
Department of Bioengineering, Ege University

30 June 2017

Assoc. Prof. Dr. Jens ALLMER
Supervisor, Department of Molecular Biology and Genetics
İzmir Institute of Technology

Assoc. Prof. Engin ÖZÇİVİCİ
Head of the Department of
Biotechnology and Bioengineering

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I thank my advisor, Jens Allmer, for giving me the opportunity to work in bioinformatics field. He has been a great advisor who changed my perspective about many things, may it be sports or science. He has always been supportive and I have learnt a lot from him.

I also thank my committee members, Prof. Dr. Anne Frary and Prof. Dr. Bahattin Tanyolaç, for their time and advices towards betterment of this thesis.

I am thankful to all of my friends for their understanding to my busy schedule. Also, I would like to express my gratitude towards my friends from the laboratory. They have shown great patience towards my lack of bioinformatics knowledge when I first joined them, and my endless bad jokes. I am truly thankful to them. Special thanks to Nihan Atak for her most precious support, especially during the preparation of this thesis. She has always motivated me and kept me going.

Thanks to TÜBİTAK for financial support via project 113E326.

Last but not least, I thank my family, who always encouraged me to follow my dreams. Thanks to them, today I am working in a field that I love.

ABSTRACT

MINING THE TOXOPLASMA GONDII GENOME FOR MICRORNA REGULATORY PATTERNS

Toxoplasma gondii is a parasite that causes mental retardation, blindness or near-blindness, and decreased psycho-motor performance if the patient is congenitally infected. There have been efforts to vaccinate humans against this parasite, yet it was not achieved. Therefore, a better understanding of *Toxoplasma gondii* can be provided by examining its microRNA regulation.

MicroRNAs are known to regulate messenger RNAs and prevent translation. This results in different effects in different biological pathways. In this study, the *Toxoplasma gondii* genome was used to predict precursor and mature microRNAs, while experimentally validated microRNAs were taken into consideration. This was further explored in terms of microRNA targeting, with the known genes of *Toxoplasma gondii*. Furthermore, RNA Sequencing data of this organism was obtained and analysed in terms of gene expression and possible microRNA expression outcomes. Combining gene expression analyses with targeting predictions, it was possible to create a microRNA - gene interaction network.

Gene expression analyses showed that there was no differentially expressed genes, microRNAs or interactions between two developmental stages of *Toxoplasma gondii*, tachyzoite and bradyzoite. This result was added to interactions to determine up and down regulations. Then, all of these interactions were connected where they intersect, to create a regulation network of microRNAs.

This network was further explored and compared to random networks of the same size. It was seen that the biological network contains many larger sized cliques. This knowledge can be further analysed in future work, to create drug leads that will target vital pathways of *Toxoplasma gondii*.

ÖZET

TOXOPLASMA GONDII GENOMUNDAN DÜZENLEYİCİ MİKRORNA ŞABLONLARININ ÇIKARILMASI

Toxoplasma gondii, doğuştan aktarıldığında zeka geriliği, körlük ya da psikomotor performansında düşüşlere sebebiyet vermektedir. İnsanları bu parazitten korumak adına aşı çalışmaları yapılmış, fakat başarılı olunamamıştır. Bu yüzden, mikroRNA düzenlemeleriyle *Toxoplasma gondii*'yle ilgili daha çok bilgi sağlanması amaçlanmıştır.

MikroRNAların, haberci RNA'ları düzenleyerek protein oluşturmalarını engelledikleri bilinmektedir. Bu düzenleme, farklı biyolojik yolaklarda farklı etkilere sebebiyet vermiştir. Bu çalışmada, *Toxoplasma gondii* genomu, deneysel olarak doğrulanmış mikroRNA'ları da göz önünde bulundurularak, öncü ve olgun mikroRNA'ların tahmininde kullanılmıştır. Bu, *Toxoplasma gondii*'nin bilinen genleriyle, mikroRNA hedeflemesi açısından da incelen-miştir. Bu organizmanın RNA dizileme verisi, gen ve olası mikroRNA ifadeleri incelenmek üzere elde edilmiş ve analiz edilmiştir. Gen ifadesi analizi, hedefleme tahminleriyle birleştirilerek mikroRNA - gen etkileşimleri çıkartılmıştır.

Gen ifadesi analizi, *Toxoplasma gondii*'nin iki gelişme safhası olan tachyzoite ve bradyzoite arasında gen, mikroRNA veya etkileşim ifade farklılıklarının olmadığını göstermiştir. Bu sonuçlar etkileşimlere eklenerek yukarı ve aşağı düzenlemeler belirlenmiştir. Sonrasında bu etkileşimler, kesişim noktalarından bağlanarak mikroRNA düzenleyici ağı oluşturulmuştur.

Bu ağ, aynı büyüklükteki rastlantısal ağlarla karşılaştırılmıştır. Biyolojik ağın daha büyük kliklere sahip olduğu görülmüştür. Bu bilgi gelecek çalışmalarla, *Toxoplasma gondii*'nin hayati yolaklarını hedefleyecek ilaç öncüleri yaratılması için incelenebilir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

LIST OF ABBREVIATIONS

miRNA . . . . . microRNA
nt . . . . . nucleotides
3'UTR . . . . . Three prime untranslated region
DGCR8 . . . . . DiGeorge critical region 8
pre-miRNA . . . . . precursor microRNA
ago . . . . . Argonaute
RISC . . . . . RNA-Induced Silencing Complex
T. gondii . . . . . *Toxoplasma gondii*
CDC . . . . . Centers for Disease Control and Prevention
NHGRI . . . . . National Human Genome Research Institute
NGS . . . . . Next-Generation Sequencing
SOLiD . . . . . Sequencing by Oligo Ligation and Detection
RNA-Seq . . . . . RNA Sequencing
SRA . . . . . Sequence Read Archive
AWS . . . . . Amazon Web Services
MCCV . . . . . Monte-Carlo Cross Validation
RPKM . . . . . Reads per kilobase per million mapped reads
FPKM . . . . . Fragments per kilobase per million mapped reads
NKMN . . . . . Nucleotides per kilobase of transcript per million nucleotides mapped
l2fc . . . . . Log_2 Fold Change

CHAPTER 1

INTRODUCTION

Advances in sequencing technologies led to rapid and cheap sequencing of genomes. What cost around $100,000,000 in 2001, costs $1,000 today (Wetterstrand, 2016). Due to this decrease, it is estimated that genomic sequence data in 2025 will amount to 20 times of total size of all the videos combined on YouTube by 2025 (Stephens et al., 2015). Therefore, available data for scientific pursuits only increase, which creates more opportunity to analyse different biological phenomena.

MicroRNAs, which are small regulatory units, gained popularity with new sequencing technologies (Eminaga et al., 2013). On the other hand, *Toxoplasma gondii*, even though studied immensely in its early years of discovery (Dubey, 2008), fell to the list of five neglected parasites (CDC, 2017).

In this study, *Toxoplasma gondii* related sequencing data was explored in terms of microRNA regulation. Computational methods were employed to analyse and predict regulation caused by microRNAs in this parasite, and to establish a complete microRNA regulation network within confidence levels. In order to introduce terms that were part of this study, microRNAs, *Toxoplasma gondii*, and next-generation sequencing are summarized in the following sub-sections.

1.1. MicroRNAs

MicroRNAs were the medium to form the regulation network in this study. Hence, information about microRNAs is summarized in three subsections.

1.1.1. History and Roles

MicroRNAs (miRNAs) are 18 to 24 nucleotide (nt) long, non-coding RNAs (Bartel, 2009). Since the discovery of miRNAs, many researchers started studying these little RNAs. MiRNAs were first discovered in *C.elegans* in 1993 (Lee et al., 1993). Back then,

they were not named as miRNAs, they were only described as anti-sense RNA-RNA interaction. Regardless, it was found out that this small RNAs bind to three prime untranslated regions (3'UTRs) of specific messenger RNAs (mRNAs) (Lee et al., 1993).

Nowadays, miRNAs are known to regulate many mRNAs either by degrading them or repressing their translation (Ha and Kim, 2014). More than 60% of human mRNAs are estimated to be controlled by miRNAs (Friedman et al., 2009). This kind of extensive regulation by miRNAs results in them affecting biological roles such as differences in development, cell signalling, apoptosis, and immune responses (Tüfekci et al., 2014). Since biological processes are controlled by miRNAs, it is not hard to associate them with many different consequences such as cancer (Farazi et al., 2013), cardiovascular diseases (Romaine et al., 2015), inflammatory responses (Thounaojam et al., 2014), neurodegenerative diseases (Abe and Bonini, 2013) and autoimmune diseases (Saito et al., 2014).

1.1.2. Biogenesis

The biogenesis of miRNAs differs between various organisms (Millar and Waterhouse, 2005; Axtell et al., 2011), yet some of the structures in this process remain the same (Figure 1.1). First, primary miRNAs (pri-miRNAs) are mainly transcribed with RNA polymerase II (Lee et al., 2003), with some cases seen with RNA polymerase III employment (Borchert et al., 2006). These pri-miRNAs are further processed by a microprocessor that contains Drosha, an RNase III enzyme, and DGCR8 (DiGeorge critical region 8) in humans, or Pasha in invertebrates to create precursor miRNA (pre-miRNA) in the nucleus (Xie and Steitz, 2014; Wahid et al., 2010). However, no homologs of Drosha or its cofactors were found in plants which indicates this step may be absent (Wahid et al., 2010). Pre-miRNAs are transported out of the nucleus by exportin-5 (Yi et al., 2003). In plants, however, this transportation occurs with a homologue of exportin-5, namely HASTY (Bollman et al., 2003). In animals, transported pre-miRNAs are further cut by Dicer, an endonuclease cytoplasmic RNase III, to create mature miRNAs (Ketting et al., 2001). Then these mature miRNAs are loaded on Argonaute (ago) protein and create the RNA-Induced Silencing Complex (RISC) (Carmell et al., 2002). This procedure is, again, different in plants as they lack Dicer (Reinhart et al., 2002). Plants use Dicer-like proteins to create mature miRNAs and these mature miRNAs are reported to be created in nucleus

(Papp et al., 2003). Then RISC loading occurs to finalize the synthesis.

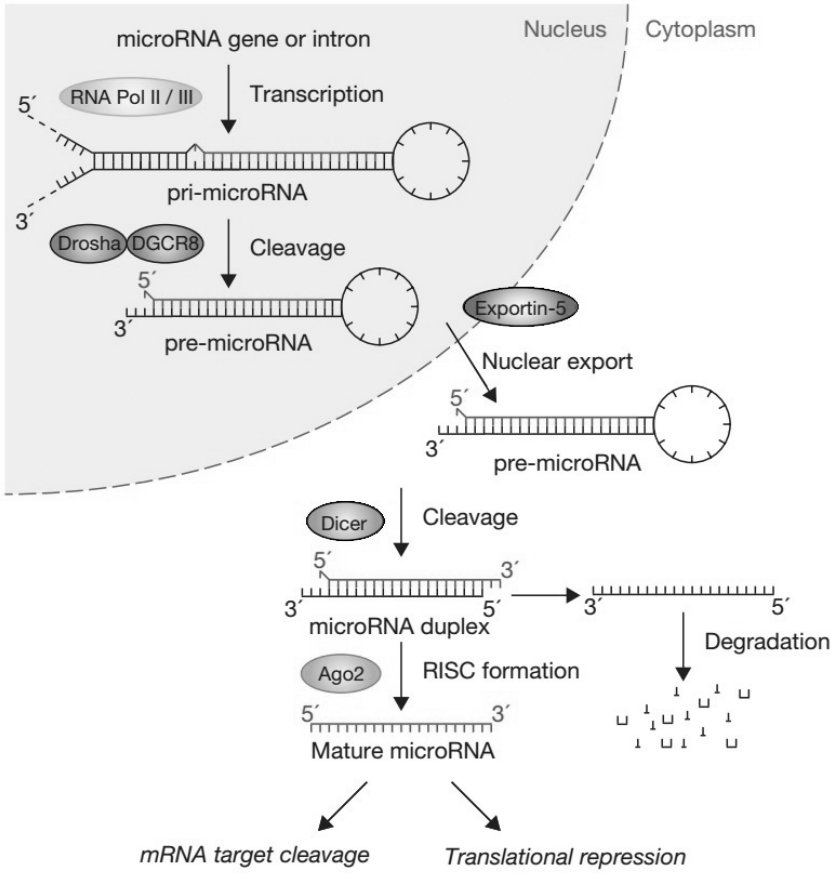

Figure 1.1. General biogenesis pathway of miRNA. Primary structure is transcribed by either RNA polymerase II or III, then it is cleaved into pre-miRNA structure by Drosha, if present in the organism. This pre-miRNA structure is then exported out of nucleus by Exportin-5, or by HASTY in plants. Dicer or Dicer-like proteins cleave pre-miRNA structures into mature miRNAs and mature miRNAs form RNA-Induced Silencing Complex (RISC) with Argonaute (Ago) proetins. This image was taken from another study, and edited for simplicity. (Source: Winter et al. (2009))

1.1.3. Genomic Locations

MiRNAs are reported to be synthesized from intergenic regions as well as intronic regions (Lau et al., 2001; Lee et al., 2003). In this respect, miRNAs that are produced from different genomic locations can be classified in three different groups, which are,

intergenic miRNAs with their own promoters, intronic miRNAs that are synthesized from introns, and exonic miRNAs that overlap with exons (Olena and Patton, 2009).

As was explained, mature miRNAs come from pre-miRNAs. Generally, mature miRNAs are known to come from one of the arms (3' or 5') of the hairpin shaped pre-miRNA structure (Du, 2005), and in some cases from both arms (Glazov et al., 2008). However, it was shown that some miRNAs may come from loop regions of pre-miRNAs as well (Winter et al., 2013). In this study, miRNAs that may be present in the loop regions were also taken into consideration.

1.2. *Toxoplasma gondii*

Toxoplasma gondii (*T. gondii*) was first identified in 1908 by Nicolle and Manceaux in a hamster-like tissue (Dubey, 2008). Even though there have been many studies about *T. gondii*, and its pathogenic effect, toxoplasmosis, today *T. gondii* is identified as one of the five neglected parasitic infections by Centers for Disease Control and Prevention (CDC) (CDC, 2017). Around one third of the human population worldwide is estimated to be chronically infected by *T. gondii* (Liu et al., 2015). *T. gondii* causes toxoplasmosis in people with congenital infection, which was diagnosed to cause mental retardation, blindness or near-blindness, and decreased psycho-motor performance in early studies (McCabe and Remington, 1988). However recent studies show that *T. gondii* infection may actually cause even more symptoms (Table 1.1) (Flegr et al., 2014).

T. gondii is an infectious parasite that uses felids as definitive hosts and other warm blooded animals as intermediates (Dubey, 2004). Transmission can happen congenitally, in fecal - oral route, or via undercooked / raw meat consumption (Dubey, 2008). Usually, cats ingest *T. gondii* which is in one of the three infectious stages: tachyzoite, which is a form in which *T. gondii* quickly multiplies itself; bradyzoite, which is a dormant state where *T. gondii* remains in a cyst in infected tissues; or sporozoites, which are oocysts and shed in feces (Dubey, 1996; Dubey et al., 1998). Human infection may occur via horizontal transmission in bradyzoite and sporozoite stages, whereas in tachyzoite stage, vertical transmission happens (Figure 1.2) (Tenter et al., 2000). There have been vaccination studies for *T. gondii*, however protection from this parasite in humans was not achieved (Jongert et al., 2009).

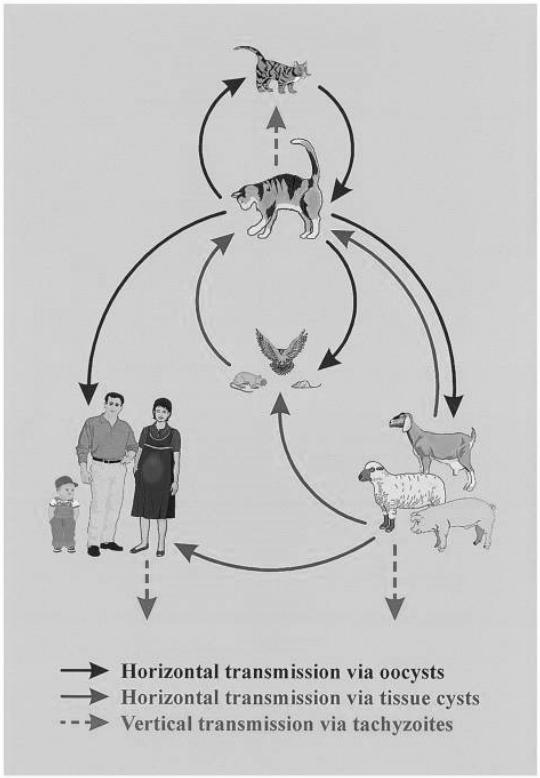

Figure 1.2. Transmission of *T. gondii*. *T. gondii* has 3 infectious stages in its life cycle. While in tachyzoite stage, it can only be transmitted to off springs. However, in bradyzoite (tissue cysts) and sporozoite (oocysts) stages, it can be transferred between species. (Source: Tenter et al. (2000))

1.3. Next-Generation Sequencing

Sequencing is the method to determine the exact order of bases in a DNA or RNA. Sequencing method was standardized by Edward Sanger in 1977 with the method, currently named, Sanger sequencing (Sanger et al., 1977). With the completion of human genome project (Venter et al., 2001), this sequencing method was improved by many different companies with the set goal of reducing the cost of the human genome sequencing to $1000 by National Human Genome Research Institute (NHGRI) (Schloss, 2008). This challenge led to massively parallel sequencing (also known as high-throughput sequencing) methods which is called Next-Generation Sequencing (NGS) (Grada and Weinbrecht, 2013).

NGS platforms are used for various studies today, such as, variant discovery,

Table 1.1. List of reported *T. gondii* related symptoms. References to symptoms can be found in the source study. (Source: Flegr et al. (2014))

Disease/Clinical Entity	
Hearing loss	Ulcerative colitis
Psychosis; schizophrenia; bipolar disorder	Crohn's disease
Mood disorders; suicide; depression (?)	Abdominal hernia
Obsessive - compulsive disorder	Hepatitis, including HCV infection
Attention/concentration deficit hyperactivity disorder	Granulomatous liver disease
Anorexia	Liver cirrhosis; granulomatous liver disease; impaired liver function
Autism spectrum disorders	Primary biliary cirrhosis; biliary atresia; cholestatic disorders
Down's syndrome	Diabeter mellitus type 1 and 2
Alzheimer's disease	Goitre; iodine deficiency
Parkinson's disease	Hashimoto's thyroiditis
Migraine; other headaches	Graves' disease; throid adenoma
Idiopathic intracranial hypertension	Rheumatoid arthritis; Still's disease
Pseudotumor cerebri	Polymyositis
Aseptic meningitis	Systemic sclerosis
Mollaret meningitis	Systemic lupus erythematosus
Epilepsy	Wegener's granulomatosis; other vasculitides
Aphasia and apilepsy (Landau - Kleffner syndrome	Hypothalamo-pituitary dysfunction; panhypopituitarism
Facial nerve palsy (Bell's palsy)	Cryoglobulibemia
Central diabetes insipidus; syndrome of inappropriate antidiuretic hormone secretion	Ocular toxoplasmosis (retinochorioiditis; uveitis; blurred vision; floaters; macular scars; nystagmus; strabismus; reduced visual acuity; blindness; scleritis; papillitis; retinal necrosis; vasculitis; retinal detachment; vitritis; congenital cataract; neuroretinitis; atrophic optic papilla; retinitis pigmentosa)
Breast cancer	Glaucoma
Anti-phospholipid syndrome	Ovarian dysfunction
Non-Hodgkin's lymphoma	Uterine atrophy
Brain tumors (meningioma; ependymoma; glioma)	Impaired reproductive function (T.gondii was present in testicles, epididymis, seminal vesicles, prostate gland in rams, and caused abnormalities in sperm motility, viability and concentration rates, weight of epididymis in rats, orchitis)
Neoplasia	Neprotic syndrome; lipoid nephrosis
Melanoma	Schönlein - Henoch purpura
Congenital toxoplasmosis (encephalitis; chorloretinitis; neonatal mortality)	Glomerulonephritis (various forms; including these with development of fibrosis); impaired kidney function
Carcinoma of female genitalia, including cervical tissue	Atherosclerosis; obesity; cardiovascular deaths; all-cause mortality
Chronic heart failure; myocarditis; arrhythmia	Diverse abnormalities in aggregate personality; including aggresive behaviour in animals and humans
Inflammatory bowel disease	

sequencing of transcripts of an organism, and profiling genome-wide epigenetic marks (Metzker, 2010). Most of the sequencing platforms employ a different sequencing method, and for the task at hand, different sequence preparation method is used (Goodwin et al., 2016). Also, the output from these applications (Figure 1.3), which are fragments of original DNA or RNA (reads), vary in length and output presentation type among methods (Liu et al., 2012). This leads to many different procedures for different topics to explore. Hence, only the sequencing of transcripts, which is called RNA-Seq, in Applied Biosciences SOLiD (Sequencing by Oligo Ligation and Detection) will be introduced, because of the RNA-Seq data that were used in this study.

1.3.1. RNA Sequencing

To perform RNA Sequencing (RNA-Seq), RNAs should be isolated from source cell. Then, depending on the focus of the research, selection of these RNAs takes place (Kukurba and Montgomery, 2015). This is called library preparation and there are various designs to prepare a library for different purposes as can be seen in Table 1.2. This procedure is then followed by conversion to complementary DNA, amplification of these complementary DNAs with sequencing platform specific adapters, and using the sequencing method depending on the sequencing platform with the prepared library (Figure 1.4).

Table 1.2. Library design for RNA-Seq. In the library preparation step of the RNA-Seq process, an RNA library is created with different criterion for different study foci. (Source: Kukurba and Montgomery (2015))

Library Design	**Usage**	**Description**
Poly-A selection	Sequencing mRNA	Selects for RNA species with poly-A tail and enriches for mRNA
Ribo-depletion	Sequencing mRNA, pre-mRNA, ncRNA	Removes ribosomal RNA and enriches for mRNA, pre-mRNA, and ncRNA
Size selection	Sequencing miRNA	Selects RNA species using size fractionation by gel electrophoresis
Duplex-specific nuclease	Reduce highly abundant transcripts	Cleaves highly abundant transcripts, including rRNA and other highly expressed genes
Strand-specific	De novo transcriptome assembly	Preserves strand information of the transcript
Multiplexed	Sequencing multiple samples together	Genetic barcoding method that enables sequencing multiple samples together
Short-read	Higher coverage	Produces 50-100 bp reads; generally higher read coverage and reduced error rate compared to long-read sequencing
Long-read	De novo transcriptome assembly	Produces >1000 bp reads; advantageous for resolving splice junctions and repetitive regions

RNA-seq can be done with different sequencing platforms (Adiconis et al., 2013; Goodwin et al., 2016), however the SOLiD sequencing method is reported to have 99.85% accuracy in sequencing, after a filtration process (Liu et al., 2012). This high accuracy is due to two base encoding of SOLiD sequencing. This sequencing procedure starts with SOLiD specific adapters in prepared library, binding to the universal sequencing primer, which is complementary to the adapters. Then these primers are elongated from their 3' end by 8mer oligonucleotides with fluorescent label on the 4th and 5th positions, that matches the RNA fragment. These fluorescent labels are predetermined for different bases and are screened while sequencing occurs. 8mers are cleaved between the 5th and 6th position and 6th through 8th bases are washed away. After washing, another matching

8mer is ligated (which would give the information of the 9th and 10th base this time) and this process continues until the 25th base knowledge is acquired. After this, a second round of sequencing starts with -1 starting position of the primer (which would give 3rd and 4th base in the first elongation this time, 8th and 9th in second and so on). These rounds are repeated until whole sequence is known with the help of fluorescent screening. Except for the first and last bases, each base gets sequenced twice, which reduces errors in sequencing. Details of this procedure can be found in the study of Mardis (2008). At the ending of sequencing, a sequence file is produced with color encoded values (Ondov et al., 2008). This type of encoding is called color-space, and instead of base characters, color codes are given along with their quality scores (Figure 1.3).

```
@SRR1542923.13
T2330320021013222202123031312002020010113212O
+
KGKKFF:GGE6DGGF:IHHHGHIIHGEJJEGIIJGHGJJGEBDG
@SRR1542923.15
T10122001100310302320102102233011132222021312101321 3
+
KJJJJJJKKJJJKKJJKIIKKJJJJJJFFJJJJJJIIIIIIIJJFH:FF@
@SRR1542923.17
T331010333111301033330023013313131212003101 23
+
JIIIIIIKKGIJEEJJJIIIJJFFJJJIIJJIGGIIFIGFFEFB
@SRR1542923.19
T13203300002021010130310103020111203000000102 02
+
JKKKKKKKKKKKKKKKKKKKKJJKKKKKKKJIIJJGJJBGDEBEE
@SRR1542923.21
T13220303311103300223313002102011232132212310121122
+
KKKKKKKKKKKKKKKKKKKKKKKKKKIKKIHKKKKKKJJKKIJJIIAIG@
```

Figure 1.3. Example of color-space sequencing output. In the output file, each sequenced fragment (read) consists of 4 lines. Starting with '@' and '+' are the identifiers of the read, where '+' line may be empty for different sequencing platforms. 2nd line of each read contains sequence information. In color-space however, only first base is given in actual character, rest are encoded with corresponding color code. 4th line of a read contains sequencing quality score of the read, which can be 94 different characters from '!' to '~' (from 33rd American Standard Code for Information Interchange [ASCII] code to 126th one).

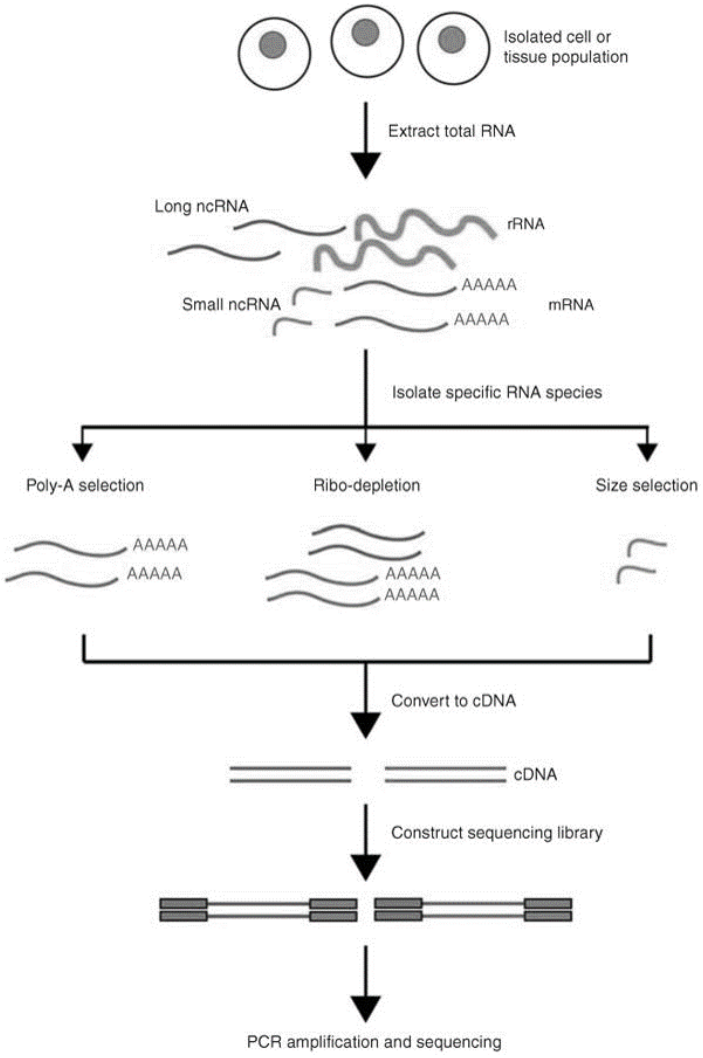

Figure 1.4. RNA-Seq. To sequence RNAs, first all of the RNAs in a cell or a tissue are extracted. Then, these RNAs are selected according to focus of the study that will be conducted. After selection, cDNAs are created from these RNAs and they are amplified for sequencing. (Source: Kukurba and Montgomery (2015))

1.4. Machine Learning

Machine learning is a computational approach which employs various algorithms onto previous data, 'learns' from it, and creates a mathematical model out of properties (called 'features') of these data (Baştanlar and Özuysal, Baştanlar and Özuysal). Basi-

cally, there are two approaches to achieve these kind of models with machine learning, which are unsupervised learning and supervised learning (Sætrom and Snøve, 2007). Unsupervised learning is employed when there is little to no prior knowledge about the data (D'haeseleer, 2005). On the other hand, supervised learning requires labelled data, which would allow algorithms to create a model that will classify unlabelled data depending on the features of these labeled groups (Libbrecht and Noble, 2015).

In this study, unsupervised learning was used in gene expression analysis, to determine relation in the gene expression data. Supervised learning was also used, with one of the most popular algorithms called Random Forests (Breiman, 2001), to predict mature miRNAs from pre-miRNA structures. Other predictions involved in this study were made with tools (Dai and Zhao, 2011; Allmer and Saçar Demirci, 2016), hence their algorithms were not mentioned.

1.5. Aim

Although there have been many studies since the 1900s, *T. gondii* is still not understood well enough to protect humans effectively from infection. With NGS, it is possible to get more data about genes of an organism which helps towards more complete understanding of different biological pathways. As miRNAs are known to regulate genes, it is crucial to understand miRNA roles in *T. gondii* for effective protection. The aim of this study was to create a gene expression analysis supported miRNA regulatory network from publicly available RNA-Seq data, which will help towards understanding regulation patterns of *T. gondii*.

CHAPTER 2

METHODOLOGY

Due to the computational nature of the study, the methodology is split into subsections to provide less complicated, step-by-step explanations.

2.1. Data

The reference genome for *T. gondii* (ToxoDB-25_TgondiiME49) and its annotation file of known transcripts and genes were downloaded from toxodb.org (Gajria et al., 2007) to match the extracted hairpins from previous work (Saçar Demirci et al., 2016) which used the same genome file. To filter our data from possible contamination, human reference genome (*Homo sapiens*, GRCh38) was downloaded from Ensembl (Herrero et al., 2016). For gene expression analysis, RNA-Seq data that contains different strains (Croken et al., 2014) were downloaded from Sequence Read Archive (SRA) (Leinonen et al., 2011). Samples contained three different strains of *T. gondii* (CTG, PLK and RH); SRR1542919-24 belonged to RH strain, SRR1542925-30 belonged to PLK strain and rest belonged to CTG. Each strain contained two different developmental stage (tachyzoite and bradyzoite). Developmental stage samples were equal in number for each strain and first half of the samples belonged to tachyzoite whereas last half was bradyzoite (e.g. SRR1542919-21 were tachyzoite RH samples and SRR1542922-24 were bradyzoite RH samples)

There were already 339 mature miRNA sequences available and described in another study (Wang et al., 2012). These mature miRNAs were obtained from Supplementary File 7 of the said study, to train a prediction model.

2.2. MicroRNA Detection

Since there are only a number of validated miRNAs of *T. gondii*, computational predictions were made to increase possible undiscovered miRNAs. This prediction was

done in three steps, according to the biogenesis stages of miRNAs.

2.2.1. Pre-MicroRNA Detection

For analysis, hairpins from previous work (Saçar Demirci et al., 2016) were obtained, which were created by computationally fragmenting and folding the reference genome (ToxoDB-25_TgondiiME49). Instead of using websites that provide feature calculation services (Yones et al., 2015; Bağcı and Allmer, 2016), a Java code was written to employ Amazon Web Services (AWS) for calculation. izMiR Framework (Allmer and Saçar Demirci, 2016) was used to create *T. gondii* specific model, at 1000 fold Monte-Carlo Cross Validation (MCCV) (Xu and Liang, 2001) by using 70% of the hairpin data for training and 30% for testing. From the previous work (Saçar Demirci et al., 2016), pre-miRNAs containing 292 of the 339 known mature miRNAs (Wang et al., 2012) were used as positive samples. Missing known mature miRNAs (47 in all) were extended by 50 nt to both directions on the reference genome and hairpins of these mature miRNAs were added to positive samples, resulting in a total of 683 pre-miRNAs. Pseudo pre-miRNAs (Ng and Mishra, 2007) were used as negative data.

2.2.2. Mature MicroRNA Detection

There was only a small number of known mature miRNAs for *T. gondii* (339), so a general mature miRNA prediction model was created by using all mature miRNAs listed in miRTarBase (Release 6.0) (Chou et al., 2016) which resulted in 4316 mature miRNA sequences available in miRBase (Kozomara and Griffiths-Jones, 2014). A negative data set was created as was proposed in another study (Gkirtzou et al., 2010); by shifting mature sequences by half of their length within the hairpin sequences. To describe mature miRNAs, 101 features were calculated such as: start and end positions of mature sequence (2), central loop start and end points (2), hairpin length, miRBase hairpin length, stem length, mature length, maximum loop length (5), number of matches and mismatches in the mature sequence region (2), single nucleotide counts (4), dinucleotide counts (16), trinucleotide counts (64), distances of start and end positions to 3', 5', loop start and loop end (6). Random forest machine learning algorithm was used to train a model with 70%

learning - 30% testing data in 1000 fold MCCV. From these 1000 fold MCCV (Figure 2.1), the model with the highest accuracy (0.932) was chosen to apply to predicted pre-miRNAs.

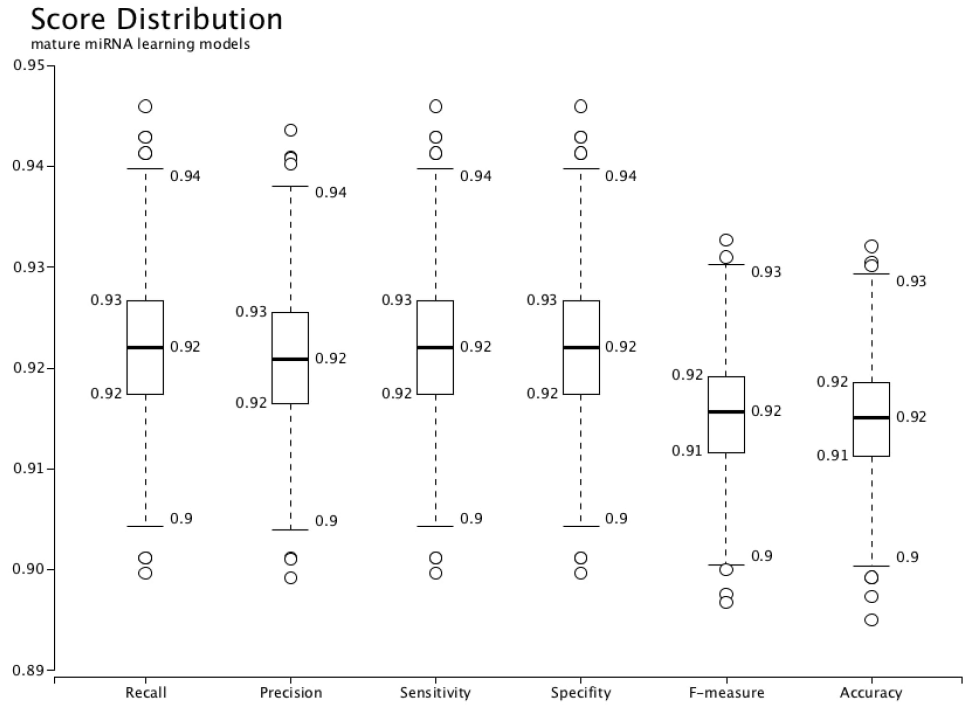

Figure 2.1. Score distributions of 1000 machine learned models established using 1000 fold Monte Carlo cross validation.

2.2.3. MicroRNA Targeting

For targeting predictions, psRNATarget 2011 release (Dai and Zhao, 2011) was used in default settings. All of the genes that were described for *T. gondii* were extracted from the reference genome and they were used as target sites, since UTRs are not yet established.

2.3. Expression Anaylsis

Expression analysis was done for both genes and miRNAs, because of the regulatory interaction of these two. Then, whole interactions were explored in terms of expression analysis, to deduce differences between strains and developmental stages in a broader perspective.

2.3.1. Gene Expression

Downloaded RNA-seq data were first cleaned from their adapter sequences using Cutadapt tool (Martin, 2011). Then, low quality regions were trimmed from the data using Sickle tool (Joshi and Fass, 2011). Quality trimming was done with the threshold of 30 quality score. If the reads were smaller than 30 nt long after trimming, they were discarded (Table 2.1). FastQC tool (Andrews, 2010) was used to check quality levels before and after trimming. Resulting clean reads were mapped onto the human reference genome (GRCh38) using Tophat v1.4.1 (Trapnell et al., 2009). This mapping application was done in order to filter out possible human contamination in the reads, as *T. gondii* strains were reported to be grown in human foreskin fibroblast cell lines in the source study (Croken et al., 2014). Because of the colorspace characteristics of samples, a large amount of reads was filtered and an older version of Tophat had to be used instead of current version v2.1.1. A java script was developed to count mapped reads sorted by their locations with the help of the downloaded annotation file for *T. gondii*. Sorting of reads was done on KNIME Analytics Platform (Berthold et al., 2009). Genes that had fewer than 5 reads mapped onto them were considered as not expressed and filtered out. Mapped nucleotide normalization method was adopted (see 2.3.4) to normalize mapped read counts. Normalized counts were further filtered by their expression among their strains and developmental stages. Genes that were not expressed in at least 70% of the samples in their respective strains or developmental stages were filtered out.

Table 2.1. Read statistics. For all 18 samples downloaded from SRA, pre-processed and processed read numbers can be seen.

Samples	Raw Reads	Cleaned Reads	Mean Clean Read Length	Deleted Reads (%)	Reads Mapped on Human (%)	Toxo Mapped Reads (%)
SRR1542919	58,730,137	2,077,159	37.15	96.46	1.05	84.49
SRR1542920	43,055,026	11,581,307	44.23	73.10	0.91	72.09
SRR1542921	44,958,415	14,617,199	53.89	67.49	0.70	75.73
SRR1542922	43,264,535	1,375,655	37.31	96.82	0.75	85.54
SRR1542923	53,074,533	13,476,636	44.13	74.61	1.01	69.58
SRR1542924	55,077,053	18,329,562	55.04	66.72	0.77	76.15
SRR1542925	36,994,224	1,423,290	37.57	96.15	3.27	88.33
SRR1542926	48,595,529	12,544,189	44.20	74.19	1.25	77.00
SRR1542927	55,934,799	17,482,709	53.82	68.74	1.12	81.83
SRR1542928	74,716,539	2,727,669	37.70	96.35	3.20	89.18
SRR1542929	51,517,301	12,437,480	44.01	75.86	2.61	79.16
SRR1542930	41,089,401	13,451,657	53.60	67.26	1.42	83.09
SRR1542931	211,425,021	7,886,857	37.41	96.27	1.89	87.22
SRR1542932	44,043,513	9,678,918	43.69	78.02	2.06	68.16
SRR1542933	248,076,128	80,172,404	54.23	67.68	1.88	78.25
SRR1542934	51,790,061	1,704,048	37.07	96.71	2.26	83.78
SRR1542935	55,535,624	14,076,802	44.19	74.65	4.72	71.03
SRR1542936	38,718,995	13,031,003	54.12	66.34	3.74	75.64
Mean	69,810,935	13,781,919	45.19	79.64	1.92	79.24

2.3.2. MicroRNA Expression

A custom annotation file was created from the strand and genomic location knowledge of predicted pre-miRNAs in this study. With this annotation file, it was possible to ascertain the amount of miRNAs expressed by employing the steps defined in 2.3.1. Due to less mapping to miRNA regions, the mapped read count threshold was lowered to 2 from 5. MiRNAs that had less than 2 reads mapped onto them were filtered out. As was mentioned in 2.3.1, if a miRNA was not expressed in at least 70% of the samples in their respective strains or developmental stages, it was filtered out.

2.3.3. MicroRNA - mRNA Interactions

In this study, an interaction is defined as a miRNA co-expressed with at least one of its target mRNAs in the same sample. MicroRNAs often originate from genes and these source genes can be used to extend the interaction to the gene level by associating them with metabolic or regulatory pathways. Thus a complete interaction is defined by the source gene, miRNA, and its target(s). MicroRNAs that did not come from a known gene were filtered out in this study.

In order to establish a regulatory network and to further analyse the interactions within *T. gondii*, Cytoscape v3.4.0 (Shannon et al., 2003) was used. Interactions were built with source genes and target genes represented as nodes and miRNAs as edges. A network was built with only the expressed genes and miRNAs (after they were filtered by their respective criteria) if they showed expression in at least one sample. In the created network, total amount of interaction of a gene with other genes was used to determine the node size. Total expression amount (total of all samples, in NKMN) was used to color the genes with a color gradient. In this gradient, yellow color showed the lowest amount of expression and red showed highest. There were many nodes between expression amount of 1000 to 5000 NKMN. To increase the differentiation between these nodes by color, two additional breaks were added to color gradient, turquoise and dark blue, on 1000 and 5000 NKMN.

Normalized interaction ratios (see 2.3.4) were presented in $\log_2$ scale to show ratios with higher target gene expression as positive values and higher source gene expressions as negative. Edge colors were set to show this interaction values, where red colors show negative values and green showing positives. Regardless of their sign, edge width was set to get thinner as it gets closer to zero. Cliques were detected using the Cytoscape plug-in MClique (v1.2). Since extensive numbers of cliques were found they were compared to randomized networks using the Network Randomizer (v1.1.2) plug-in in Cytoscape. Ten random networks were created using the same nodes and edges of the microRNA network. MClique was, again, used to detect cliques in the randomized networks.

2.3.4. Normalization

Generally, in RNA-Seq related studies, reads per kilobase per million mapped reads (RPKM) or in paired end samples, fragments per kilobase per million mapped reads (FPKM) are calculated to normalize the expression values. However, our data showed differences even before quality and adapter trimming processes. Therefore, an approach to normalize according to actually mapped nucleotide number was employed. With the written Java script, it was possible to save the length of mapped reads while counting them. Knowledge of lengths of all genes and miRNAs was available within respective annotation files, and FastQC tool (Andrews, 2010) provided total amount of nucleotides

in each sample. Using these, nucleotides per kilobase of transcript per million nucleotides mapped (NKMN) was applied.

NKMN method was calculated by total mapped nucleotides per gene or miRNA, divided by total nucleotide number of corresponding gene or miRNA and total nucleotides in the sample. Since the ratio was aimed to show per kilobase per million, it was multiplied by a billion to bring the values into an intuitive range with the following formula.

$$\frac{\sum(\text{Mapped Nucleotides on Gene or miRNA})}{\sum(\text{Nucleotides of Gene or miRNA}) \times \sum(\text{Nucleotides in Sample})} \times 10^9$$

For the interactions, an interaction ratio was calculated. Since the lengths of reads were changing among samples and this affected miRNA expression counts and their normalization, expression of their source genes was used to represent their abundance in interactions. Therefore, this ratio was calculated by dividing target gene expression amount by source gene expression amount for the interaction. Then, these ratios were normalized by the median value of interaction ratios of the sample to which they belonged. After this, the median value of all of the median values of samples were taken, and normalization is further extended by dividing all of the ratios by this final median value.

2.3.5. Differential Expression Analysis

R platform (Team, 2016) was used to determine differences in expression, as well as interaction ratios. NKMN values between different strains (RH vs. PLK, RH vs. CTG, and PLK vs. CTG) and between developmental stages (Tachyzoite vs. Bradyzoite) were converted to $\log_2$ fold changes. Student's t-test was performed for each gene, miRNA and interaction among different strains and developmental stages. P values were obtained from these test and adjusted according to Benjamini-Hochberg (Benjamini and Hochberg, 1995). Thresholds were chosen as 0.05 for p-value, and 2 for $\log_2$ fold changes.

2.4. Annotation of Genes and MicroRNAs

The annotation file of *T. gondii* contained gene annotation and protein names from these genes. However, gene annotation was cryptic (e.g. gene TGME49_293600) so pro-

tein product names were used where possible. In cases where a protein was synthesized from multiple genes, gene accession numbers (e.g. 293600) were added to protein names (e.g. RPL27_293600). For genes that did not have known protein products, BLAST (v2.4.0+) (Camacho et al., 2009) was used to align these genes with genes of other *T. gondii* strains. Similarity above 75% (with mismatch + gap <4) in these alignments was accepted to be used as new annotation. Genes that did not fulfill any of these conditions were left with their gene annotation from the annotation file.

Predicted pre-miRNAs were initially annotated with numbers only, starting from one to the total number of predicted miRNAs. Then after filtering pre-miRNAs according to criteria described in 2.2.1, this annotation was extended to their source strand (Pos for positive and Neg for negative), prediction number and the chromosome to which they belonged (e.g. Neg_263687_TGME49_chrII). Further annotation was done by aligning all mature miRNAs in miRBase (Kozomara and Griffiths-Jones, 2014) and the previously described 339 *T.gondii* miRNAs (Wang et al., 2012) to our predicted pre-miRNAs using BLAST (v2.4.0+) in blastn-short mode. In alignments with above 75% similarity (with mismatch <4), miRNA names were kept as new names to our predicted miRNAs with addition of our prediction number to be able to track the miRNA back when required (e.g. tgo-novel-12-9_21502). Those that did not fulfill this criterion were kept unchanged. To be able to differentiate mature miRNAs coming from same pre-miRNA, an identifier number was attached to annotations (e.g. Neg_263687_TGME49_chrII_2).

CHAPTER 3

RESULTS AND DISCUSSION

Due to the nature of this study, expression results, and differences between these expressions are discussed separately for miRNAs, genes and interactions, in the light of obtained results. Then a regulation network was created based on expression and miRNA targeting analysis.

3.1. MicroRNA Detection

In a previous study (Saçar Demirci et al., 2016) the *Toxoplasma gondii* ME49 genome was folded and hairpin like structures (approximately 5 million) were extracted. Since many of these hairpins are unlikely to be pre-miRNAs, a machine learning model using izMiR framework (Allmer and Saçar Demirci, 2016) was trained and used for the assessment of the putative hairpins. In total, 1,227,917 pre-miRNAs were predicted from these hairpin structures and these were filtered by their confidence scores (>0.99) using the izMiR model. This filtering resulted in 4,589 confident pre-miRNAs. These pre-miRNAs were further checked whether they are part of a gene or not, and intergenic pre-miRNAs were not taken into account. About 300 candidate pre-miRNAs were affected by this filtering leaving 4,240 hairpins for further analyses. Expression with at least 2 mapped reads in at least one of the samples was the final requirement for pre-miRNAs and 2,484 passed this filtering step. Pre-miRNAs were further processed into mature miRNAs in the miRNA genesis pathway and this process was mimicked by fragmenting the 4,589 confident candidate hairpins into 24 nt long sequences with 6 nt overlaps. Since the length of mature sequences is generally smaller or equal to 24 nt, the majority of mature sequences should be in the generated candidate pool with this approach. This pool was too large for further analysis, therefore, a machine learning model was established to discriminate the candidates. Mature candidates with a minimum of 15 nt long sequences and a model prediction score of at least 1.0 were used for further analysis. A total of 4,234 mature sequences passed the filtering by the machine learned model. 973 of the mature sequences overlapped with the hairpin loop and were removed while 89 included

the complete loop and were retained as loop-miRs (Winter et al., 2013). 1,058 mature miRNAs are located on the 5' arm while 2,114 are located on the 3' one. The confident pre-miRNA candidates were compiled into a genome feature format file to enable their expression analysis using standard workflows.

3.2. Gene Expression

For the expression analysis of *T. gondii* genes and miRNAs, a set of RNA-seq samples was acquired. Pre-processing of the downloaded RNA-Seq samples produced varying read lengths (Table 2.1). This difference in length needed to be taken into account during normalization and nucleotide mapping rate rather than read or transcript mapping

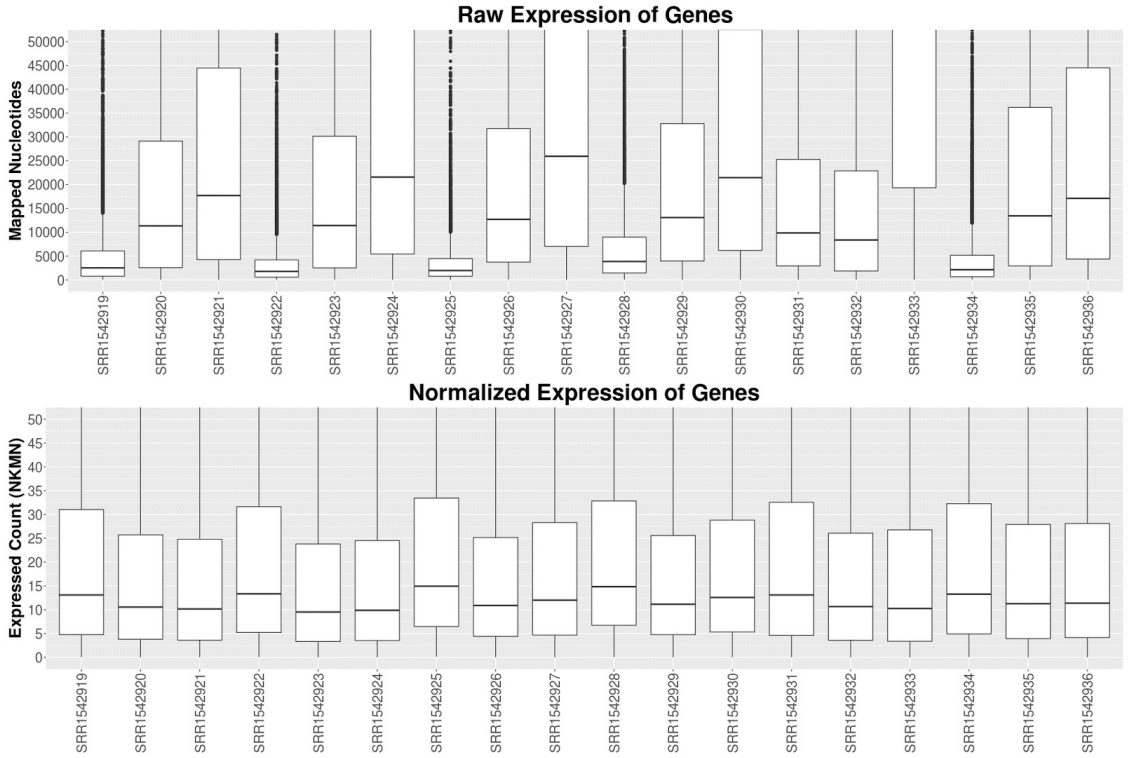

Figure 3.1. Distribution of normalized genes. The values shown on the y-axis are the resulting numbers from the formula presented in normalization method. Normalization was done for each gene in each sample. Each sample was shown with its accession number in the x-axis. The distribution of normalized mapped nucleotides were found to have closer median values than raw counts.

rates was considered. After normalization, the general distribution of gene expression appeared similar among samples despite variation in average read length indicating the effectiveness of the NKMN normalization approach (Figure 3.1).

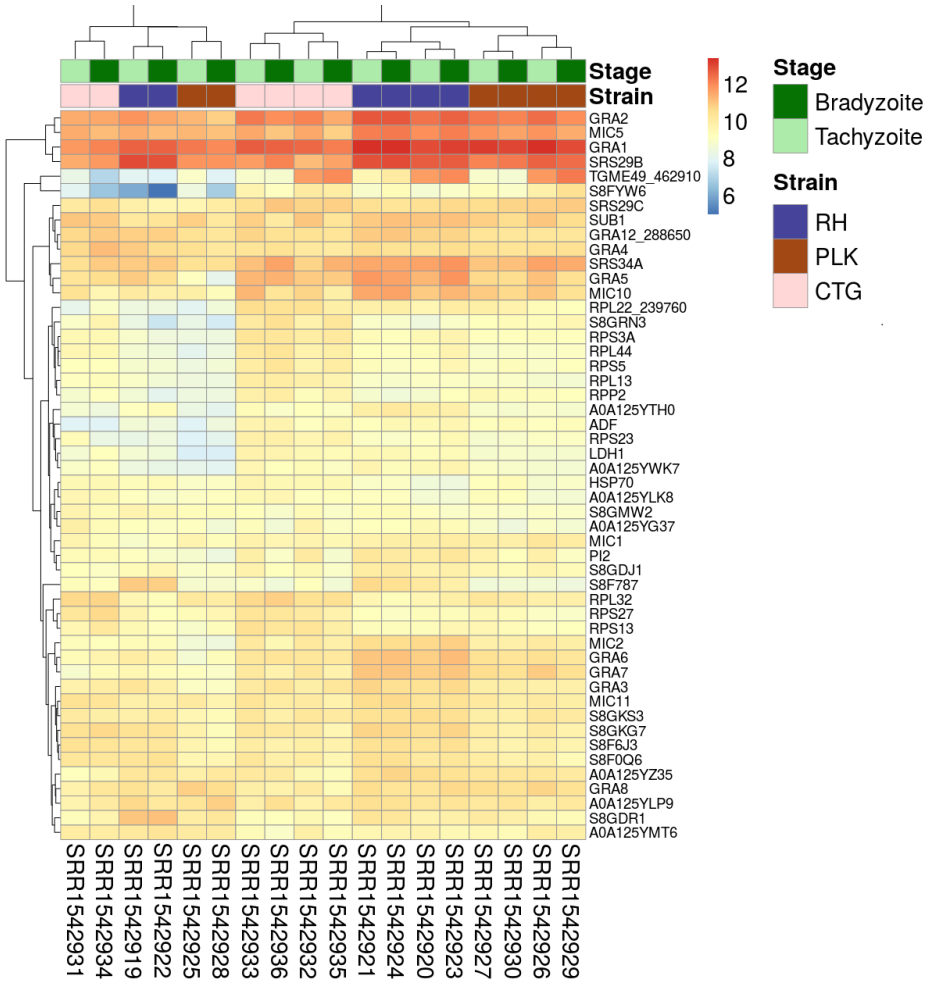

Figure 3.2. Heatmap showing the 50 genes with largest average expression among samples. The strains (CTG: pink, PLK, green, and RH: red) and developmental stages (bradyzoite: olive and tachyzoite: blue) can be seen on top of the genes. Gene identifiers are provided on a per row basis on the right and sample accessions are provided below the heatmap. Rows and columns have been clustered and expression amount is plotted in $\log_2$ scale using the pheatmap package in R.

This normalized expression among samples was compared and the most expressed 50 genes (top 50 of the average of all 18 samples) are presented in Figure 3.2. The process of picking the genes most expressed on average identifies the genes that are similarly

expressed in all samples. This is confirmed by the heatmap in Figure 3.2. There is no significant expression difference between developmental stages of *T. gondii* among the most expressed 50 genes. Samples from strains clustered together which reveals that the highly expressed genes in all samples are more uniformly expressed on a per strain basis than on a per developmental stage basis. Unfortunately, 6 samples presented slightly different behaviour (SRR15429[19,22,25,28,31,34]). Investigation into the origin of this revealed that the outliers have the shortest average read length and the largest percentage of deleted reads after pre-processing (Table 2.1). Even though there were differences in read lengths and amount of the reads in these outlier samples, it was seen that both development stages of each strain had almost the same expression amount among itself (Figure 3.2). This indifference in gene expressions among development stage was seen to be not influenced by read length or number differences. Furthermore, in both cases cluster analysis shows that strains PLK and RH have more similar gene expression among the top 50 genes.

3.3. Differential Gene Expression

Differential expression analysis was done in R using the NKMN normalized gene expression and employing t-test with Benjamini-Hochberg correction. Only genes expressed in at least 70% of the samples were considered for differential expression analysis. Out of a total of 8,920 annotated genes in *T. gondii*, 7,834 genes in strain RH, 8,047 genes in strain PLK, and 7,853 genes in strain CTG passed the 70% criteria. For the developmental stages 7,949 genes (tachyzoite) and 7,954 genes (bradyzoite) were available for differential expression analysis after filtering. For the comparison between stages and strains, only these expressed genes were taken into account, which resulted in a further decrease of comparable genes: 7,790 genes (RH vs. PLK), 7,679 genes (RH vs. CTG), 7,781 genes (PLK vs. CTG), and 7,863 genes (tachyzoite vs. bradyzoite). The $\log_2$ transformed distribution of differential expression among strains and stages is displayed in Figure 3.3.

The distribution of differential expression is least for tachyzoites vs. bradyzoites (Figure 3.3). It is also similar for RH vs. CTG and PLK while quite different for PLK vs. CTG (Figure 3.3). This further confirms the finding that RH and PLK are closer related in respect to their expressed genes than CTG (Figure 3.2). Calculated $\log_2$ fold changes

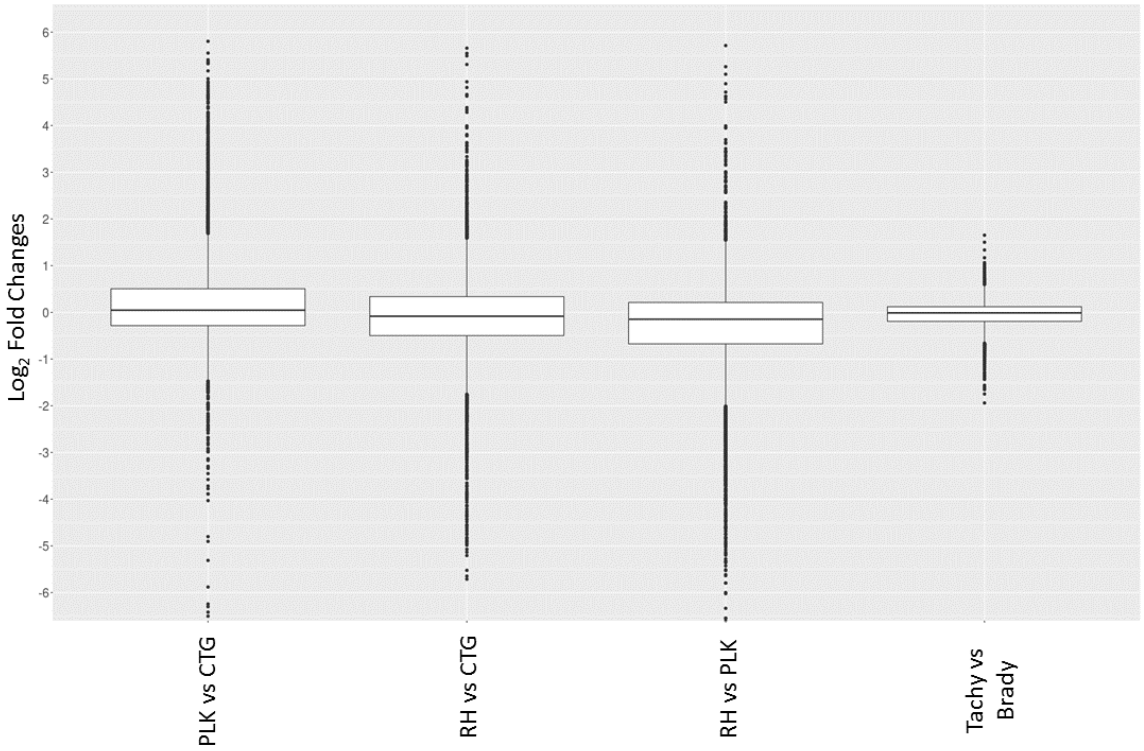

Figure 3.3. Distribution of differential gene expression ($\log_2$ transformed) between strains and developmental stages.

and adjusted p-values were used to further filter the genes for these comparisons. The selected significance threshold for p value was <0.05 whereas for the $\log_2$ fold change (l2fc), gene expressions with log fold $<$-2 or $>$2 were chosen. With these thresholds, 529 genes in the PLK vs. CTG comparison were found to be differentially expressed, whereas differentially expressed genes for RH vs. CTG amounted to 328 and for RH vs PLK to 613. There was no significantly differentially expressed gene for the comparison between tachyzoite and bradyzoite stages.

For each pair of strains, the five most differentially expressed genes per strain were chosen (Figure 3.4, Appendix A.1). Only for RH vs. CTG, the differential expression clusters stages while for PLK vs. CTG and RH vs. PLK the stages do not cluster at all. Over-expressed genes in RH are not as strongly over-expressed as for CTG and PLK. As expected for developmental stages, which did not have any significantly differentially expressed genes, the heatmap (Figure 3.4, bottom right) does not display any clustering

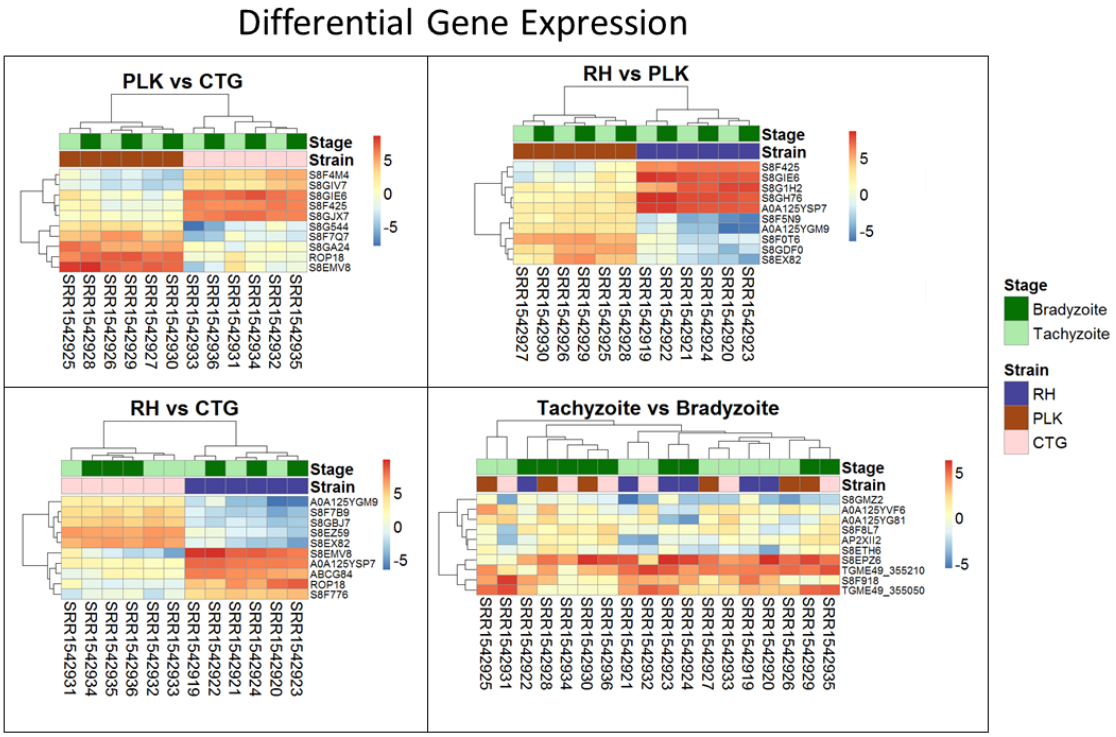

Figure 3.4. Heatmaps showing the 5 most over-expressed genes per strain (RH: blue, PLK: brown, CTG: pink) and developmental stage (bradyzoite: dark green, tachyzoite: mint). Each pair presents 10 genes of which 5 are over-expressed in one of the strains/stages and 5 in the other. Genes are hierarchically clustered based on the expression among replicates. Over-expression was analysed with pooled replicates, but for a better overview, all measurements are shown in columns including their hierarchical clustering. Actual values of these maps and chosen genes can be seen in Appendix A.1.

for strains or developmental stage.

3.4. MicroRNA Expression

For the expression analysis of miRNAs, pre-miRNAs were compiled into an annotation file for usage with standard expression analysis workflows. The same NKMN normalization was applied to miRNA expression analysis. Due to aforementioned differences in read lengths, raw miRNA counts varied greatly among samples even though normalization was performed and despite the normalization being effective for genes (Figure

3.5). Therefore, miRNA normalization was not found to be effective. It is hypothesized that perhaps mature miRNAs are more likely to be sampled by shorter reads which is further confirmed by the lower mapping ratio of samples with longer reads (Table 2.1). While different reads were hypothesized to be needed, expression analysis continued to be done with RNA-seq data to have a general idea about *T. gondii* miRNA expression.

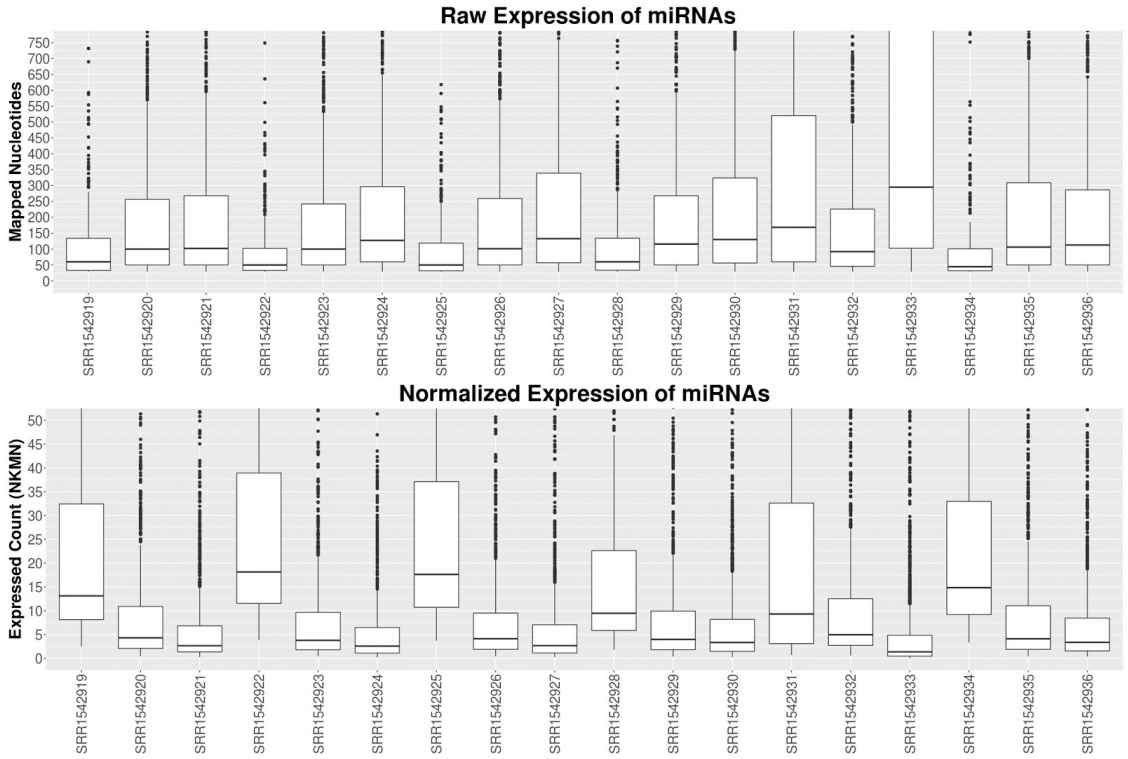

Figure 3.5. Distribution of normalized miRNA expressions. Normalization method that was employed to genes were applied to miRNA expressions. It was seen that median values were varying between samples but closer among similar mean read lengths.

In a similar fashion to gene expression analysis, the on average most expressed 50 miRNAs were identified for a general idea of expression among miRNAs and samples (Figure 3.6). A similar picture emerges for miRNAs as for genes with the exception that CTG and PLK are closer related in terms of expression in general. The same samples which were outliers for genes (overall less expression, Figure 3.2) show the opposite behaviour for miRNAs (overall more expression, Figure 3.6). Also, these samples confirm the closer relationship between RH and PLK seen for genes. Similar to gene expression, neither the development stage nor the strain show significant overall differences in ex-

pression for the 50 most expressed genes (on average). It is noteworthy, that among the most expressed 50 miRNAs, 43 were novel miRNAs predicted in this study whereas only 7 of them show high similarity with known miRNAs (Figure 3.6).

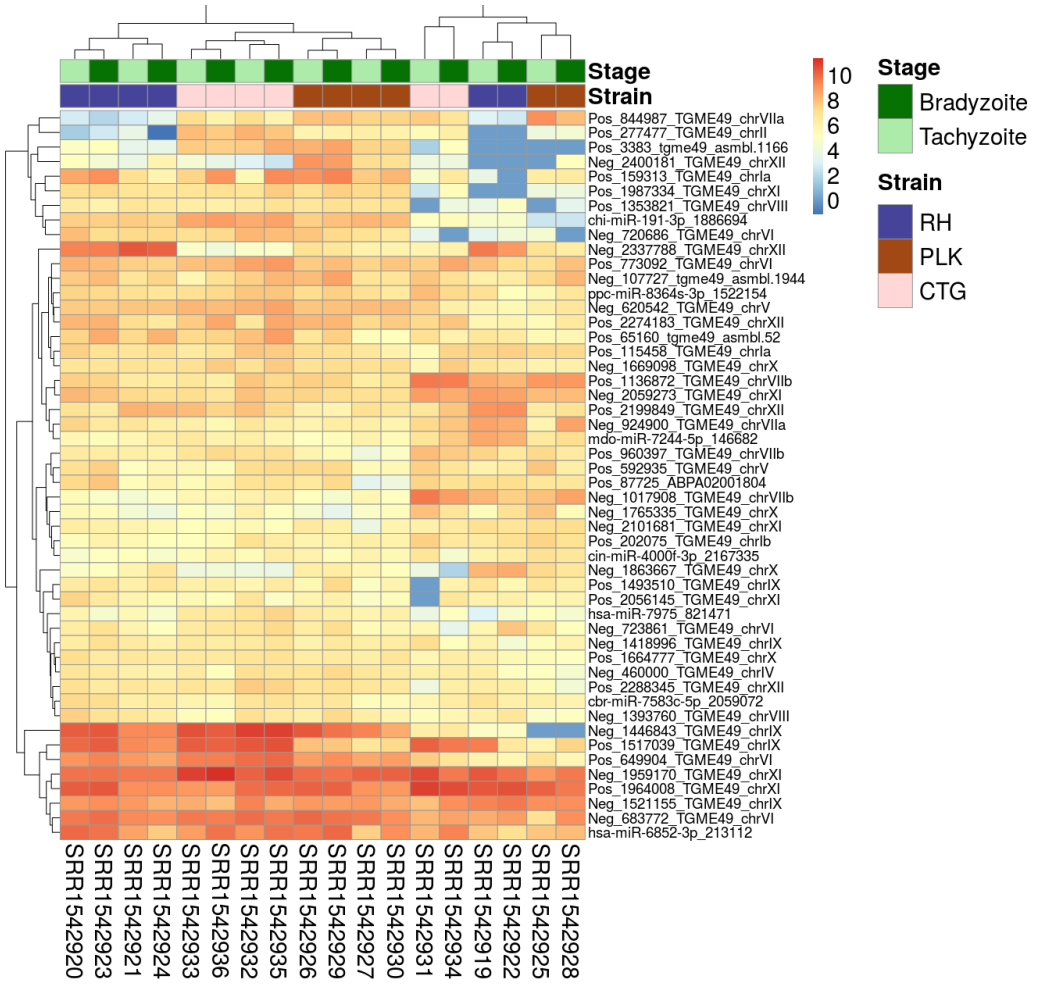

Figure 3.6. Heatmap showing the 50 miRNAs with largest average expression among samples. The strains (CTG: pink, PLK, green, and RH: red) and developmental stages (bradyzoite: olive and tachyzoite: blue) can be seen on top of the genes. Gene identifiers are provided on a per row basis on the right and sample accessions are provided below the heatmap. Rows and columns have been clustered and expression amount is plotted in log2 scale using the pheatmap package in R.

3.5. Differential MicroRNA Expression

From the initial 1,227,917 predicted miRNAs with a model confidence score >0.99, 4,589 miRNAs remained for further analysis. Similar to what was done during gene differential expression analysis, miRNAs were required to be expressed in at least 70% of the samples. For the *T. gondii* strains, this led to 398 miRNAs (RH), 515 miRNAs (PLK), 401 miRNAs (CTG); and for the developmental stages 448 miRNAs (bradyzoite), and 447 miRNAs (tachyzoite) remained. These numbers further decreased for comparison groups: 272 miRNAs (RH vs. PLK), 258 (RH vs. CTG), 289 (PLK vs. CTG), and 328 (tachyzoite vs. bradyzoite).

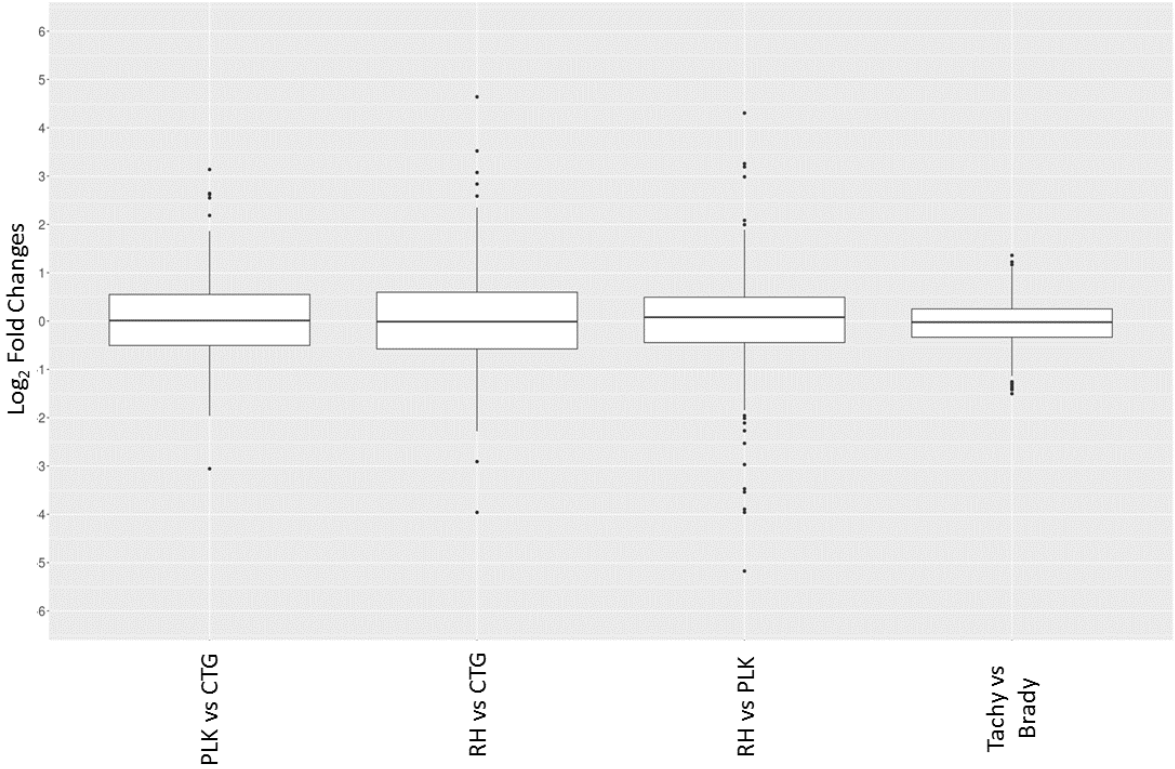

Figure 3.7. Distribution of differential miRNA expression (log_2 transformed) between strains and developmental stages.

For these miRNAs, Benjamini-Hochberg corrected t-test was applied and log_2 fold changes were calculated using R. The same threshold values as the differential gene expression analysis (p-value <0.05, l2fc>2 or l2fc<-2) were applied to differential miRNA

expression analysis. Only 2 miRNAs were found to be differentially expressed between PLK and CTG, another 2 miRNAs between RH and CTG and 5 in RH vs. PLK. No significantly differentially expressed miRNAs were found for tachyzoite vs. bradyzoite stages, thereby, confirming the findings for the above gene expression analysis. The distribution of $\log_2$ fold changes for miRNAs can be seen in Figure 3.7.

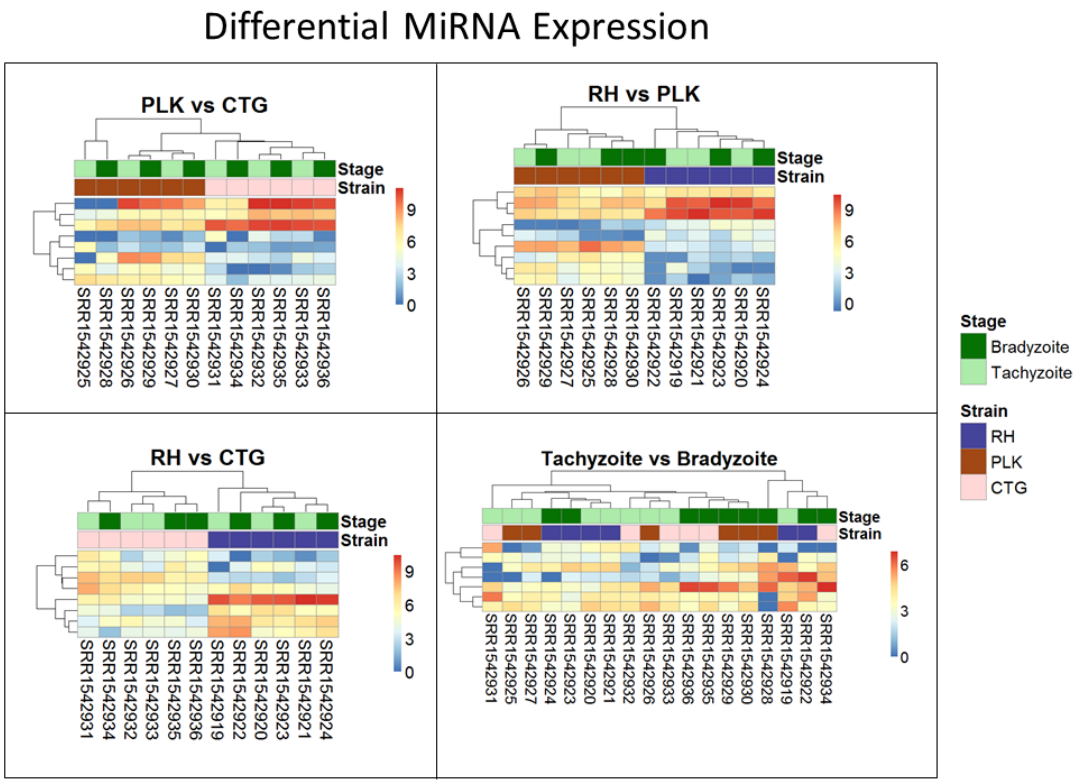

Figure 3.8. Heatmaps showing all differentially expressed microRNAs for pairs of strains (RH: blue, PLK: brown, CTG: pink) and developmental stages (bradyzoite: dark green, tachyzoite: mint). MicroRNA expression is hierarchically clustered based on the expression among replicates. Overexpression was analysed with pooled replicates, but for a better overview, all measurements are shown in columns including their hierarchical clustering. Actual values and miRNAs can be seen in Appendix A.2.

The $\log_2$ fold change distributions among strains and stages is quite similar for miRNAs (Figure 3.7) with the exception of differential expression for tachyzoites vs. bradyzoites which shows a very small inter quartile range (Figure 3.7).

Clustering of strains is dominating clustering as compared to developmental stage (Figure 3.8, Appendix A.2). Four of the miRNAs were annotated via similar sequences

in miRBase but unfortunately, they are either of plant origin or their targets are not annotated so that a cross annotation is not possible in this case. Many of the significantly differentially expressed miRNAs were detected in this study.

3.6. MicroRNA - mRNA Interactions

Since miRNA - mRNA interactions contain both expression analysis and network generation, they are explained in different subsections to be able to keep track of separate results.

3.6.1. Expression and Differential Expression

For miRNAs to be functionally active, they need to be co-expressed with their target mRNAs. It is, therefore, important to ensure that both miRNA and target are expressed in the same sample to be able to conclude anything about miRNA regulation. Above, pre-miRNAs and mature miRNAs were detected and their expression was confirmed. Gene expression was also established for the same samples. Therefore, it is possible to analyse miRNA and mRNA co-expression in this study. As a note, even though miRNA expression was explored in this study, generally miRNA expression analyses require specifically prepared libraries (Eminaga et al., 2013) (also see Table 1.2). However, the samples used in this study were prepared to detect mRNAs rather than miRNAs which led to low detection of miRNAs and almost no detection of their differential expression. To overcome this challenge, all miRNAs that do not originate from an annotated gene were discarded. For the remaining miRNAs (4,240) the expression of their source genes was used to represent their expression. Naturally, more reads will be mappable to mRNAs than much shorter miRNAs, which makes the approach chosen here more robust, as well. Thus, an interaction for this study is defined by a source gene and a target gene connected by a miRNA and co-expressed in the same sample.

Overall, 4,240 miRNAs and 8,920 (all annotated *T. gondii*) mRNAs were available for interaction analysis. If all interactions were possible, this would lead to approximately 40 million interactions. However according to targeting prediction that was done there were initially 161,970 interactions. Then these interactions were filtered by expres-

sions. If one component of the interaction (source gene, miRNA or target gene) was not expressed, then the interaction was filtered out. Out of a total of 161,970 interactions 65,602 were found to be co-expressed in this manner.

The ratio of target gene expression divided by source gene expression was used for the analysis of the differential expression of interactions. Interactions needed to exist in at least 70% of the samples in order to qualify for differential expression analysis.

Out of the total 65,602 interactions found, 63,120 of them were expressed in the samples of the RH strain. PLK had 63,778, CTG strain 62,494, bradyzoite stage 62,994, and tachyzoite stage had 62,369 interactions in their respective samples. As before, when considering differential expression of interactions, these numbers further trimmed down to 62,867 (RH vs PLK), 61,923 (RH vs. CTG), 62,441 (PLK vs. CTG), and 61,874 (tachyzoite vs. bradyzoite). T-test and $\log_2$ fold change calculations were performed for

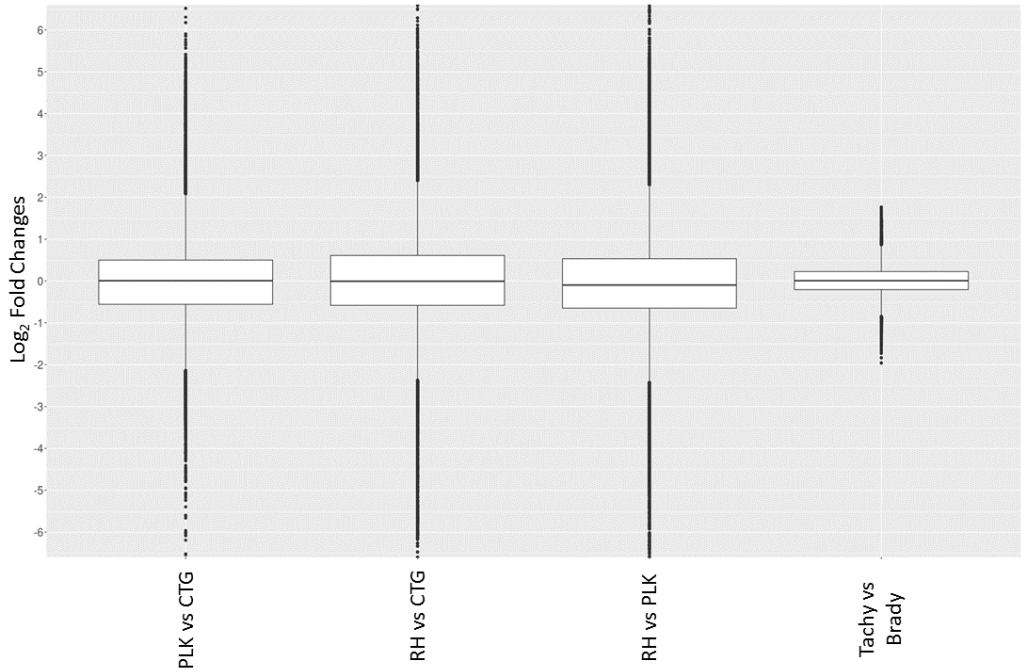

Figure 3.9. Distribution of differential expression of interactions (log2 transformed) between strains and developmental stages.

the remaining interactions. Thresholds were kept the same ($p < 0.05$, l2fc $<$-2 or l2fc $>$2) for the assessment of differential expression among strains and development stages. As can be seen in Figure 3.9, $\log_2$ fold changes between developmental stages did not vary significantly which supports the findings for differential expression of genes and miRNAs.

The distributions look similar to the distributions of differential gene expression (Figure 3.3) as can be expected since an interaction is defined by the ratio of the expression of a pair of genes. After significance filtering, 4,502 interactions were found to be differentially expressed in PLK vs. CTG, 5,488 (RH vs. CTG), 6,508 (RH vs. PLK), and none for bradyzoites vs. tachyzoites. Most of the top differentially expressed interactions are new

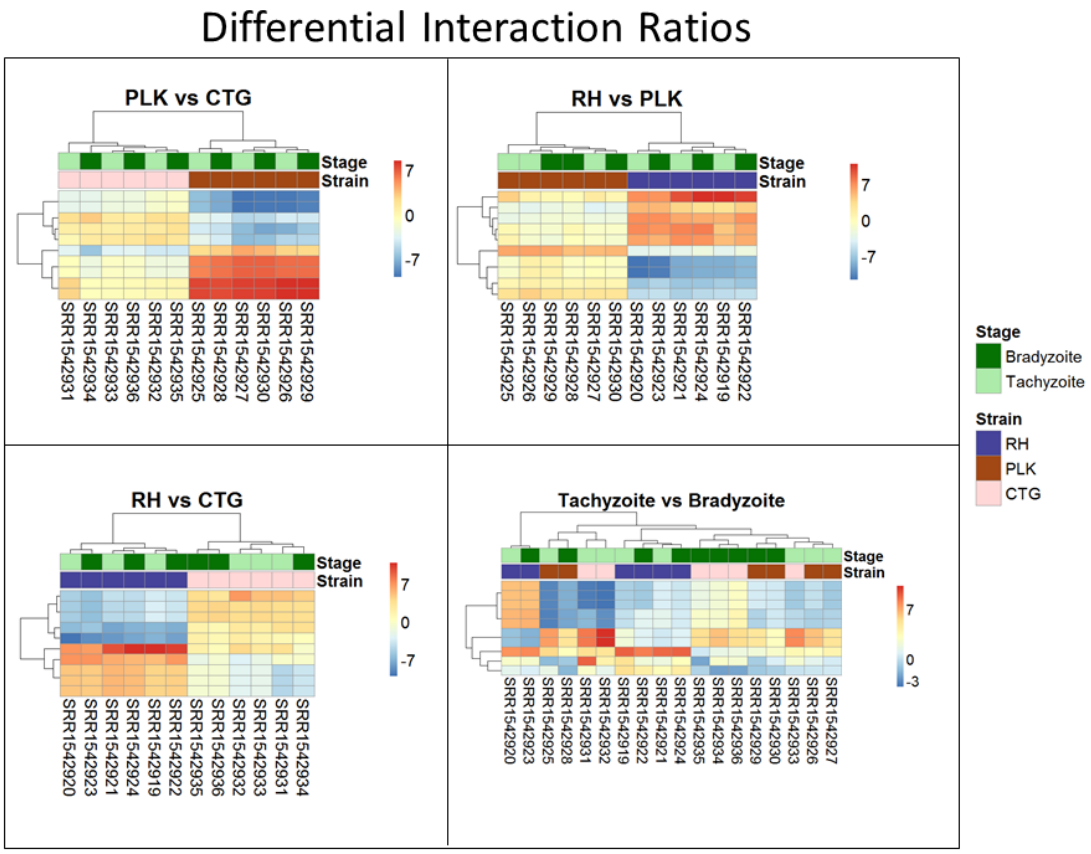

Figure 3.10. Heatmaps showing the 5 most over-expressed interactions per strain (RH: blue, PLK: brown, CTG: pink) and developmental stage (bradyzoite: dark green, tachyzoite: mint). Each pair presents 10 interactions of which 5 are over-expressed in one of the strains/stages and 5 in the other. Interactions are hierarchically clustered based on the expression among replicates. Over-expression was analysed with pooled replicates, but for a better overview, all measurements are shown in columns including their hierarchical clustering. Actual values and chosen interactions can be seen in Appendix A.3.

detections in this study. For the comparison between RH and CTG, however, more than half of the top differentially expressed interactions involved miRNAs similar to mouse

miRNAs. The comparison among strains leads to clustering of strains before stages (Figure 3.10, Appendix A.3). This is different for comparison between developmental stages where the clustering is not as expected which may be due to missing of actual significant differential expression.

It is interesting to note, that 10s of miRNAs, 100s of genes, and 1000s of interactions were significantly differentially expressed among strains but not between developmental stages. The interactions consist of co-expressed miRNAs and genes. From these results it can be deduced that few miRNAs and genes can lead to a wide variety of expressed (including differentially expressed) interactions which may have a large influence on the resulting phenotype.

3.6.2. Regulatory Network

An interaction network was formed using the 65,602 interactions that were expressed in total (Figure 3.11). This network contained a total of 5,126 nodes that consisted of expressed source and target genes. 173 of the nodes were only miRNA sources (3.37%) while 4297 were only targets (83.83%). 656 of the nodes (12.80%) were acting as both sources and targets for different interactions. 28 of the nodes (0.55%) were sources for miRNAs (43) targeting themselves i.e.: self-regulating.

While protein expression is not available for this study, taking into account highly connected regulatory components within the overall regulatory network may also add significance to the proposed interactions.

Therefore, the network was searched for cliques which are sub-graphs that are fully connected i.e.: each node is connected to all other nodes in the sub-graph via an edge (miRNA). Since an interaction consists of two nodes and an edge, such trivial cliques were ignored. Excluding cliques with less than 3 nodes, a total of 64,349 cliques were found in the interaction network. The largest cliques in the biological network consisted of 11 nodes, and there were only 2 such big cliques found. To assess whether the interaction network could be meaningful, nodes and edges were randomized using the NetworkRandomizer plug-in in Cytoscape to create 10 random networks. While the total number of cliques varied between random networks, none of them formed cliques with more than 4 nodes (Figure 3.12). In 10 random networks, 3 had the occurrence of one 4-node clique while in the interaction network, 13,134 4-node cliques were observed

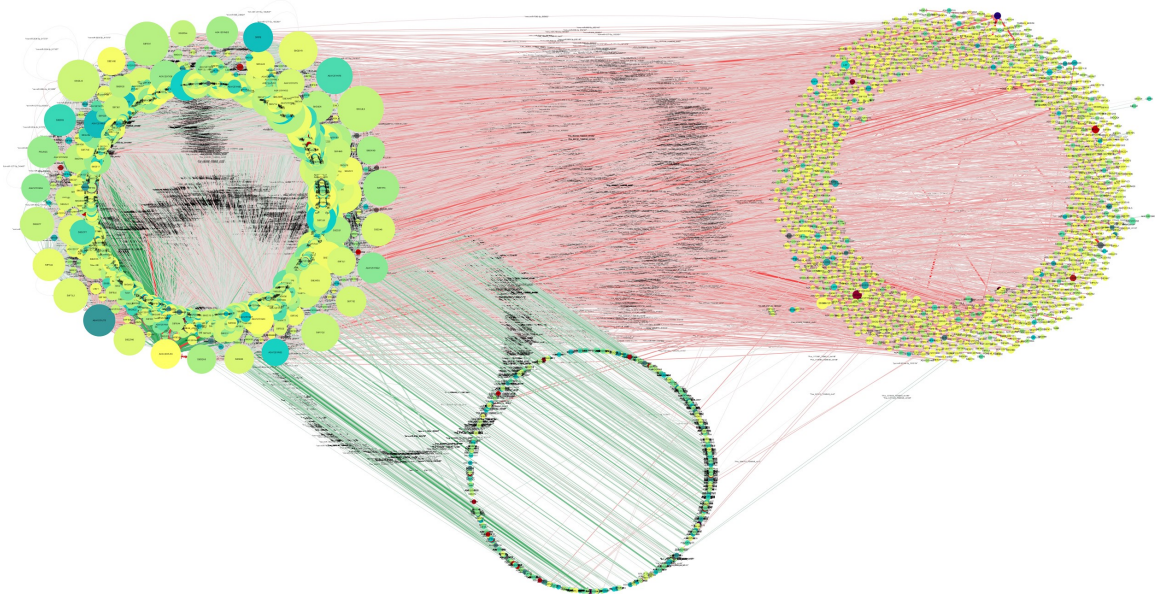

Figure 3.11. MicroRNA interaction network. Nodes represent genes; node size depends on the node degree and node color on the associated gene expression. Edges represent miRNA driven regulation between the genes connected by the edge (source gene – miRNA —– target gene). Edge color and width represent the normalized targeting ratios.

(Figure 3.12). Therefore, the cliques in biological network is significantly different and unique. These cliques may be biologically meaningful, as the sizes and amounts were created from miRNA targeting prediction.

The largest two cliques in the interaction network consisted of 11 nodes with 119 and 120 edges, respectively. Such strong coupling and the large amount of cliques suggest biological meaning for these regulatory components. The most targeted node in the interaction network is S8GNL0 with 209 incoming interactions. The most outgoing interactions are from the gene S8F823 amounting to 1,312 targets (524 distinct ones) from 4 distinct miRNAs (mmu-miR-466g_277687_1: 193 interactions and tch-miR-1277-5p_275830_1, tch-miR-1277-5p_275830_2, and tch-miR-1277-5p_275830_3: 373 interactions each). Considering both targeting and being targeted at the same time, the most interactive gene is S8F823 with 12 miRNAs targeting it to its additional 1,312 targets (with additional multiple mature miRNAs targeting the same gene), totaling 1,324 interactions.

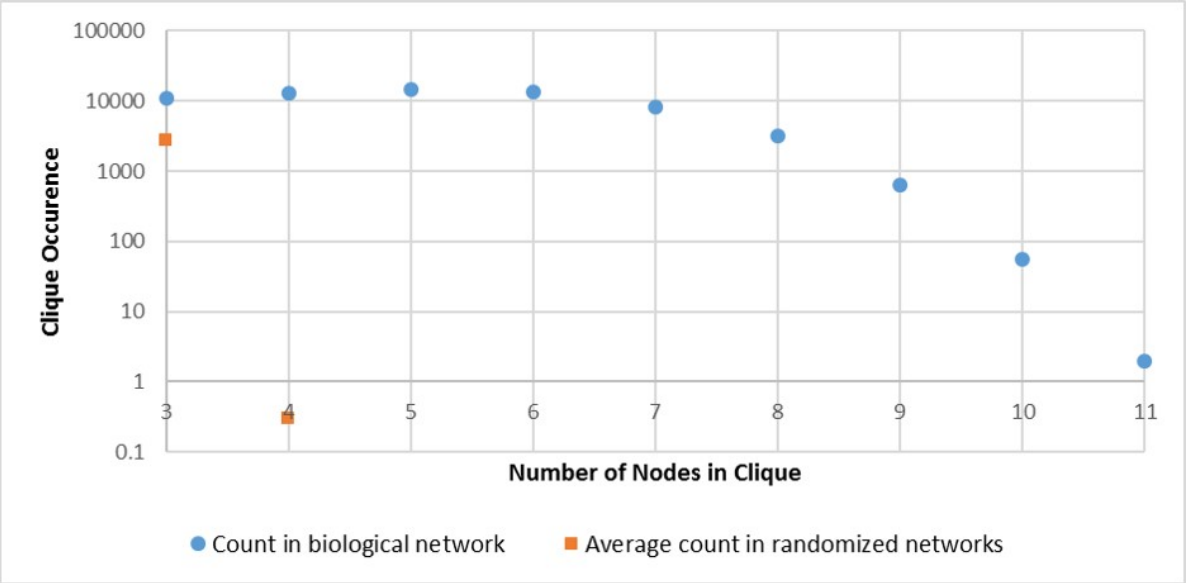

Figure 3.12. Clique Occurrence. In the biological network created in this study, there are multiple cliques containing more than 4 nodes. It was seen that the highest number of cliques were formed with 5 nodes (14621). In created 10 random networks, average amount of cliques was only high for 3 node ones. 4 node cliques were seen once in 3 different random network only.

Most of the cliques found in the interactions network were of size 5 and, therefore, such a clique was chosen to be presented (Figure 3.13).

The selected clique contains one gene (S8F559) which displays self-regulation via two miRNAs and is, thereby, also the most targeted gene in this clique with seven incoming miRNA edges (Figure 3.13). A clique with five nodes needs a minimum of 10 edges to be fully connected. The selected clique contains 19 edges with 7 of them being redundant for the clique criterion and two being self-regulatory edges which are also not considered for clique detection. Some of the miRNAs (6) in this clique are similar to miRNAs from tch (tree shrew) which is a squirrel like animal and thus a host for *T. gondii*. Other cliques (not shown) are enriched in miRNAs from, for example, human or rat.

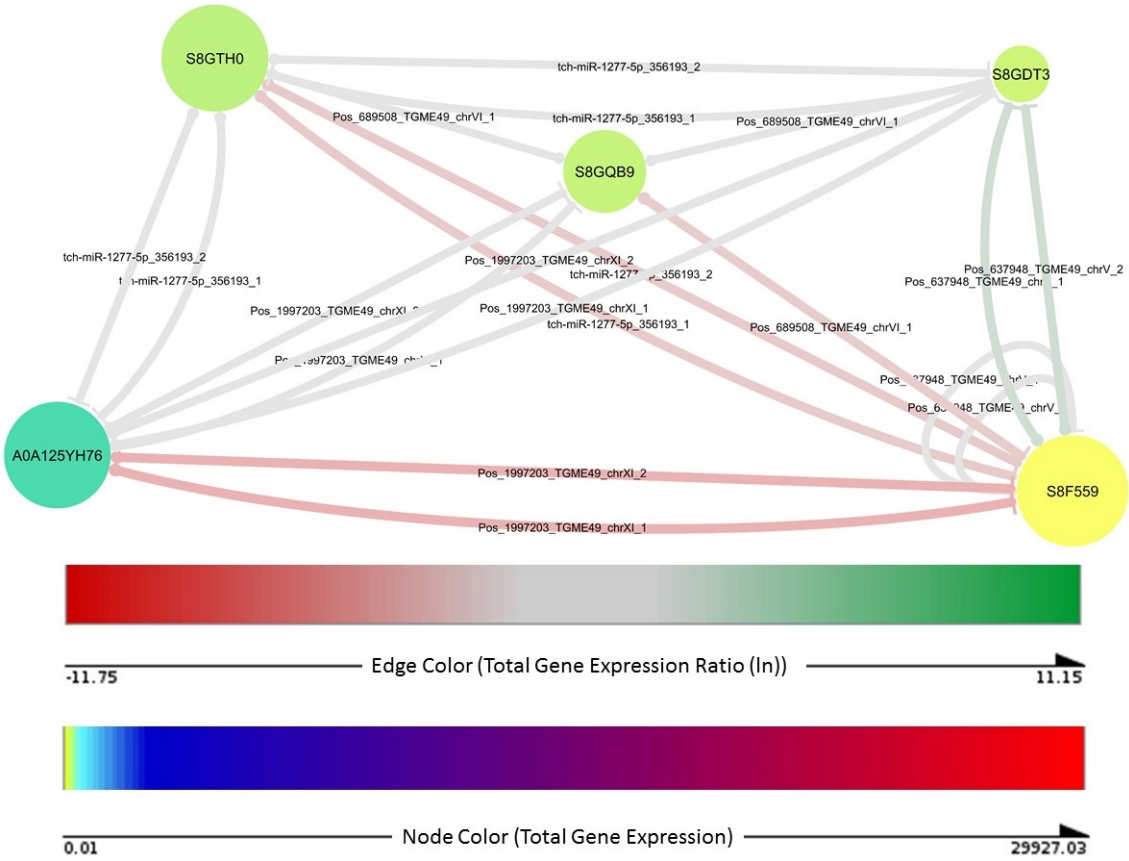

Figure 3.13. Representative clique from the interaction network. Nodes (5) represent genes. Edges (19) represent mature miRNAs which are part of their source gene. Colour of the nodes shows total normalized gene expression. Colour of the edges shows gene expression ratio, which is defined as the natural logarithm of the target gene expression divided by the source gene expression. Thickness of edges emphasizes extreme gene expression ratio. Source gene of the miRNA is indicated by a circle at the edge whereas the targeted side is modelled as a T-shaped tip.

CHAPTER 4

CONCLUSION

Little is known about the miRNA-based regulation in *T. gondii*. Therefore, pre-miRNAs, their associated mature miRNAs, and the mRNA targets were predicted from the genome. A publicly available RNA-seq dataset investigating three *T. gondii* strains (RH, PLK, and CTG) and two developmental stages (tachyzoite and bradyzoite) was used to analyse expression of the detected miRNAs and their targets. In an attempt to add further confidence, miRNAs and their targets were analysed together as interactions. 65,602 expressed interactions were found between the 4,240 miRNAs and 8,920 annotated mRNAs. Currently, 339 miRNAs have been described for *T. gondii* of which a number (47 out of 339) of them was disputed previously (Saçar Demirci et al., 2016). Here, 4,240 miRNAs (containing 305 of the 339 known ones) and their targets co-expressed in the same sample are presented.

With the available data, it was possible to create a miRNA driven gene regulation network. Metabolic pathways of *T. gondii* are not well defined and most of the proteins synthesized by *T. gondii* remain hypothetical. With further studies, the interactions presented in this study can be integrated into proven pathways of *T. gondii* for better understanding of the regulation happening in this intra-cellular parasite. Furthermore, interactions found in this study may lead to interesting drug targets. Combining the knowledge presented in this study, these drug targets can be chosen so that they affect any strains or developmental stages that were explored.

REFERENCES

Abe, M. and N. M. Bonini (2013, jan). MicroRNAs and neurodegeneration: role and impact. *Trends in Cell Biology 23*(1), 30–36.

Adiconis, X., D. Borges-Rivera, R. Satija, D. S. DeLuca, M. A. Busby, A. M. Berlin, A. Sivachenko, D. A. Thompson, A. Wysoker, T. Fennell, A. Gnirke, N. Pochet, A. Regev, and J. Z. Levin (2013, may). Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature Methods 10*(7), 623–629.

Allmer, J. and M. D. Saçar Demirci (2016, jul). izMiR: computational ab initio microRNA detection. *Protocol Exchange*.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Axtell, M. J., J. O. Westholm, and E. C. Lai (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology 12*(4), 221.

Bağcı, C. and J. Allmer (2016). JLab Hairpin Feature Calculator. http://jlab.iyte.edu.tr/software/mirna. Accessed on: 2017-05-02.

Bartel, D. P. (2009). MicroRNA Target Recognition and Regulatory Functions. *Cell 136*(2), 215–233.

Baştanlar, Y. and M. Özuysal (2014). Introduction to Machine Learning. 10.1007/978-1-62703-748-8_7.

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

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel (2009, nov). KNIME - the Konstanz information miner. *ACM SIGKDD Explorations Newsletter 11*(1), 26.

Bollman, K. M., M. J. Aukerman, M.-Y. Park, C. Hunter, T. Z. Berardini, and R. S. Poethig (2003, apr). HASTY, the Arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development (Cambridge, England) 130*(8), 1493–504.

Borchert, G. M., W. Lanier, and B. L. Davidson (2006, dec). RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology 13*(12), 1097–1101.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). BLAST plus: architecture and applications. *BMC Bioinformatics 10*(421), 1.

Carmell, M. A., Z. Xuan, M. Q. Zhang, and G. J. Hannon (2002, nov). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes & development 16*(21), 2733–42.

CDC (2017). Neglected Parasitic Infections (NPIs). https://www.cdc.gov/parasites/npi/. Accessed on: 2017-05-14.

Chou, C.-H., N.-W. Chang, S. Shrestha, S.-D. Hsu, Y.-L. Lin, W.-H. Lee, C.-D. Yang, H.-C. Hong, T.-Y. Wei, S.-J. Tu, T.-R. Tsai, S.-Y. Ho, T.-Y. Jian, H.-Y. Wu, P.-R. Chen, N.-C. Lin, H.-T. Huang, T.-L. Yang, C.-Y. Pai, C.-S. Tai, W.-L. Chen, C.-Y. Huang, C.-C. Liu, S.-L. Weng, K.-W. Liao, W.-L. Hsu, and H.-D. Huang (2016, jan). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research 44*(D1), D239–D247.

Croken, M. M., Y. Ma, L. M. Markillie, R. C. Taylor, G. Orr, L. M. Weiss, and K. Kim (2014, nov). Distinct Strains of Toxoplasma gondii Feature Divergent Transcriptomes Regardless of Developmental Stage. *PLoS ONE 9*(11), e111297.

Dai, X. and P. X. Zhao (2011, jul). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research 39*(suppl), W155–W159.

D'haeseleer, P. (2005, dec). How does gene expression clustering work? *Nature Biotechnology 23*(12), 1499–1501.

Du, T. (2005, sep). microPrimer: the biogenesis and function of microRNA. *Development 132*(21), 4645–4652.

Dubey, J. (2004, dec). Toxoplasmosis - a waterborne zoonosis. *Veterinary Parasitology 126*(1-2), 57–72.

Dubey, J. P. (1996). Toxoplasma Gondii. http://www.ncbi.nlm.nih.gov/pubmed/21413265.

Dubey, J. P. (2008, nov). The History of Toxoplasma gondii - The First 100 Years. *Journal of Eukaryotic Microbiology 55*(6), 467–475.

Dubey, J. P., D. S. Lindsay, and C. A. Speer (1998, apr). Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts. *Clinical microbiology reviews 11*(2), 267–99.

Eminaga, S., D. C. Christodoulou, F. Vigneault, G. M. Church, and J. Seidman (2013, jul). Quantification of microRNA Expression with Next-Generation Sequencing. In *Current Protocols in Molecular Biology*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Farazi, T. A., J. I. Hoell, P. Morozov, and T. Tuschl (2013). MicroRNAs in Human Cancer. pp. 1–20.

Flegr, J., J. Prandota, M. Sovičková, and Z. H. Israili (2014, mar). Toxoplasmosis - A Global Threat. Correlation of Latent Toxoplasmosis with Specific Disease Burden in a Set of 88 Countries. *PLoS ONE 9*(3), e90203.

Friedman, R. C., K. K.-H. Farh, C. B. Burge, and D. P. Bartel (2009, jan). Most mammalian mRNAs are conserved targets of microRNAs. *Genome research 19*(1), 92–105.

Gajria, B., A. Bahl, J. Brestelli, J. Dommer, S. Fischer, X. Gao, M. Heiges, J. Iodice, J. C. Kissinger, A. J. Mackey, D. F. Pinney, D. S. Roos, C. J. Stoeckert, H. Wang, and B. P. Brunk (2007, dec). ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic Acids Research 36*(Database), D553–D556.

Gkirtzou, K., I. Tsamardinos, P. Tsakalides, and P. Poirazi (2010, aug). MatureBayes: A Probabilistic Algorithm for Identifying the Mature miRNA within Novel Precursors. *PLoS ONE 5*(8), e11843.

Glazov, E. A., P. A. Cottee, W. C. Barris, R. J. Moore, B. P. Dalrymple, and M. L. Tizard (2008, may). A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome Research 18*(6), 957–964.

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016, may). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 17*(6), 333–351.

Grada, A. and K. Weinbrecht (2013, aug). Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology 133*(8), 1–4.

Ha, M. and V. N. Kim (2014). Regulation of microRNA biogenesis. *Nature reviews. Molecular cell biology 15*(8), 509–524.

Herrero, J., M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. J. Searle, R. Amode, S. Brent, W. Spooner, E. Kulesha, A. Yates, and P. Flicek (2016). Ensembl comparative genomics resources. *Database : the journal of biological databases and curation 2016*.

Jongert, E., C. W. Roberts, N. Gargano, E. Förster-Waldl, E. Förster-Wald, and E. Petersen (2009, mar). Vaccines against Toxoplasma gondii: challenges and opportunities. *Memorias do Instituto Oswaldo Cruz 104*(2), 252–66.

Joshi, N. and J. Fass (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle.

Ketting, R. F., S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk

(2001, oct). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & development 15*(20), 2654–9.

Kozomara, A. and S. Griffiths-Jones (2014, jan). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research 42*(D1), D68–D73.

Kukurba, K. R. and S. B. Montgomery (2015, nov). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols 2015*(11), pdb.top084970.

Lau, N. C., L. P. Lim, E. G. Weinstein, and D. P. Bartel (2001, oct). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.) 294*(5543), 858–62.

Lee, R. C., R. L. Feinbaum, and V. Ambros (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*(5), 843–854.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim (2003, sep). The nuclear RNase III Drosha initiates microRNA processing. *Nature 425*(6956), 415–9.

Leinonen, R., H. Sugawara, and M. Shumway (2011, jan). The Sequence Read Archive. *Nucleic Acids Research 39*(Database), D19–D21.

Libbrecht, M. W. and W. S. Noble (2015, may). Machine learning applications in genetics and genomics. *Nature Reviews Genetics 16*(6), 321–332.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology 2012*, 1–11.

Liu, Q., Z.-D. Wang, S.-Y. Huang, and X.-Q. Zhu (2015, dec). Diagnosis of toxoplasmosis and typing of Toxoplasma gondii. *Parasites & Vectors 8*(1), 292.

Mardis, E. R. (2008, mar). The impact of next-generation sequencing technology on genetics. *Trends in Genetics 24*(3), 133–141.

Martin, M. (2011, may). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal 17*(1), 10.

McCabe, R. and J. S. Remington (1988, feb). Toxoplasmosis: the time has come. *The New England journal of medicine 318*(5), 313–5.

Metzker, M. L. (2010, jan). Sequencing technologies - the next generation. *Nature Reviews Genetics 11*(1), 31–46.

Millar, A. A. and P. M. Waterhouse (2005, jul). Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics 5*(3), 129–135.

Ng, K. L. S. and S. K. Mishra (2007, jun). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics 23*(11), 1321–1330.

Olena, A. F. and J. G. Patton (2009). Genomic organization of microRNAs. *Journal of Cellular Physiology*.

Ondov, B. D., A. Varadarajan, K. D. Passalacqua, and N. H. Bergman (2008, dec). Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics 24*(23), 2776–2777.

Papp, I., M. F. Mette, W. Aufsatz, L. Daxinger, S. E. Schauer, A. Ray, J. van der Winden, M. Matzke, and A. J. M. Matzke (2003, jul). Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant physiology 132*(3), 1382–90.

Reinhart, B. J., E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel (2002, jul). MicroRNAs in plants. *Genes & development 16*(13), 1616–26.

Romaine, S. P. R., M. Tomaszewski, G. Condorelli, and N. J. Samani (2015, jun). MicroRNAs in cardiovascular disease: an introduction for clinicians. *Heart 101*(12), 921–928.

Saçar Demirci, M. D., C. Bağcı, and J. Allmer (2016). Differential Expression of Toxoplasma gondii MicroRNAs in Murine and Human Hosts. In *Non-coding RNAs and Inter-kingdom Communication*, Chapter Non-coding, pp. 143–159. Cham: Springer International Publishing.

Sætrom, P. and O. Snøve (2007). Robust Machine Learning Algorithms Predict MicroRNA Genes and Targets. pp. 25–49.

Saito, Y., H. Saito, G. Liang, and J. M. Friedman (2014, oct). Epigenetic Alterations and MicroRNA Misexpression in Cancer and Autoimmune Diseases: a Critical Review. *Clinical Reviews in Allergy & Immunology 47*(2), 128–135.

Sanger, F., S. Nicklen, and A. R. Coulson (1977, dec). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America 74*(12), 5463–7.

Schloss, J. A. (2008, oct). How to get genomes at one ten-thousandth the cost. *Nature Biotechnology 26*(10), 1113–1115.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (2003, nov). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research 13*(11),

2498–504.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson (2015, jul). Big Data: Astronomical or Genomical? *PLOS Biology 13*(7), e1002195.

Team, R. C. (2016). R: A Language and Environment for Statistical Computing. https://www.r-project.org/.

Tenter, A. M., A. R. Heckeroth, and L. M. Weiss (2000, nov). Toxoplasma gondii: from animals to humans. *International journal for parasitology 30*(12-13), 1217–58.

Thounaojam, M. C., K. Kundu, D. K. Kaushik, S. Swaroop, A. Mahadevan, S. K. Shankar, and A. Basu (2014, may). MicroRNA 155 Regulates Japanese Encephalitis Virus-Induced Inflammatory Response by Targeting Src Homology 2-Containing Inositol Phosphatase 1. *Journal of Virology 88*(9), 4798–4810.

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

Tüfekci, K. U., R. L. J. Meuwissen, and e. Genç (2014). The Role of MicroRNAs in Biological Processes. pp. 15–31.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter,

S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu (2001, feb). The sequence of the human genome. *Science (New York, N.Y.) 291*(5507), 1304–51.

Wahid, F., A. Shehzad, T. Khan, and Y. Y. Kim (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochimica et Biophysica Acta - Molecular Cell Research 1803*(11), 1231–1243.

Wang, J., X. Liu, B. Jia, H. Lu, S. Peng, X. Piao, N. Hou, P. Cai, J. Yin, N. Jiang, and Q. Chen (2012). A comparative study of small RNAs in Toxoplasma gondii of distinct genotypes. *Parasites & Vectors 5*(1), 186.

Wetterstrand, K. A. (2016). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Winter, J., S. Jung, S. Keller, R. I. Gregory, and S. Diederichs (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature cell biology 11*(3), 228–234.

Winter, J., S. Link, D. Witzigmann, C. Hildenbrand, C. Previti, and S. Diederichs (2013, may). Loop-miRs: active microRNAs generated from single-stranded loop regions. *Nucleic Acids Research 41*(10), 5503–5512.

Xie, M. and J. A. Steitz (2014). Versatile microRNA biogenesis in animals and their viruses. *RNA biology 11*(6), 673–81.

Xu, Q.-S. and Y.-Z. Liang (2001, apr). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems 56*(1), 1–11.

Yi, R., Y. Qin, I. G. Macara, and B. R. Cullen (2003, dec). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development 17*(24), 3011–6.

Yones, C. A., G. Stegmayer, L. Kamenetzky, and D. H. Milone (2015, dec). miRNAfe: A comprehensive tool for feature extraction in microRNA prediction. *Biosystems 138*, 1–5.

APPENDIX A

PLOTTED DIFFERENTIAL EXPRESSION VALUES

Table A.1. Top 5 Differentially Expressed Genes

Differential Expression - Genes					
PLK vs CTG			RH vs CTG		
Gene	**l2fc**	**padj**	**Gene**	**l2fc**	**padj**
S8EMV8	8.63	3.29E-03	S8EMV8	9.91	1.84E-03
S8F7Q7	7.19	1.72E-03	ROP18	6.76	4.87E-04
ROP18	6.86	3.47E-04	A0A125YSP7	5.66	6.01E-06
S8G544	5.80	2.25E-02	ABCG84	5.55	3.92E-04
S8GA24	5.55	2.19E-04	S8F776	5.49	4.94E-04
S8GIV7	-5.88	1.53E-04	A0A125YGM9	-6.81	1.09E-02
S8GJX7	-6.24	2.87E-05	S8GBJ7	-7.15	1.44E-04
S8F4M4	-6.30	1.01E-03	S8F7B9	-7.31	3.73E-04
S8GIE6	-6.42	2.29E-03	S8EZ59	-7.43	9.50E-04
S8F425	-6.50	9.60E-04	S8EX82	-8.72	1.46E-03
RH vs PLK			**Tachyzoite vs Bradyzoite**		
S8GIE6	7.56	1.02E-03	A0A125YVF6	1.65	9.26E-01
S8F425	7.03	5.81E-04	S8F918	1.50	9.18E-01
S8GH76	5.71	7.61E-05	TGME49_355050	1.34	9.42E-01
S8G1H2	5.26	9.70E-04	A0A125YG81	1.33	9.85E-01
A0A125YSP7	5.10	1.85E-06	TGME49_355210	1.17	9.18E-01
S8GDF0	-6.34	6.57E-04	S8EPZ6	-1.60	9.59E-01
S8F0T6	-6.55	8.10E-04	S8F8L7	-1.62	9.62E-01
S8EX82	-6.55	1.33E-03	S8GMZ2	-1.65	9.83E-01
S8F5N9	-6.62	2.48E-03	S8ETH6	-1.75	9.49E-01
A0A125YGM9	-6.66	8.89E-03	AP2XII2	-1.94	9.26E-01

Table A.2. Differentially Expressed miRNAs

Differential Expression - miRNAs		
PLK vs CTG		
miRNA	**l2fc**	**padj**
Pos_1017361_TGME49_chrVIIb	3.23	5.28E-02
Neg_2400181_TGME49_chrXII	2.58	9.79E-01
Neg_1863667_TGME49_chrX	2.46	4.82E-02
Pos_1740891_TGME49_chrX	2.26	8.04E-01
Pos_541500_TGME49_chrV	-1.80	9.79E-01
Pos_277477_TGME49_chrII	-1.84	3.30E-01
Neg_1446843_TGME49_chrIX	-2.02	9.79E-01
Pos_1517039_TGME49_chrIX	-3.74	1.58E-02
RH vs PLK		
Neg_2337788_TGME49_chrXII	3.66	3.62E-05
Pos_1632774_TGME49_chrX	3.13	1.08E-02
Pos_1517039_TGME49_chrIX	2.56	3.01E-01
Neg_460000_TGME49_chrIV	1.93	7.48E-01
mmu-miR-6414_2286979	1.83	5.37E-01
Neg_557604_TGME49_chrV	-3.65	6.65E-02
tae-miR1128_633204	-4.18	7.47E-02
Pos_844987_TGME49_chrVIIa	-5.00	9.63E-05
RH vs CTG		
Neg_2337788_TGME49_chrXII	4.87	3.36E-03
gra-miR7484c_2030510	2.94	2.39E-01
Neg_1863667_TGME49_chrX	2.79	2.39E-01
tch-miR-1277-5p_564989	2.54	1.14E-01
Pos_2184537_TGME49_chrXII	-2.32	6.37E-01
Neg_1765335_TGME49_chrX	-2.45	9.93E-01
Neg_1417681_TGME49_chrIX	-2.65	2.39E-01
Pos_844987_TGME49_chrVIIa	-3.85	4.82E-03
Tachyzoite vs Bradyzoite		
Pos_2115001_TGME49_chrXI	1.51	9.90E-01
Pos_541500_TGME49_chrV	1.47	9.90E-01
Pos_2164687_TGME49_chrXII	1.45	9.90E-01
Pos_1978651_TGME49_chrXI	-1.39	9.90E-01
hsa-miR-4524b-3p_1680787	-1.68	9.90E-01
Neg_1739554_TGME49_chrX	-1.73	9.90E-01
Pos_1842340_TGME49_chrX	-1.74	9.90E-01

Table A.3. Top 10 Differentially Expressed Interactions

Differential Expression - Interactions		
PLK vs CTG		
Interaction	**l2fc**	**padj**
S8GRN8—Pos_570377_TGME49_chrV_1—S8G1H7	7.96	1.69E-05
S8GRN8—Pos_570377_TGME49_chrV_2—S8G1H7	7.96	1.69E-05
S8F6M7—ptr-miR-3149_468546_1—A0A125YQV8	7.73	3.21E-04
S8F6M7—ptr-miR-3149_468546_2—A0A125YQV8	7.73	3.21E-04
S8GRN8—Pos_570377_TGME49_chrV_1—S8F675	7.29	7.18E-05
S8GB10—tch-miR-1277-5p_2146654_1—S8GRN8	-6.20	5.93E-05
LAP—Neg_1539412_TGME49_chrIX_1—S8GIV7	-6.52	3.50E-04
LAP—Neg_1539412_TGME49_chrIX_2—S8GIV7	-6.52	3.50E-04
S8GG66—gga-miR-3523_1617267_1—S8GIV7	-6.55	7.67E-04
S8F6X6—Pos_1102181_TGME49_chrVIIb_1—S8F4M4	-7.38	1.01E-03
RH vs PLK		
S8FCX5—Neg_452509_TGME49_chrIV_1—A0A125YSP7	8.60	1.68E-06
S8FC20—hsa-miR-6728-5p_210073_1—S8EQB0	7.81	3.22E-04
S8F0T6—Neg_1386367_TGME49_chrVIII_1—A0A125YLZ6	7.11	4.23E-05
S8FCX5—Neg_452509_TGME49_chrIV_1—S8F1G6	6.92	1.38E-05
S8F0T6—Neg_1386367_TGME49_chrVIII_1—S8G8G7	6.85	2.23E-05
ABCG84—Neg_1516947_TGME49_chrIX_1—S8EZM6	-7.81	1.83E-07
ABCG84—Neg_1516947_TGME49_chrIX_1—A0A125YP51	-8.49	3.39E-07
ABCG84—Neg_1516947_TGME49_chrIX_1—S8F591	-8.79	1.84E-07
A0A125YI56—gra-miR7484c_2030510_1—A0A125YNV2	-9.47	3.35E-05
A0A125YI56—gra-miR7484c_2030510_2—A0A125YNV2	-9.47	3.35E-05
RH vs CTG		
S8FCX5—Neg_452509_TGME49_chrIV_1—A0A125YSP7	7.69	2.13E-05
S8F591—mmu-miR-7063-3p_585890_1—S8GML9	7.44	8.65E-04
S8F591—mmu-miR-7063-3p_585890_2—S8GML9	7.44	8.65E-04
S8F591—mmu-miR-7063-3p_585890_3—S8GML9	7.44	8.65E-04
S8FC20—hsa-miR-6728-5p_210073_1—S8EQB0	7.02	9.07E-04
S8F7Q2—mmu-miR-466m-3p_679113_1—S8EZ59	-8.78	1.53E-04
S8F7Q2—mmu-miR-466m-3p_679113_2—S8EZ59	-8.78	1.53E-04
S8FE58—Neg_284035_TGME49_chrII_1—S8EZ59	-9.36	1.75E-05
ABCG84—Neg_1516947_TGME49_chrIX_1—S8F081	-9.46	1.06E-05
S8GL47—mmu-miR-467g_1104875_1—S8F7B9	-9.77	6.79E-06
Tachyzoite vs Bradyzoite		
S8F559—Pos_637948_TGME49_chrV_1—S8EZ59	1.77	9.60E-01
S8F559—Pos_637948_TGME49_chrV_2—S8EZ59	1.77	9.60E-01
A0A125YKL6—gga-miR-1627-3p_500385_1—S8F5J3	1.75	8.17E-01
S8FC20—hsa-miR-6728-5p_210073_1—S8EWB5	1.74	9.78E-01
S8F559—Pos_637948_TGME49_chrV_1—S8G387	1.72	9.88E-01
S8F056—Neg_1782640_TGME49_chrX_2—S8EY17	-1.84	9.98E-01
S8F056—Neg_1782640_TGME49_chrX_3—S8EY17	-1.84	9.98E-01
S8F056—Neg_1782640_TGME49_chrX_1—AP2IX5	-1.96	9.61E-01
S8F056—Neg_1782640_TGME49_chrX_2—AP2IX5	-1.96	9.61E-01
S8F056—Neg_1782640_TGME49_chrX_3—AP2IX5	-1.96	9.61E-01