Dual RNA-Seq analysis of *Mus musculus* and *Leishmania donovani* transcriptomes

ALEXANDRA HART

MASTER OF SCIENCE BY RESEARCH

UNIVERSITY OF YORK

BIOLOGY

DECEMBER 2017

ABSTRACT

Parasitic protozoa of the genus *Leishmania* cause a spectrum of disease, affecting 12 million people worldwide. This project aimed to investigate the effect of *Leishmania donovani* infection on the gene expression of healthy/WT (Black 6) and immunocompromised (RAG2KO) mice. Differences in the gene expression of parasites in inoculum and tissue were also elucidated.

WT and RAG mice were infected using an *L. donovani* inoculum, and were euthanised after 28 days. Harvested spleens (and the inoculum) were used to generate RNA samples, from which mRNA was isolated and purified. Transcriptome data was generated using dual RNA-Seq approaches from the mRNA samples. After appropriate pre-processing, data underwent a number of bioinformatic analyses to explore differential gene expression, such as Gene Ontology, Gene Set Enrichment, and KEGG Pathway analysis.

Comparison of different mouse spleen transcriptomes revealed that even in uninfected mice, WT mice more highly express genes related to immunoglobulins when compared with their immunocompromised counterparts. Healthy mice were found to react to infection through the induction of inflammatory response, and the production of NOX generating species. RAG mice still upregulated immunoglobulin-related genes in response to infection, despite their inability to generate antibodies, T-cells, and B-cells. However, RAG modulation of haeme and iron metabolism may contribute to defence against the parasites despite a lack of acquired immunity.

Differences in the amastin, the key glycoprotein on the surface on intracellular-stage parasites, are apparent between the inoculum and tissue parasites, which may reflect microenvironment adaptation. Additionally, tissue-derived parasites showed significant upregulation of genes related to gene expression control, such as histones and DNA-packing.

These experiments are among the first attempts to in vivo transcriptome sequence mice and *Leishmania* simultaneously, a powerful approach giving insight to action and reaction. However, these techniques are not without challenge, such as low parasite read counts.

LIST OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

CHAPTER 2: METHODS

CHAPTER 3: RESULTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

CHAPTER 2: METHODS

CHAPTER 3: RESULTS

LIST OF ACCOMPANYING MATERIAL

CHAPTER 3: RESULTS

A1: Lists of differentially expressed genes

A2: Lists of gene set enrichment analysis results

A3: Lists of gene ontology results

ACKNOWLEDGEMENTS

Many thanks are necessary for all of the help and guidance I received during this project. First and foremost, I would like to thank Professor Jeremy Mottram, for his endless hard work patching up the holes in my experimental design, lab work, thesis, and well – everything!

Thank you to the Genomics and Bioinformatics Technology Facility, and the York Bioinformatics club, especially Dr Sally James and Dr John Davey, for answering all of my silly questions and for their support and encouragement.

Thank you to the Mottram lab group, for their fantastic advice, dazzling array of research, and constant supply of cake.

Thank you to my friends and family, who have put up with my complaining for the entire duration of this project.

Lastly, I would like to extend a special thanks to Dr Sarah Forrester, without whom I would never have completed this thesis.

DECLARATION

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

CHAPTER 1: INTRODUCTION

1.1 LEISHMANIASIS

1.1.1 OVERVIEW

The leishmaniases are a broad spectrum of diseases caused by protozoan parasites of the genus *Leishmania*, affecting humans and animals across a distribution covering five continents, mainly in the Middle East and South America (Pace 2014). Symptoms range from mild, self-curing cutaneous lesions to potentially fatal visceral organ damage (Reithinger et al. 2007; Kaye and Scott 2011). Borne by sandfly vectors, and with latent populations in rodent and dog reservoirs, over 350 million people are at risk, with 12 million affected globally (Pace 2014; Okwor and Uzonna 2016; WHO 2016). Leishmaniasis contributes a significant burden on global health; among parasitic diseases, leishmaniasis is the second biggest cause of mortality and the third highest cause of morbidity (in terms of DALYs) (Reithinger et al. 2007; Pace 2014).

1.1.2 CLINICAL DISEASE AND SUBTYPES

Leishmaniasis manifests in one of three disease tropisms, causing cutaneous, visceral, or mucocutaneous disease, depending on the infecting *Leishmania* species. Cutaneous leishmaniasis (CL), predominantly caused by *L. major*, is the most commonly occurring form of the disease, causing approximately 1.2 million new cases each year (Pace 2014). Though rarely fatal, CL can cause severe scarring and disfigurement of areas exposed to the sandfly vector, with psychological and social implications (Reithinger et al. 2005; Kedzierski et al. 2006; Okwor and Uzonna 2016). Mucocutaneous leishmaniasis (MCL) is relatively rare, with only 25,000 new cases each year, and occurs when the host immune response allows parasites to escape cutaneous lesions, months or years after infection (Pace 2014). MCL is most often caused by *L. braziliensis*, though can be caused by other *Leishmania* species (Kedzierski et al. 2006; Reithinger et al. 2007; Maretti-Mira et al. 2012). Visceral leishmaniasis (VL or kala-azar) is the most severe form of the disease, and is responsible for 0.4 million new infections every year. VL occurs in both the New World, where the causative agent is *L. infantum* (also known as *L. chagasi*), and the Old World, where *L. donovani* is responsible (WHO 2016). Untreated VL is often fatal, and patients develop severe disease with symptoms such as splenic and hepatic organomegaly (WHO 2013; Pace 2014). Drug treatment for VL is often lengthy and toxic, and severe post-treatment complications can occur, such as post kala-azar dermal leishmaniasis, or PKDL (WHO 2013; Pace 2014; Tschoeke et al. 2014).

Table 1: A summary of the different forms of leishmaniasis. Although each form can be caused by several *Leishmania* species, the most common are listed.

Type of leishmaniasis	Signs and symptoms	Location of symptoms	Typical causative agents
Cutaneous (CL)	Cutaneous lesions which can lead to severe scarring	Areas of skin exposed to sandfly bites such as the face, arms and hands	*L. major*
Mucocutaneous (MCL)	Destruction of mucosal membranes and nearby soft tissue, which can lead to disfigurement	Mucosal membranes, typically starting in the nose and progressing through the respiratory tract	*L. braziliensis*
Visceral (VL)	Fever, diarrhoea, organomegaly	Visceral organs, especially the liver and the spleen	*L. infantum/chagasi* and *L. donovani*
Post kala-azar dermal (PKDL)	Rash, papules, lesions, loss of pigmentation	Anywhere on the skin	*L. infantum/chagasi* and *L. donovani*

1.1.3 EVOLUTION AND TAXONOMY

Although technically eukaryotes, *Leishmania* diverged from other eukaryotes very early on. *Leishmania* are classified in the Kinetoplastea class, characterised by circular DNA organelles known as kinetoplasts. Along with human parasites *Trypanosoma brucei* and *Trypanosoma cruzi*, *Leishmania* are considered members of the order Trypanosomatida (Akhoundi et al. 2016).

Modern *Leishmania* taxonomy is still unresolved. However, there is some agreement on the division of the genus into subgenus *Viannia*, containing New World *Leishmania* species such as *L. braziliensis* and *L. guyanensis*, and subgenus *Leishmania*, which contains some Old World paleotropical species including *L. donovani, L. infantum* and *L. major*, and some New World neotropic species such as *L. mexicana* (Akhoundi et al. 2016).

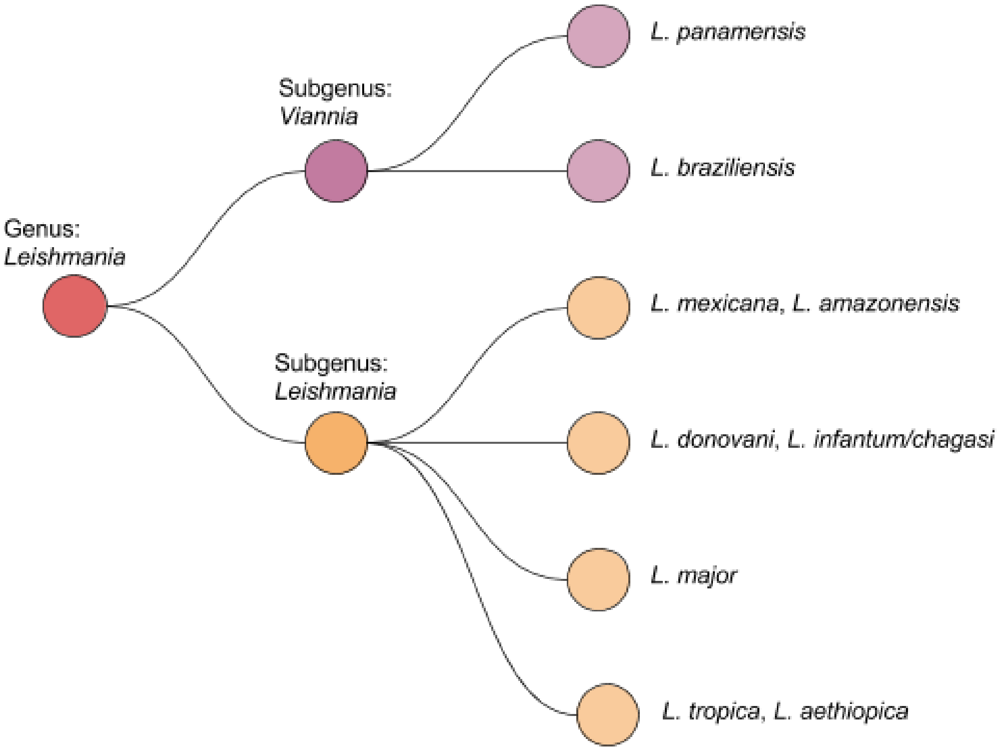


Figure 1. A simplified taxonomy of the major human-infective *Leishmania* species, adapted from Akhoundi et al. 2016.

1.2 LIFE CYCLE AND INFECTION

1.2.1 OVERVIEW

Leishmania species require two hosts to complete their life cycles, one mammalian, and one invertebrate. Across these stages, *Leishmania* may take the form of long, slender promastigotes in infective (metacyclic) and non-infective (procyclic) forms, or rounded intracellular amastigotes (Gluenz et al. 2010; Pace 2014).

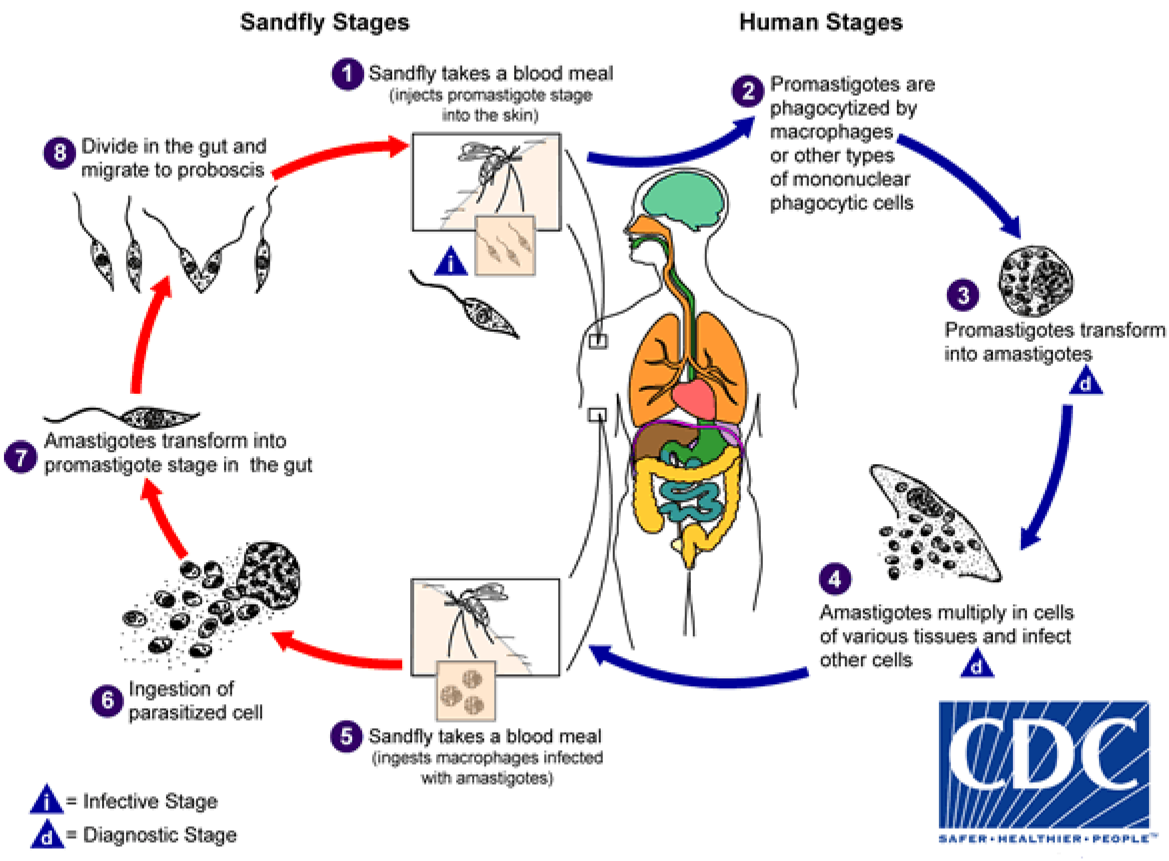

Sandfly Stages
Human Stages
1 Sandfly takes a blood meal (injects promastigote stage into the skin)
2 Promastigotes are phagocytized by macrophages or other types of mononuclear phagocytic cells
3 Promastigotes transform into amastigotes
4 Amastigotes multiply in cells of various tissues and infect other cells
5 Sandfly takes a blood meal (ingests macrophages infected with amastigotes)
6 Ingestion of parasitized cell
7 Amastigotes transform into promastigote stage in the gut
8 Divide in the gut and migrate to proboscis
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™

Figure 2. Taken from the centre of disease control and prevention (www.cdc.gov). Mammalian host stages are shown in blue, with sandfly stages in red.

Different life cycle stages show distinct morphological, biochemical and transcriptomic profiles. These changes reflect adaptation to various microenvironments in the vector and host (Cohen-Freue et al. 2007; Saxena et al. 2007; Franco et al. 2012; Moradin and Descoteaux 2012). As promastigotes differentiate into amastigotes, the parasites change shape, reduce their flagellum to a small tip, replace their surface lipophosphoglycan (LPG) coat with amastin, and undergo a number of gene expression changes, for example, changing their metabolic pathway preference from glucose and proline to amino acid and fatty acid beta oxidation (Gluenz et al. 2010; Franco et al. 2012; Fiebig et al. 2015). In order to tightly manage the differentiation between life cycle stages, gene expression is under fine control, and many subsets of genes are differentially expressed between stages. For example, genes upregulated in amastigotes are typically transporters, surface proteins, signalling molecules and cell growth related. Genes upregulated in promastigotes are mostly related to motility, respiration and biosynthesis (Saxena et al. 2007; Rastrojo et al. 2013; Fiebig et al. 2015).

Leishmania do not usually regularly undergo sexual reproduction, instead relying on a number of binary fissions at different life cycle stages to expand their population in the vector and the host (Killick-Kendrick 1990; Ravel et al. 2006). Despite the theoretical advantages of sexual recombination in a host-parasite system, there are costs associated with it, and *Leishmania* appears to prefer asexual reproduction. However, the presence of hybrid genotypes in some natural populations, as well as linkage equilibrium studies, suggest that sexual reproduction does occur, albeit at very low frequencies (Victoir and Dujardin 2002; Ravel et al. 2006; Kazemi 2011; Cantacessi et al. 2015).

1.2.2 LEISHMANIA INFECTION OF THE MAMMALIAN HOST

The first stage of the life cycle in the mammalian host begins when sandflies of the genus *Lutzomyia* (in the New World) and *Phlebotomus* (Old World) take a blood meal from a mammalian host and are induced to regurgitate a bolus of parasites, through parasite secretion of Promastigote Secretory Gel (PSG), into the bite wound (Reithinger et al. 2007; Pace 2014). Infective metacyclic promastigotes, previously inhabiting the fly gut, enter the site along with a complex mixture of immunomodulators (Maxwell-Silverman and Reiner 2012). Pro-inflammatory components of the inoculum, such sandfly saliva, recruit white blood cells to the site for predation, while *Leishmania*-secreted microvesicles known as exosomes are taken up by host cells to induce a pro-parasite phenotype and interfere with the developing immune response by targeting cell signalling pathways (Reithinger et al. 2007; Kaye and Scott 2011; Maxwell-Silverman and Reiner 2012; Atayde et al. 2015). The parasites manage to avoid complement lysis in blood through steric hindrance, achieved by long surface lipophosphoglycan (Franco et al. 2012).

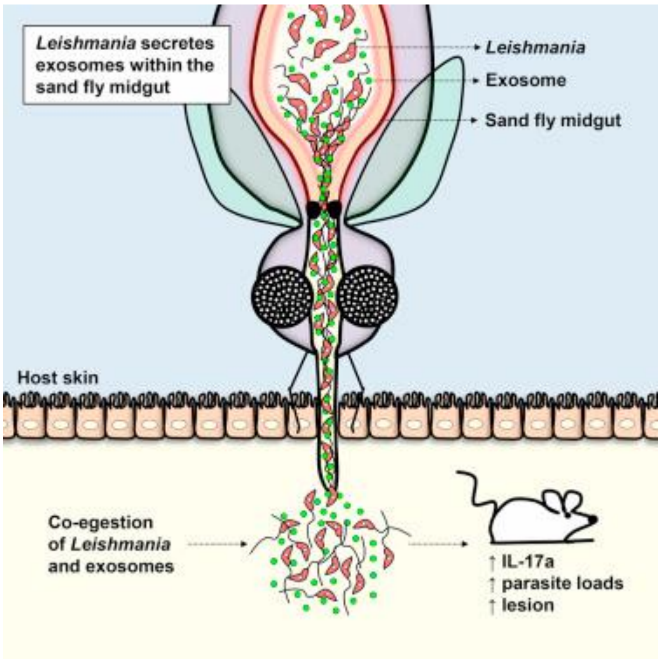

Leishmania secretes exosomes within the sand fly midgut
Leishmania
Exosome
Sand fly midgut
Host skin
Co-egestion of *Leishmania* and exosomes
↑ IL-17a
↑ parasite loads
↑ lesion

Figure 3. Taken from Atayde et al. 2015. The sand fly vector injects pro-inflammatory saliva components and *Leishmania*-secreted exosomes with the parasite bolus.

1.2.3 UPTAKE OF PARASITES INTO HOST CELLS

Through a variety of mechanisms, promastigotes are taken up into cells that are recruited to the invasion site, such as macrophages or dendritic cells (Reithinger et al. 2007; Kaye and Scott 2011; Maxwell-Silverman and Reiner 2012). Care must be taken by the promastigotes when entering the cells, as incorrect uptake can activate cellular antimicrobial defences such as reactive oxygen species production, leading to parasite death (Franco et al. 2012). Instead, promastigotes enter the cell through pathways that avoid triggering such reactions.

The major uptake pathways used by *Leishmania* rely on zippering phagocytosis mechanisms, and do not activate antimicrobial defences (Kaye and Scott 2011; Franco et al. 2012). For example, mannose-binding lectin binds to the *Leishmania* LPG coat, allowing for the uptake via the C3 convertase and C3b generation complement receptor pathway (Franco et al. 2012; Moradin and Descoteaux 2012). Several other pathways/receptors are also used, such as the mannose-fucose receptor, attachment of C-reactive protein to LPG and the CRP receptor, and recruitment of lysosomes to the cell surface via damage caused by flagellar beating in a manner similar to *Trypanosoma cruzi* (Forestier et al. 2011; Franco et al. 2012).

1.2.4 EVADING THE IMMUNE SYSTEM

An alternate non-activatory pathway into a host macrophage, known more commonly as the trojan horse mechanism, involves uptake into a non-macrophage white blood cell such as a neutrophil, inducing apoptosis, and being taken up by persisting in the apoptotic blebs into a macrophage, the preferred host cell (Kaye and Scott 2011). By hiding within the immune cells themselves, and producing low levels of pathogen-associated molecular patterns (PAMPs), *Leishmania* can effectively obscure their presence from the host immune system.

In addition to concealing themselves inside host cells, *Leishmania* are also master regulators of host immunity (Jaramillo et al. 2011; Kaye and Scott 2011). Gp63, a *Leishmania* surface protease, cleaves mTOR, a host kinase involved in translation regulation. The downstream consequence of this is a generalised inhibition of host protein synthesis, creating a more permissive environment for infection (Jaramillo et al. 2011). Other experiments have shown that *Leishmania* are able to manipulate the phenotype and activation status of their host cells in order to reduce production of antimicrobial NO_x species (Franco et al. 2012).

1.2.5 INTRACELLULAR SURVIVAL

The host cell is no safe haven, and the promastigotes must prevent fusion of the endosome that contains them with host cell lysosomes, which contain hydrolytic enzymes strong enough to destroy bacterial pathogens (Kaye and Scott 2011; Franco et al. 2012; Moradin and Descoteaux 2012). The promastigotes prevent fusion by preventing the maturation and acidification of the phagosome for long enough to differentiate into the next life cycle stage, developing into amastigotes which show immunity to the harsh environment they are contained within (Reithinger et al. 2007; Kaye and Scott 2011; Moradin and Descoteaux 2012). The amastigotes then develop the endosome into a parasitophorous vacuole, in which they draw nutrients from the host cell and multiply by binary fission (Reithinger et al. 2007). Pathology is thought to be caused by a complicated set of cellular and immune interactions including neutrophils, natural killer cells, and dendritic cells (Reithinger et al. 2007).

1.2.6 TRANSMISSION TO SANDFLY HOSTS

The next stages of the *Leishmania* life cycle occur in the sandfly host. Cells parasitised by amastigotes are ingested with a blood meal taken from an infected human host (Reithinger et al. 2007). Depending on their life cycle stage, *Leishmania* may inhabit a number of digestive tract microenvironments. Though typically considered only minor life cycle stages, paramastigotes are found inhabiting the mouthparts and oesophogus, haptomonads at the stomodeal/cardiac valve, and leptomonads and nectomonads can be found at various gut locations (Killick-Kendrick 1990; Schlein 1993).

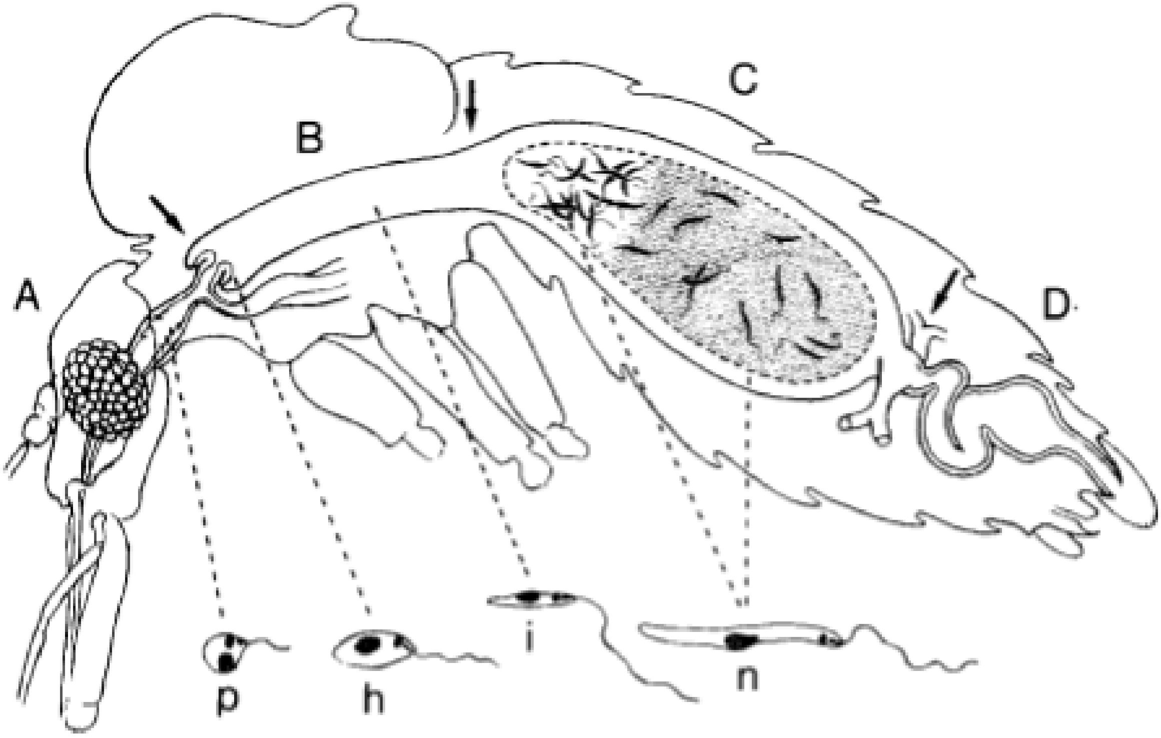


Figure 4. Taken from Schlein 1993. The gut of the sand fly is divided by the arrows into A, foregut, B, thoracic midgut, C, abdominal midgut and D, hindgut. The development of *Leishmania* promastigote stages are indicated as p, paramastigote, h, haptomonad, i, infective promastigote, and n, nectomonad.

Once inside the fly gut, contained within the midgut peritrophic membrane, amastigotes differentiate into non-infective procyclic promastigotes which undergo replication (Killick-Kendrick 1990). The parasite LPG surface coat grants protection against the hydrolytic digestive enzymes of the sandfly digestive tract (Franco et al. 2012). To avoid being excreted, promastigotes transform into nectomonads, which attach to the fly gut epithelium using flagella (Killick-Kendrick 1990; Schlein 1993; Gluenz et al. 2010). Galectins are expressed in the sandfly midgut, unique to each species. Parasite LPG is the critical molecule for attachment to galectin, which also shows polymorphism between species (Kamhawi et al. 2004; Franco et al. 2012; Abrudan et al. 2013). Since attachment to the midgut is essential for survival, the compatibility of parasite LPG to host galectin dictates vector species tolerance (Kamhawi et al. 2004). In order to migrate to the foregut, the nectomonads differentiate into leptomonads, which replicate again before finally transforming into metacyclic promastigotes, ready for infection (Killick-Kendrick 1990; Schlein 1993).

1.3 DIAGNOSIS, TREATMENT AND CONTROL

1.3.1 OVERVIEW

Current treatments and strategies for control are dependent upon region (Rai et al. 2013; WHO 2013). Diagnosis relies mostly on direct visualisation of parasites through culture or microscopy, or through detection of parasite DNA through PCR-based methods. Anti-*Leishmania* antibody serology tests exist with mixed reliability (Reithinger et al. 2007; WHO 2010; Moore and Lockwood 2011; WHO 2013). The most commonly used drug treatments include pentostam, glucantime, amphotericin B, pentamidine, paromomycin and miltefosine for both CL, VL and their variants (Croft et al. 2006; Perry et al. 2013; Pace 2014). Treatments are often expensive and considered a major contributing factor in the poverty of affected patients. Patients are often too sick to work and thus unable to afford treatments; leishmaniasis also leaves patients susceptible to opportunistic infections. (WHO 2013; Okwor and Uzonna 2016).

Table 2: An overview of key treatments for leishmaniasis.

Drug treatment	Disease	Mechanism	Caveat
Pentavalent antimonials e.g Pentostam and glucantime	Cutaneous, visceral and mucocutaneous	Currently unknown; thought to be related to inhibition of metabolism and protein synthesis	Severe hepatic and renal toxicity; painful administration
Amphotericin B	Visceral and cutaneous	Induction of pore formation in membranes	Expensive in liposomal form
Pentamidine	Second-line for visceral, cutaneous and diffuse cutaneous	Inhibition of polyamine synthesis and mitochondrial membrane potential interference	Resistance develops quickly, so treatments must be used sparingly
Miltefosine	Visceral, cutaneous and diffuse cutaneous	Interference with apoptosis and sterol biosynthesis	High relapse rate, particularly on the Indian Subcontinent

1.3.2 CURRENT TREATMENT PROGRAMMES

Pentavalent antimonials such as pentostam (sodium stibogluconate, or SSG) and glucantime (meglumine antimoniate) have been used to treat leishmaniasis for over 60 years (Croft et al. 2006). The drugs are administered through injection, and have severe toxic side effects (Reithinger et al. 2007; Perry et al 2013). The exact mechanisms through which antimony-based treatments are thought to work are currently unknown, though research suggests inhibition of protein synthesis, glycolysis and metabolism (Croft et al. 2006; Perry et al. 2013).

As with many anti-parasitic drugs, amphotericin B was initially developed as an antifungal before being adapted for anti-leishmanial use. The major sterol in mammalian cells is cholesterol, but in target organisms such as *Leishmania*, the major sterol is ergosterol instead; Amphotericin B has a high affinity for ergosterol, binding and causing pore formation in membranes (Croft et al. 2006).

Pentamidine has been used as a second-line treatment for VL, CL and diffuse CL, and is thought to have antimicrobial activity through inhibition of polyamine synthesis and interfering with mitochondrial membrane potential (Croft et al. 2006).

The antibiotic paromomycin works through an unknown mechanism in *Leishmania*, though in bacteria it interferes with biosynthesis through binding to the 16S rRNA (Croft et al. 2006).

Originally trialled as an anti-cancer drug, miltefosine interferes with sterol biosynthesis pathways to kill *Leishmania* parasites (de Morais-Teixeria et al. 2011; Rai et al. 2013).

1.3.3 APPLICATION OF COMBINATION THERAPIES

Although initially used alone, many drugs are now being used in combination therapies. A number of advantages exist for combination therapies over monotherapies, for example, shorter treatment regimens using smaller doses are typically cheaper and more effective than monotherapies (van Griensven et al. 2010).

Smaller doses of each individual drug reduce the associated toxic side effects and shortens the treatment duration, increasing treatment regime compliance. Combination therapy is also associated with lower rates of mortality and complications, when compared with single drug treatments (van Griensven et al. 2010). Combination therapies may also be more effective in patients with co-infections such as TB or HIV, where ‘classical’ single-drug treatments often fail (van Griensven et al. 2010; Trinconi et al. 2014).

Treatments together can have additive effects, working synergistically to achieve better results than they would have individually. While not all treatment combinations achieve this effect, some display a high level of activity enhancement, for example, amphotericin B and miltefosine, or SSG with paromomycin (Croft et al. 2006; Trinconi et al. 2014).

1.3.4 TREATMENT RESISTANCE

As with antibiotics, the time from deployment resistance is highly variable, and dependent on a number of factors. Some treatments, such as pentavalent antimonials, have been used as a front-line drug for over 60 years with resistance only developing recently - whereas some much more recently licensed treatments, paromomycin or miltefosine, for example, had reports of resistance almost immediately (Croft et al. 2006).

The majority of resistance is thought to be attributable to overuse and misuse of drugs. Widely used pentavalent antimonials are available over-the-counter which is likely a major contributor to overuse (Croft et al. 2006; van Griensven et al. 2010). A large proportion (73%) of patients were found to seek medical advice from non-qualified practitioners, who may give inappropriate usage instruction. One survey found that only 26% of patients followed official treatment guidelines, for example, patients often took breaks from treatment citing fears about renal toxicity (Reithinger et al. 2005; Croft et al. 2006). Poor healthcare infrastructure in countries where leishmaniasis is common means that delays in diagnosis and treatment are common, allowing infection to more thoroughly establish and making treatment less effective (van Griensven et al. 2010).

Combination therapy is a common approach to tackling resistance in infectious disease, for example, use of dual therapies is becoming increasingly common in the treatment of malaria (Trinconi et al. 2014). In the case of leishmaniasis, use of combination therapies to deter the development of resistance is expected to be relatively effective, given that current drug treatments span a number of chemical classes and targets. A consequence of delaying the appearance of resistance using combination therapies is the extension of the usefulness of current drugs as effective treatments (van Griensven et al. 2010). Ideally, however, resistance can be dealt with by simply developing new drugs for new targets, to help avoid cross-resistance (Croft et al. 2006).

1.3.5 TREATMENT EFFICACY

A number of host factors are known to influence treatment efficacy, such as pharmacokinetics and immune status. (Reithinger et al. 2007; Moore and Lockwood 2011; Perry et al. 2013; Rai et al. 2013; Fernandes et al. 2016) In particular, pentavalent antimonial drugs and pentamidine appear affected by the host T-cell response, which is important in HIV co-infections and other patients suffering from immunosuppression. Other drugs such as miltefosine and amphotericin B appear unaffected by T-cell response (Croft et al. 2006).

Leishmania factors may also affect treatment potency (Croft et al. 2006; Rai et al. 2013; Fernandes et al. 2016). Natural variance in species and strain will affect the parasite’s ability to tolerate drug pressure, for example, antimonial drugs such as glucantime appear to be especially effective against some species (Croft et al. 2006). A number of mechanisms exist to cope with drugs such as decreased uptake, increased export, metabolic inactivation and sequestration (Croft et al. 2006; Perry et al. 2013).

1.3.6 APPROACHES TO VACCINES AND CURRENT DEVELOPMENTS

Unusually for a parasitic disease, leishmaniasis is thought to be controllable by vaccination (Kedzierski et al. 2006; Gillespie et al. 2016). Compared with other parasitic diseases, *Leishmania* has a relatively simple life cycle, and patients show general resistance to reinfection (Kedzierski et al. 2006). Leishmanisation, a method used historically in the Soviet Union, Iran and Israel, involving injection of live parasites, has seen recent safety measures introduced such as drug sensitivity and suicide genes for a controlled infection. Infections were typically performed on the leg to avoid facial scarring (Kedzierski et al. 2006; Reithinger et al. 2007; Kumar and Engwerda 2014; Gillespie et al. 2016).

More modern attempts at developing a vaccine have considered DNA vaccines, recombinant proteins, sandfly salivary proteins, and attenuated parasites (Kedzierski et al. 2006; Reithinger et al. 2007; Kumar and Engwerda 2014). Use of dead promastigotes in clinical trials were deemed safe, but ultimately failed to induce a protective immune response. Live, attenuated parasites can allow for the development of a full immune response, but the method through which they are attenuated may cause rapid elimination from the body, preventing such a response progressing (Kedzierski et al. 2006; Kumar and Engwerda 2014). There is always a minor risk in using attenuated parasites that they may regain pathogenic capacity.

No vaccine is currently licensed for leishmaniasis, though several have progressed to various stages of clinical trial. LEISH-F2 and LEISH-F3 are based on antigen epitopes expressed in bacteria, and have both progressed to phase II trials. (Gillespie et al. 2016). Ad5-KH uses an adenovirus vector to deliver two *Leishmania* antigens, and is also currently undergoing clinical trials (Maroof et al. 2012). Other vaccines, such as LiESAp-MDP and Leish-111F, show some potential but are limited by parasite genetic variation - the vaccines only provide protection against species that are genetically close to the species from which the vaccine was generated (Kumar and Engwerda 2014).

1.3.7 VECTOR CONTROL, ECOLOGICAL CONTROL, ALTERNATIVE STRATEGIES

A wide variety of strategies are applied to control both CL and VL, largely relying on insecticides and reducing potential reservoirs.

For some *Leishmania* species, humans are the sole reservoirs, such as *L. donovani* and *L. tropica* (WHO 2010). In the case of many others, particularly in the Mediterranean, stray dogs and a variety of other wild mammals may act as reservoirs (Kedzierski et al. 2006; WHO 2010; Pace 2014; WHO 2016). Mass culling of stray dogs has historically been shown to be ineffective. Instead, to control spread via dog reservoirs, insecticide-impregnated collars has been used to great success in reducing VL incidence in countries such as Brazil and Iran (WHO 2010; Cantacessi et al. 2015). Destruction of reservoir habitats proximal to settlements has also been effective, such as the destruction of gerbil burrows (WHO 2010).

Vector control has many facets. WHO recommends an integrated approach in order to interrupt leishmaniasis transmission, using chemicals, environmental management and personal protection (WHO 2010). Use of chemical insecticide sprays in domestic and animal housing is an effective way to reduce exposure to sandflies, in addition to impregnated and sprayed nets (Pace 2014).

1.4 GENE EXPRESSION SYSTEMS IN KINETOPLASTIDS

1.4.1 OVERVIEW

Despite being taxonomically eukaryotic, *Leishmania* and other kinetoplastids show highly divergent genome organisation and regulation of gene expression when compared with features associated with classical eukaryotes (Kazemi 2011; Rastrojo et al. 2013; Tschoeke et al. 2014; Fiebig et al. 2015).

1.4.2 GENOME ORGANISATION

Kinetoplastids show highly divergent methods of controlling gene expression when compared with other eukaryotes, and this is reflected by their genome organisation. Genes are arranged in tandem arrays, typically containing between 10 and 100 genes not necessarily related by function or homology. These genes lack introns, promoters and enhancers, and are transcribed together as a single giant polycistron (Clayton and Shapira 2007; Saxena et al. 2007; Kramer 2012; Cantacessi et al 2015). In addition to unusual organisation, polycistronic transcripts also undergo unusual post-transcriptional processing. Mature mRNA is generated through trans-splicing, a process which attaches a 39nt ‘splice leader’ (SL) sequence to the 5’ end of the transcripts, and polyadenylates the 3’ end (Clayton and Shapira 2007; Kazemi 2011; Siegel et al. 2011; Kramer 2012; Rastrojo et al. 2013).

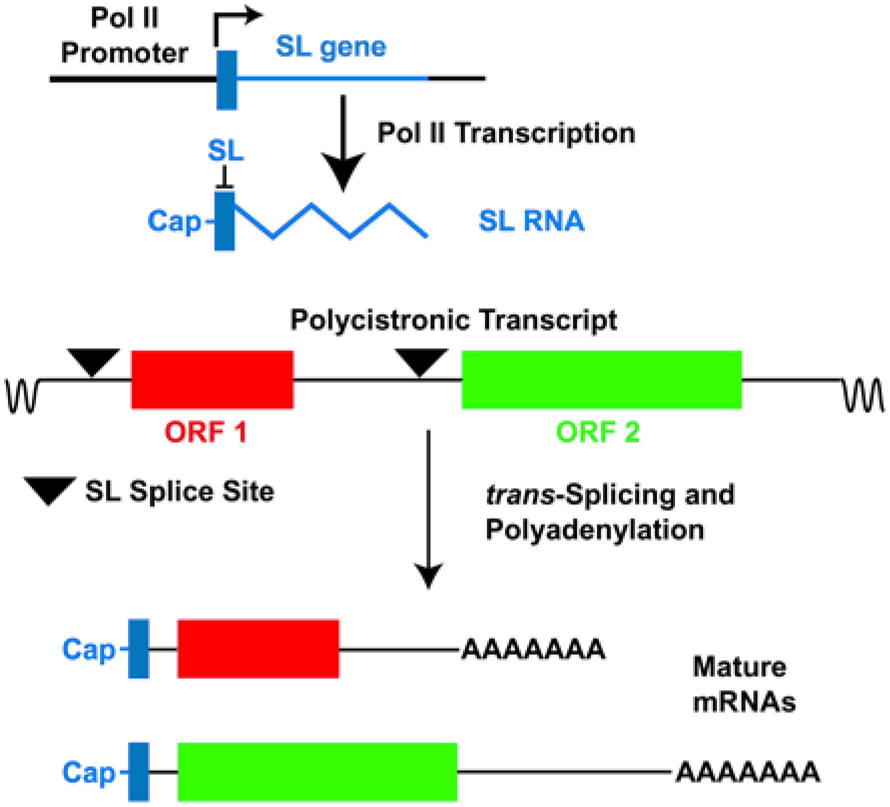


Figure 5. Taken from Landfear 2003. Splice leader (SL) RNAs are transcribed from SL genes. Multiple ORFs are transcribed together in a polycistronic transcript, which are separated into individual mature mRNAs through trans-splicing. This process involves the addition of the SL to the 5' end of the transcript, and polyadenylation of the 3' end.

1.4.3 REGULATION OF TRANSCRIPTION

Around 8% of the genes in the human genome code for DNA-binding proteins, on top of other DNA-regulatory elements such as transcription factors, promoters and enhancers (Kramer 2012). In kinetoplastids, chromatin-remodelling enzymes are massively under-represented, with RNA polymerase II promoters and enhancers are almost entirely absent as protein-coding gene expression is regulated through alternative methods (Clayton and Shapira 2007; Kazemi 2011; Cantacessi et al. 2015). Instead, transcriptional start and stop sites are marked by epigenetic variation such as histone protein variants, base variants such as base J, and histone modifications (Siegel et al. 2011; Van Luenen et al. 2012). Exceptions to the absence of promoter elements do exist - the only well characterised RNA polymerase II promoter is for the splice leader RNA gene, for example (Kazemi 2011).

1.4.4 A COMBINED APPROACH TO GENE REGULATION

Due to the general lack of promoter and enhancer elements, polycistron transcription occurs at approximately the same rate, causing a basal level of gene transcription (Kramer 2012; Fiebig et al. 2015). Given that kinetoplastids are still able to react to environmental stimuli and undergo differentiation, gene regulation must occur post-transcriptionally (Clayton and Shapira 2007). This is likely achieved through a combination of mRNA stability mediated by elements in the 3' UTR, signalling cascades, riboswitches and RNA thermometer-like elements, protein localisation, stability and degradation, and phosphorylation events (Clayton and Shapira 2007; Cohen-Freue et al. 2007; Siegel et al. 2011; Kramer 2012; Cantacessi et al. 2015).

1.4.5 GENOMIC FLEXIBILITY

The first *Leishmania* genome to be sequenced was *L. major*, revealing a genome of 32.8Mbp, with 8,311 protein-coding genes predicted (Ivens et al. 2005; Tschoeke et al. 2014). The genomes of *Leishmania* species show an unusually high degree of conservation when compared to other microbes that have been diverged for similar periods of time; *L. major*, *L. infantum* and *L. braziliensis* average 8,300 protein-coding genes, of which 99% were highly syntenic (Peacock et al. 2007; Mannaert et al. 2012; Tschoeke et al. 2014). Very few unique genes exist among each species, with *L. braziliensis*, *L. infantum* and *L. major* showing 47, 27 and 5 unique genes respectively (Tschoeke et al. 2014).

Old world species of *Leishmania* such as *L. major*, *L. donovani* and *L. infantum* have 36 chromosomes; however, this is not true for all *Leishmania* species, as others such as *L. braziliensis* and *L. mexicana* have 35 and 34 respectively, due to chromosome fusion events. In *L. mexicana* group species, chromosomes 8 and 29, and 20 and 6 have fused; in *L. braziliensis* group species, 20 and 34 have fused (Ivens et al. 2005; Peacock et al. 2007; Mannaert et al. 2012; Lachaud et al. 2014). The exact number of chromosome copies varies between chromosome, species, strain and population, but the haploid genome totals between 29 and 33 Mb in size (Cantacessi et al. 2015). Across 6 chromosomes studied using fluorescent in-situ hybridisation, *L. infantum*, *L. tropica* and *L. amazonensis* appear primarily disomic, while *L. donovani* is more heterogeneous, with different chromosomes showing monosomy, disomy and trisomy (Lachaud et al. 2014). *L. braziliensis* is primarily triploid, with some tetrasomic chromosomes, and a single hexasomic chromosome (Mannaert et al. 2012).

In addition to their unusual genome organisation and expression system, *Leishmania* are also unusual even within trypanosomatids in their tolerance to chromosome number variation and genomic plasticity (Kazemi 2011; Cantacessi et al. 2015; Fiebig et al. 2015). In most eukaryotes, euploidy such as haploidy or diploidy is common, and variation is often indicative of disease. Through gene dosage consequences, changes in ploidy can be severely detrimental, even lethal. However, *Leishmania* appear tolerant to such changes, with no measurable consequence in terms of cell growth (Mannaert et al. 2012; Lachaud et al. 2014).

Populations of *Leishmania*, even within the same infection, exhibit mosaic aneuploidy, in which chromosome copy number variation exists between individuals, from monosomy to pentasomy. (Lachaud et al. 2014; Cantacessi et al. 2015). A number of potential functions have been suggested for the amplification of genes or chromosomes in *Leishmania*. These changes in gene dosage can be brought about through whole chromosome duplication, whole genome duplication, or just a translocation of the target gene to another chromosome (Mannaert et al. 2012). The ability of *Leishmania* to vary their genotype/karyotype, and therefore their phenotype, within a clonal population is considered an inherent advantage (Sterkers et al. 2012).

Variation in gene dosage by ploidy or copy number change is likely a useful mechanism in adapting to different environments and microenvironments, in response to the changing conditions within a host (Mannaert et al. 2012; Sterkers et al. 2012; Rogers et al. 2014). It has also been suggested that gene dosage may contribute to the differences in tissue tropism between *Leishmania* species, as well as regulating pathogenicity and virulence (Sterkers et al. 2012; Cantacessi et al. 2015)

Several studies have noted chromosomal amplification as a response to the presence of anti-leishmanial drugs, both in vivo and in vitro. Deliberate induction of drug resistance is associated with chromosomal and genes-specific amplification in experimental studies, though the chromosomes amplified in response to drug resistance induction have been found to revert in some cases, after the pressure in question was relieved (Mannaert et al. 2012; Lachaud et al. 2014). Resistance to sodium arsenate and methotrexate has been observed with gene amplification between 2 and 20-fold. Amplifications have also been implicated in the resistance of Amphotericin B and pentostam (Kazemi 2011). The extent of resistance may correlate with the number of copies of the gene conveying resistance; this would agree with other experimental data implying that a correlation between post-transcriptional dosage and chromosome copy number exists (Downing et al. 2011; Mannaert et al. 2012).

The mechanisms underlying ploidy and copy number variation changes in *Leishmania* are poorly characterised, though techniques exist to study changes at both single cell and population level, such as fluorescent in-situ hybridisation, or FISH (Sterkers et al. 2012). Given the inconsistency of changes in ploidy, the mechanism is unlikely to be a simple cause-and-effect – i.e the mechanism will likely involve multiple steps, proteins, and genes (Rogers et al. 2014).

1.5 MICE AS MODEL ORGANISMS FOR IMMUNOLOGICAL STUDY

1.5.1 OVERVIEW

Mus musculus domesticus is the subspecies of house mouse commonly used as a model animal in scientific laboratories. Considered an "unsung hero" of biology, the contribution mice have made to our current understanding of immunology is immense (Viney et al. 2015). Commonly used strains such as Black 6 and BALB/c mice were derived from the fancy mouse trade during the 1920's, but have undergone considerable selective pressure to increase growth rate and reduce the age at which they sexually mature (Viney et al. 2015).

In terms of the structure of their immune system, mice are broadly similar to humans, with both innate and acquired branches, and analogous organs (Metas and Hughes 2004). However, the mouse immune system is less aggressive than the human immune system; mice tend to tolerate pathogens rather than attempt clearance, in order to reduce immunopathology (Zschaler et al. 2014). However, when compared with wild mice, laboratory mice are known to have hyperreactive cytokine responses to the presence of pathogens, likely due to a lack of natural exposure to such threats. The relative lack of response in wild mice is likely advantageous, as the reduced inflammation avoids immunopathology (Viney et al. 2015; Abolins et al. 2017).

Although the overall architecture of the mouse immune system is akin to that of humans, there are some significant differences. Across the entire genome, only 75% of mouse genes have 1:1 human orthologues, and immune genes are no exception (Belizário 2009). Cytokines show differential functions, as well as several human cytokines being absent, and novel cytokines being present in mice (Zschaler et al. 2014). The proportions of immune cells are also different; mice blood is lymphocyte rich, consisting of 70-95% lymphocytes and 10-25% neutrophils (Metas and Hughes 2004). Laboratory mice show a low proportion of activated CD4+ and CD8+ T-cells, B-cells, macrophages and DCs when compared with wild mice, likely as a consequence of their sterile environment (Abolins et al. 2017). Patterns of surface proteins, such as CD markers and toll-like receptors, also vary when comparing human and mouse immune cell types; cells undergo slightly different development and maturation processes (Metas and Hughes 2004). Mice make IgA, IgD, IgE and IgM, but have several distinct subclasses absent in humans (Metas and Hughes 2004). As with the ratio of activated cells, serum concentrations of immunoglobulins in laboratory mice have been found up to 370x lower than their wild counterparts (Abolins et al. 2017).

1.5.2 INFLUENCE OF STRAIN ON IMMUNOLOGY

Despite considerable inbreeding, basic, unmodified laboratory mouse strains are considered immunocompetent, displaying extensive variation on immune phenotype. Differences between strains can lie in single point mutations or through variation in complex, multilocus traits (Sellers et al. 2012). Strains are known to vary both in components of the innate immune system, such as pattern recognition receptors (Toll-like receptors, C-lectin receptors, etc.), and in the adaptive immune system, in proportions of T-cells and B-cells, and the use of regulatory microRNAs (Sellers et al. 2012). For example, Black 6 mice disproportionately respond with Th1 cells, known for their importance in the clearance of intracellular pathogens, which other strains, such as BALB/c, have a Th2 bias in their response. These differences among strains amount to differences in susceptibility, resistance, response and disease progression (Sellers et al. 2012).

Use of immunodeficient mice has been paramount in the study of reductive immunology. Defects in the immune system can be natural, transgenic, genetically engineered, or induced through mutation (Belizário 2009). Typically, immunodeficient strains have faulty major histocompatibility complexes, defective T-cells or B-cells, or knockouts of particular cytokines, receptors or transcription factors (Belizário 2009). Genetically engineered mice with defective adaptive immune systems commonly have issues with chronic inflammatory disorders, such as colitis (Sellers et al. 2012).

RAG mice are immunodeficient knockout mice with defective adaptive immune systems. The RAG genes (recombination activating genes), as the name suggests, are DNA recombinases crucial in the activation of VDJ recombination, used in the generation of T- and B-cell receptor diversity, and the production of immunoglobulins. Faults in receptor and immunoglobulin generation prevent the development and maturation of T-cells and B-cells, but do not affect innate immunity or other physiological or behavioural aspects (Mombaerts et al. 1992; Shinkai et al. 1992; Thompson 1992; Belizário 2009). Generation of RAG mice involves the deletion of either of the *Rag1* or *Rag2* genes (Mombaerts et al. 1992; Thompson 1992; Belizário 2009). RAG mice have been used in a wide range of studies, investigating lymphocyte differentiation, immune response to tumourigenesis, metastasis, autoimmunity, chemotherapy and infectious disease (Belizário 2009).

1.5.3 MICE IN THE STUDY OF VISCERAL LEISHMANIASIS

Mice have been used as model animals in the study of leishmaniasis for over 40 years. Given how relatively easy they are to keep and to breed, experiments involving captive mice also restrict the influence of the environment (Liplodová and Demant, 2006; Loría-Cervera and Andrade-Narváez, 2014). A wide array of topics within leishmaniasis infection biology have been studied using mouse models, such as cytokines, cell types, signalling cascades, antileishmanial defences, disease progression, and vaccine development (Loría-Cervera and Andrade-Narváez, 2014). However, the data from mouse studies should not be directly used to support conclusions about human disease - not only are the host species different, genetically and immunologically, but experimental conditions do not mimic those of natural infections, and as such the results should not be expected to behave as a natural human infection would (Loría-Cervera and Andrade-Narváez, 2014; Loeuillet et al. 2016).

A number of experimental design choices can affect end results, which may to some extent explain discrepancies in data even when using the same animal model (Loría-Cervera and Andrade-Narváez, 2014; Loeuillet et al. 2016). The genetic background of both the host and the parasite - from species to strain - is known to influence disease outcome. The transmission of parasites is critically important to the disease manifestation. Site of inoculation (e.g subcutaneous, intraperitoneal), inclusion of immunogenic sandfly saliva, parasite life cycle stage, origin of parasites in terms of culture, and parasite dose have all been documented to influence transmission success (Loría-Cervera and Andrade-Narváez, 2014; Loeuillet et al. 2016). For example, the number of parasites administered can change the course of an infection. In BALB/c mice, low numbers of parasites inoculated subcutaneously caused a Th1 response leading to disease resolution. A higher dose of parasites was found to induce a Th2 response, causing chronic disease (Ahmed et al. 2003).

Despite the potential for inconsistent results, mouse work on leishmaniasis has yielded a number of important discoveries (Liplodová and Demant, 2006; Loría-Cervera and Andrade-Narváez, 2014). For example, in mice, disease outcome can largely be predicted by measurable immunological features. A Th1 immune response leads to the production of interferon gamma, a pattern seen in leishmaniasis-resistant mice. Susceptible mice instead show a Th2 response, producing IL-4 and allowing disease progression (Ahmed et al. 2003).

Mice have also been used to study the effect of host genetic background on infection, and to investigate how specific phenotypic traits and genes relate to infections. Mice have extremely well-studied genomes, and the genetic makeup of each strain is well known, making mice excellent candidates for studying genetic background (Liplodová and Demant, 2006; Loría-Cervera and Andrade-Narváez, 2014).

Susceptibility to leishmaniasis in mice was found to be multigenic, especially for CL. Two genes in particular were found to be important in determining resistance to infection; *Slc11a1* and *H2*. *Slc11a1* is known to mediate the host's early defence against infection, found at the Lsh locus on mouse chromosome 1. Wild type mouse strains such as CBA show resistance associated with this locus, but BALB/c and C57BL/6 both appear to have mutations in this gene, causing susceptibility (Bradley et al. 1979; O'Brien et al. 1980). *H2* encodes the Major Histocompatibility Complex (MHC), involved in antigen presentation and complement defences. In mice, *H2* decides the progression of late leishmaniasis, given its involvement in the adaptive immune system. A number of different *H2* alleles exist in mice, some of which confer resistance to leishmaniasis even in strains designated susceptible by their *Slc11a1* mutation (Roberts et al. 1997).

1.7 PROJECT CONTEXT: CRACK IT

1.7.1 OVERVIEW

The CRACK IT initiative awards grants to research that facilitates replacement, reduction and refinement of animal use in research and business, organised and granted by the national council for the 3 Rs. Under the project title of 'Virtual Infectious Disease Research', 5 experiments took place to feed data and analyses into a computational simulation of *Leishmania* infection, intended to reduce the number of animals used by replacement with a hypothesis-testable model. This project will analyse samples generated by the fifth CRACK IT experiment, which examines the effect of host immune pressure on *Leishmania*; previous experiments on *Leishmania* have explored pharmacokinetics, immune response and the effect of parasite strain on the host. This particular experiment was designed to study the effect of host immunocompetence and genetic background on both the host and parasite, using genomic and transcriptomic data.

1.7.2 MODELLING INFECTION

The previous model of *Leishmania* infection was built using Petri Net mathematical modelling, intended for use in drug evaluation. The model, based on *L. donovani* infection, simulated the formation of granulomas in the liver, and was highly accurate in estimating experimental results compared to data generated in vivo mouse infections (Albergante et al. 2013; Timmis et al. 2016).

Data and analyses from the CRACK IT 5 experiment will be fed into a new model, which is intended to simulate *L. donovani* infection in spleen tissue. The spleen was chosen as the focus instead of the liver as the spleen is considered clinically more relevant, and a more useful indicator of clinical disease.

Although similar, there are key differences between the immune systems of humans and mice. Computational models can still be useful in understanding disease such as infection features, expected cell population dynamics and downstream consequences of changes.

1.8 PROJECT AIMS

The nature of the data, together with the choice of processing and analysis methods, allows for simultaneous investigation of both host and parasite transcriptome and genome.

- To characterise the differences in the transcriptomes of wild type and immunocompromised RAG2 KO mice.
- To compare the transcriptomes of uninfected mice with those infected with *L. donovani*.
- To ascertain the relationship between the RNA profile of *L. donovani* inoculum samples and parasites isolated after 28 days infection of a RAG2 KO mouse
- To examine the effect of an adaptive immune system on the parasites can be seen by comparing the samples from WT and RAG mice.

CHAPTER 2: METHODS

2.1 INOCULUM PRODUCTION AND INOCULATION

2.1.1 PARASITE CULTIVATION

Dr Helen Ashwin performed all animal work, parasite isolation, and RNA extractions. One 6-8 week old B6.RAG2KO.CD45.1Cg mouse was infected with *Leishmania donovani* amastigotes via lateral tail vein intravenous injection. After 3 months, the mouse was euthanised and its spleen harvested.

2.1.2 INOCULUM PREPARATION

Extracted spleens were stored in incomplete RPMI media. A glass homogeniser was used to gently lyse the tissue to a single-cell suspension. The suspension was transferred into a 50ml tube and RPMI media was added until the volume equalled 20ml. The tube was centrifuged at 800 RPM for 5 minutes, after which the pellet was discarded but the supernatant retained. The amount of supernatant was measured and a clean 50ml falcon tube was internally coated with 25mg of saponin per 20ml supernatant. The supernatant was added to the saponin-coated tube and mixed. After being left for 5 minutes at RT, the tube was centrifuged at 3100 RPM for 10 minutes. The pellet was examined for RBCs; if the RBCs were present, the previous two steps were repeated. Else, the supernatant was discarded and the pellets underwent 3 wash steps by resuspension in 25ml RPMI and then centrifuging for 10 minutes at 3100 RPM. After the final wash step, the pellet was resuspended in 20ml RMPI media. Keeping the needle inside a falcon tube to prevent contamination, the suspension was passed through a 26-guage (brown) needle 2-3 times, using a 10ml syringe, to break up clumps of parasites. The amastigotes were counted using a sterile parasite counter, and resuspended to a concentration of 1.5 x 10^8 parasites per ml. The inoculum was kept at 37°C to prevent the amastigotes differentiating into promastigotes.

2.1.3 INFECTION PROCEDURE

A total of 19 mice were used in this experiment (table 3). 6-8 week old mice were inoculated by lateral tail vein intravenous injection, with 200µl of inoculum (containing approximately 3 x 10^7 *L. donovani* amastigotes), which was first passed through a needle as in 2.1.2 to prevent clumping. Additional mice were left uninfected to serve as controls. Mice were not treated with antibiotics to avoid aberrant inflammatory and immune responses. 28-days post infection the mice were euthanised and their spleens harvested.

Table 3. An overview of the mice used in the experiment.

		Mouse Background	
Infection status		B6.CD45.1 (WT)	B6.RAG2KO.CD45.1Cg (RAG)
	Infected	5	5
	Uninfected	5	4

2.1.4 PARASITE BURDEN CALCULATION

After being harvested, spleens and livers from infected mice were tested for parasite burden. Each organ was weighed and cross-sectioned; the cut face of the tissue was dabbed against a glass microscope slide to leave an impression. The slide was stained with Giemsa in order to differentiate mouse cells and parasite cells. 1,000 mouse cells are counted for their associated amastigote count, which is then used to calculated Leishman-Donovan Units (LDU). LDU are the standard measure of tissue parasite burden for *Leishmania* infections. LDU can be calculated according to the following formula:

$$LDU = \frac{amastigote\ count\ per\ 1{,}000\ host\ cell\ nuclei}{organ\ weight\ (g)}$$

2.2 RNA SAMPLE COLLECTION AND CLEANING

2.2.1 INOCULUM RNA COLLECTION

Parasites were semi-purified from spleen samples from a mouse prepared in the manner described in 2.1.1. using the isolation procedure described in 2.1.2. RNA was extracted from purified *L. donovani* parasites using a Zymo Research Direct-zol RNA MiniPrep Kit, per manufacturer's instructions.

2.2.2 MOUSE RNA COLLECTION

RNA extraction from mouse spleen, using 5mg of tissue per sample, was performed using Qiagen miRNeasy micro kit (ID: 217084), per manufacturer's instructions. Spleen tissue collected from infected mice will contain both mouse and *Leishmania* RNA given the presence of the parasite in the host tissues.

2.2.3 DNASE TREATMENT OF RNA SAMPLES

DNAse treatment of RNA extracts was performed using Qiagen RNAse-free DNAse kit for DNAse treatment (ID: 79254), per manufacturer's instructions.

2.2.4 MAGNETIC BEAD CLEANING OF RNA SAMPLES

RNA samples were cleaned after DNAse treatment using AMPure XP beads (ID: A63881). The beads were warmed to room temperature for 30 minutes, and resuspended by gently vortexing. Beads were added to the sample at a volume ratio of 1.8:1 beads to sample, e.g. for 50µl of sample add 90µl beads. After a gentle vortex, the suspension was left for 5 minutes to allow binding. The tubes were placed on a magnetic separation rack and left for 3 minutes. With the tubes still on the rack, the supernatant was carefully removed without disturbing the beads. To wash the beads, the tube was filled with 70% ethanol, and left for 30 seconds. Then, the ethanol was removed and fresh ethanol was added again for a total of 2 wash steps. After washing, the beads were left to air dry for around 5 minutes, but observed to prevent over-drying and cracking. When the ethanol was fully evaporated, the beads were resuspended in 50µl of water or Tris EDTA (TE) buffer, first using a pipette, and then a vortex. The sample was left for 3 minutes then placed back onto the magnetic rack until the solution was clear. The clear supernatant – containing the cleaned RNA - was transferred to a clean tube.

2.3 MEASURING RNA/DNA QUALITY

Several different instruments were used to measure RNA and cDNA quality throughout the RNA cleaning and library prep process.

Before and after the RNA samples were cleaned using the magnetic bead protocol, the samples were run on an Agilent 2100 BioAnalyser, using Agilent RNA 6000 Pico kits and chips, per manufacturer's instructions. Additionally, after being prepared into cDNA libraries, the BioAnalyser was also used to measure DNA quality using Agilent High Sensitivity DNA kits and chips, per manufacturer's instructions.

The NanoDrop 1000 spectrophotometer was used per manufacturer's instructions to screen RNA samples for contaminants using 230:260 and 260:280 ratios; the Qubit 3.0 and Qubit HS RNA assay kits were used for more precise quantification.

2.4 LIBRARY PREPARATION

After cleaning and dilution, the RNA samples were prepared into cDNA libraries for Illumina sequencing by myself and Dr Sarah Forrester. mRNA was separated from 10ng - 1µg of total RNA using the NEBNext Poly(A) mRNA magnetic isolation module, and then converted into cDNA using the NEBNext Ultra RNA library prep for Illumina kit, both per manufacturer's instruction. For step 1.9, section 1.9A was followed in accordance with the choice of multiplex oligo kits. The samples were multiplexed using index primers from the NEBNext Multiplex Oligos for Illumina (Index Primers Set 1) per manufacturer's instruction (see Table 4 below for exact index use).

2.5 PIPPIN PREP AND SEQUENCING

cDNA libraries were selected for quality by concentration and BioAnalyser trace. Libraries with low concentration or libraries that deviated significantly from the manufacturer's recommended trace were discarded.

Libraries that were of an acceptable quality were further optimized by genomics technician Dr Sally James (University of York) by performing a pippin prep on a BluePippin, using an M1 marker, 2% agarose gel, selecting for 200-600bp, through use of a BEF2010 kit (per manufacturer's instruction).

After the pippin prep the libraries were sent to the University of Leeds NGS facility to be sequenced on an Illumina platform HiSeq 3000, with a read length of 125bp, an insert size of 300bp, non-directional paired-end reads. Table 4 lists the multiplex index and flow cell lane of each sample. Samples were sequenced in a single run, split across two flow cell lanes, limiting potential batch effects. A PhiX spike was included but not analysed as the RNA samples were balanced in terms of GC content and diversity.

Table 4. Lane and multiplex index planning for Illumina sequencing of each library.

Flow cell lane	Sample	Index
1	WT INF – 01	1
1	WT INF – 02	2
1	WT INF – 03	3
1	RAG INF – 12	4
1	RAG INF – 13	5
1	WT CONTROL – 06	6
1	WT CONTROL – 07	7
1	WT CONTROL – 08	8
1	INOCULUM – 03	12
2	RAG CONTROL – 16	1
2	RAG CONTROL – 17	2
2	RAG INF - 11	3
2	RAG CONTROL – 18	4
2	WT INF – 05	5
2	RAG INF – 14	6
2	INOCULUM – 01	7
2	INOCULUM – 02	8

2.6 DATA PRE-PROCESSING

The pre-processing and analysis of the sequenced libraries was performed using a combination of Python and R packages, Linux-based freeware and specific programs for the manipulation of genomic data. Further computational analysis based on transcriptomic data was performed using various web services, reliant upon principles of enrichment analysis, which look at overall patterns of gene expression across a dataset, rather than looking at individual genes, gene families or biochemical pathways.

2.6.1 CUTADAPT

The Python package CutAdapt (https://cutadapt.readthedocs.io/en/stable/, version 1.12) was used to remove residual sequencing adaptors from reads (Martin 2011). The adaptors are short nucleotide sequences used for guiding the sequences of interest onto the sequencing platform through Watson-Crick base pairing. Adaptors are sequenced along with the sequence of interest and must be removed computationally, as they cannot efficiently be removed chemically. CutAdapt does not actually remove reads from the library, such that the number of input and output reads will always be the same. Even reads that were entirely trimmed will still be listed, as having a read length of 0 (Martin 2011).

```
~/.local/bin/cutadapt \
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
-A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT \
-o output_R1_trim.fastq.gz \
-p output_R2_trim.fastq.gz \
input_R1.fastq.gz \
input_R2.fastq.gz \
```

-a specifies the forward adaptor sequence to be trimmed

-A specifies the reverse adaptor sequence to be trimmed

-o specifies the forward read output file name and directory

-p specifies the reverse read output file name and directory, after which the forward and reverse input files are listed.

2.6.2 SICKLE

After CutAdapt, the files were processed further with Sickle (https://github.com/najoshi/sickle, version 1.33), to remove low quality reads, with a PHRED score of <20 (Joshi and Fass 2011). Additionally, Sickle can remove any unpaired single reads, and pairs of reads where one read of the two passes quality and length criteria, but not the other. Low quality reads contain bases that have not reliably been called, i.e. it is likely that the correct base in the sequence has been misidentified, and such reads must be removed computationally. While CutAdapt can also be used for this process, Sickle will output singletons and pairs into separate files, and is commonly used as a follow-up program after adaptor trimming.

```
./sickle pe \
-f input_R1_trim.fastq.gz \
-r input_R2_trim.fastq.gz \
-t sanger \
-o output_R1_trim_sick.fastq \
-p output_R2_trim_sick.fastq \
-s output_R0_trim_sick.fastq \
-q 20
-l 20
```

-f specifies the forward read input file

-r specifies the reverse read input file

-t specifies the quality scoring method used to generate the fastqs. Illumina reads use the same method of scoring base quality – before processing.

-o specifies the forward read output file

-p specifies the reverse read output file

-s specifies the output of unpaired/ singleton reads

-q is the threshold for PHRED score, below which a base is removed

-l is the minimum length the reads are trimmed to before being discarded

2.6.3 ALIGNMENT OF SEQUENCING DATA

2.6.3.1 INDEX GENERATION

The STAR (Spliced Transcripts Alignment to a Reference) RNA-Seq alignment program was used to map the processed reads to a reference genome (https://github.com/alexdobin/STAR, Dobin et al. 2013). The STAR aligner is fast, easily parallelised, and splice aware, and thus useful for analysis of transcriptomic data. Reads were aligned to genomes rather than to transcriptomes as the focus of the experiment was gene expression rather than splice variants. In order to map reads to a genome using the STAR aligner, a genome index must first be generated from reference FASTA files, which contain sequence data, and annotation GTF files, which contain information about gene structure. To produce an index containing both mouse and *Leishmania* data at the same time, the files were simply concatenated. An index was generated from the *Mus musculus* GRCm38 reference genome (version 84) and the *Leishmania donovani* BPK282a1 reference genome (version 28) using the following STAR (2.5.2b) command.

```
STAR \
--runThreadN 16 \
--runMode genomeGenerate \
--genomeDir /input_genome/directory/ \
--genomeFastaFiles combined_reference_genome.fasta \
--sjdbGTFfile combined_annotations.gtf
--sjdbOverhang 123
--genomeSAindexNbases 13
--genomeChrBinNbits 13
```

--runThreadN specifies the number of threads on which to run the program

--runMode specifies what mode to run the program in – in this case, index generation

--genomeDir specifies the path to the genome directory

--genomeFastaFiles specifies the FASTA file of the reference genome

--sjdbGTFfile specifies the GTF (annotation) file of the reference genome

--sjdbOverhang specifies the length of genomic sequence using in constructing the splice junction database. For most purposes, it is equal to the read length – 1.

--genomeSAindexNbases specifies a value for scaling the index according to genome size

--genomeChrBinNbits specifies a value for reducing RAM consumption when large reference genomes are used

2.6.3.2 READ ALIGNMENT

After generating a genome index, STAR can be run in a different mode in order to align the trimmed reads of each sample to a reference genome, in this case, the concatenated reference listed above containing both the mouse and *Leishmania donovani* reference genomes (Dobin et al. 2013). The singleton reads separated by Sickle were not included in the alignment.

```
STAR \
--runThreadN 32 \
--genomeDir …/Mouse_donovani_combined/ \
--readFilesIn input_R1_trim_sick.fastq input_R2_trim_sick.fastq \
--outSAMtype BAM Unsorted \
--outFileNamePrefix input_alignment \
--outFilterScoreMinOverLread 0 \
--outFilterMatchNminOverLread 0 \
--outFilterMatchNmin 40 \
```

--readFilesIn specifies the forward and reverse input FASTQ files

--outSAMtype specifies the SAMtools file type to be outputted

--outFileNamePrefix specifies the prefix for output files

--outFilterScoreMinOverLread specifies the output to be filtered so that only alignments with the specifies score or above are returned, normalised to read length

--outFilterMatchNminOverLread specifies the output to be filtered so that only reads with matching the specified score or above are returned, normalised to read length

--outFilterMatchNmin specifies the output to be filtered so that only reads with matching the specified score or above are returned

Setting outFilterMatchNmin to 40 means that the read can be locally aligned, despite repetitive sequence, such as the splice leader (SL) sequence, at one end of the read. The read is not rejected as long as at least 40 base pairs of the read matches the reference, even if the presence of the SL sequence prevents the entire read aligning. Reads where only one of the pair map are discarded with unmapped reads.

2.6.4 READ COUNTING

After alignment, Python package HTSeq-count was used to count reads aligning to each gene on the reference (https://htseq.readthedocs.io/en/release_0.9.1/index.html, version 0.5.4p5). HTSeq-count uses a BAM/SAM (Binary or Sequence Alignment Map) file, in this case produced by the STAR alignment process, and a GTF annotation file, to determine the number of reads aligned with each exon (Anders et al. 2015). HTSeq-count was run in union mode, which allows reads to be assigned to an exon even if the full read does not map (such as those with SL sequences).

Additionally, the SAMtools software (http://samtools.sourceforge.net/, version 0.1.18) was used to manipulate, sort and convert different alignment information into formats suitable for other software (Li et al. 2009). In order to only input uniquely mapping reads into HTSeq-count, a bash script and SAMtools were used to filter and sort the alignment produced by STAR, removing non-uniquely mapping reads.

```
for b in sample1.bam sample2.bam sample3.bam

do

### Need to annotate if you put in like this. Opens the SAM file and extracts the header, and saves it to a sam
file

samtools view -H $b > $b.header.sam

### opens bam file extracts uniquely mapping reads and appends them to a file containing the header and
calls this the uniquely mapping BAM

samtools view $b | awk '{if ($5 == 255) print $0}' | cat $b.header.sam - | samtools view -Sb - >
$b.uniquely_mapping.bam

## command for the htseq counting

htseq-count -m union -f bam $b.uniquely_mapping.bam annotations.gtf > $b.counts

done
```

-m specifies which mode to run HTSeq-count in

-f specifies the format of the input data

> tells the program to save the results in a new file with the suffix .counts

This script was used to loop through sample STAR alignments, and using SAMtools, produce a file containing only the headers (-H), equivalent to gene name. After producing this file, the alignment was opened with SAMtools and filtered using awk; the alignment contains a tag in column 5 ($5) which, when equal to 255, means that the read aligned to only one location. Genes that were tagged as unique were then joined back with the header file using cat, the result of which is a file containing a list of genes and their uniquely-mapping only reads. Reads mapping to multiple locations in the reference genome were discarded from future analyses; only pairs of reads that mapped uniquely were used (i.e if one read of a pair was non-unique, the pair was discarded).

The list of uniquely mapping reads can then be fed into HTSeq-count in order to generate a table of reads counts for each gene.

2.7 DIFFERENTIAL EXPRESSION ANALYSIS

Using the read count table produced by HTSeq-count, edgeR, a Bioconductor R package, (http://bioconductor.org/packages/release/bioc/html/edgeR.html, version 3.18.1) was used to determine which genes were differentially expressed between samples (Robinson et al. 2010). EdgeR was used to fit the data to a negative binomial model, a commonly used model in RNA-Seq for combating statistical bias known as 'overdispersion' seen in the Poisson distribution (McCarthy et al. 2012). The generalised linear model function, glmFit, was used to fit the data, which was then normalised between libraries, within groups, and across the dataset, by calculating dispersion, which estimates the biological coefficient of variation (BCV) between samples. For each pairwise comparison, Log2FoldChange was calculated, and the generalised linear model applied to test and identify differentially expressed genes, with false discovery rate set to 0.05. EdgeR gives comparable results to more prevalent DESeq2, and both rely on the negative binomial model, but EdgeR handles gene expression at extremes (i.e very high or very low) differently (Schurch et al. 2016; Zhang et al. 2014).

An overview of the bioinformatic analysis techniques and the data produced from each can be seen in figure 6 below.

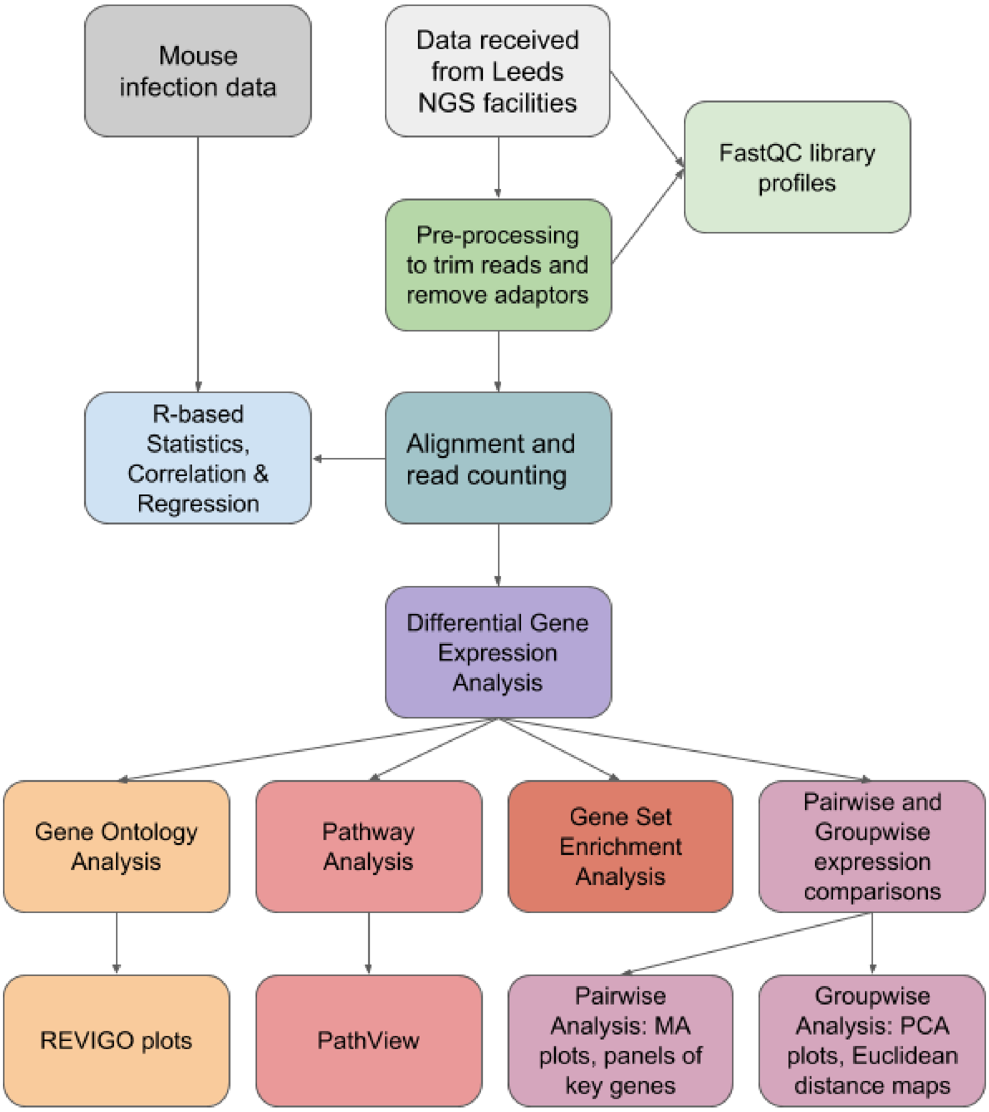

Mouse infection data
Data received from Leeds NGS facilities
FastQC library profiles
Pre-processing to trim reads and remove adaptors
R-based Statistics, Correlation & Regression
Alignment and read counting
Differential Gene Expression Analysis
Gene Ontology Analysis
Pathway Analysis
Gene Set Enrichment Analysis
Pairwise and Groupwise expression comparisons
REVIGO plots
PathView
Pairwise Analysis: MA plots, panels of key genes
Groupwise Analysis: PCA plots, Euclidean distance maps

Figure 6. Experimental overview, detailing the different analyses performed on the same transcriptome data, and the origin of each type of plot generated.

2.7.1 DATA IMPORT AND TRANSFORMATION

Three R libraries were necessary to calculate differentially expressed genes, sort files, perform statistics, and plot data.

```
library(edgeR)
library(limma)
library(fields)
```

Data were transformed into a matrix for pairwise comparison, using 'sgroup' as a factor. sgroup is a vector containing columns imported from the user-specified design, providing group names, i.e infected, uninfected, RAG, WT. as.formula(~-1+sgroup) applies pairwise comparisons, creating a matrix from a data frame, which is necessary for further calculations.

```
modelformula = as.formula(~-1+sgroup)
dmat = model.matrix(modelformula)
rownames(dmat) <- rownames(designtab)
```

2.7.2 INITIAL CALCULATIONS AND NORMALISATION

Lists of genes and counts were added to the differentially expressed gene list (before calculations) using edgeR function DGEList, and saved as object 'dgl'. Genes with zero counts (across all samples) were removed; if any sample contained one read for the gene in question then the gene was retained for analysis. Counts per million, the read count of the gene divided by the size of the library (in millions), and the mean counts per million for each group, were also calculated. CPM is used to compare expression between libraries of different sizes.

```
dgl <- DGEList(counts <- as.matrix(counttab),group=sgroup, genes=annotab, remove.zeros=T)
cpmval <- cpm(dgl)
xcpm <- log2(rowMeans(cpmval))
```

Normalisation is then applied to individual libraries, within groups, and across the dataset by dispersion calculations, again using edgeR functions, which allows the library data to be compared.

```
dgl <- calcNormFactors(dgl)
dgl <- estimateGLMCommonDisp(dgl,dmat)
dgl <- estimateGLMTrendedDisp(dgl,dmat,min.n=500)
dgl <- estimateGLMTagwiseDisp(dgl,dmat)
```

Dispersion was plotted.

```
plot(xcpm,dgl$tagwise.dispersion,cex=0.4)
points(xcpm,dgl$trended.dispersion,cex=0.4,col="green")
abline(h=dgl$common.dispersion[1],col="cyan")
```

2.7.3 CALCULATION OF DIFFERENTIAL EXPRESSION

In order to find evidence of differential expression, the data were fitted to the model, in this case, the 'fitres' function checking if object 'dgl' (the list of genes) matches the model 'glm' (generalised linear model). This process normalises data between groups for comparison. Setting the contrast to 'cmat' ensures pairwise comparisons; glmLRT performs pairwise comparisons of read counts for each gene and each sample.

```
fitres <- glmFit(dgl,dmat)
```

```
glmLRT(fitres,contrast=cmat[,j])
```

The FDR was set to 0.05 and applied as a filter to the results of the model fitting.

```
fdrcut <- 0.05

de.yes.no <- as.matrix(FDR < fdrcut)
```

All genes in the edgeR generated list had a p-value and FDR smaller than or equal to 0.05; these genes were used to produce the MA and PCA plots described below. After the differentially expressed gene list (DEG list) was generated, further criteria were manually applied to reduce less reliable and less relevant results, summarising a larger dataset into a smaller and more manageable one. Results were considered less reliable if their read counts were extremely low; small fold changes were considered less biologically relevant. In addition to the p-value and FDR cut offs, criteria were also applied to the log CPM (counts per million) and to log FC (fold change) for several further analyses. To filter out genes with very low read counts, only genes with a logCPM of 1 or greater were included. In order to exclude genes with small changes in expression level, only genes with a logFC above or equal to 1.5, or below or equal to -1.5 were used. Analyses reliant on this reduced set of data are designated as ‘filtered’.

2.7.4 PLOTTING DATA

MA plots were generated from unfiltered data for each pairwise comparison. MA plots, named after their axes (M being the expression magnitude, and A being the mean average counts per gene, or coverage) aim to compare the distribution of genomic data (i.e the overall expression of genes) between two samples. MA plots visualise bias and trends in gene expression, and indicate whether data requires further normalisation, as well as displaying which genes are significantly differentially expressed, and at what fold-change.

```
fc <- logFC

fc[(fc)>10] <- 10

fc[ fc < -10] <- -10

x11()

par(mfcol=mfcol)

for(k in 1:npanel) {

   ylab <- colnames(logFC)[k]

   i <- which(de.yes.no[,k])

   maPlot(x=NULL,y=NULL,logAbundance= xcpm, logFC = fc[,k], xlab = "log2CPM", ylab = ylab,

      de.tags= i,pch = 19, cex = 0.3, smearWidth = 0.5, panel.first = grid(),

   smooth.scatter = FALSE, lowess = FALSE,na.rm =TRUE)

   }
```

PCA (principal component analysis) plots were generated from unfiltered data to compare the relative similarity of samples. PCA transformations attempt to reduce the variance of data in multiple dimensions into as few dimensions as possible. In the case of RNA-Seq data, PCA plots visualise the overall statistical differences (such as batch effects) between the p-values and logFC of samples and groups of samples.

```
pcaplot <-
function(count,sgroup,snumber=NULL,filename="PCAplot",figtype="png",pch=16,cex=0.8,display=FALSE,col=
NULL)

{

if(is.null(snumber)) {snumber=gsub("Sample_","",colnames(count)); snumber=gsub("sample_","",snumber)}

print(snumber)

  labels=paste(sgroup,snumber,sep="_")

wx <- count

  wx[wx <= 0 ] <- 1

    pc <- princomp(wx)

  ld <- pc$loadings

  if(is.null(col)) {col <- rep(0,length(sgroup))

    for(k in 1:length(unique(sgroup))) col[sgroup %in% unique(sgroup)[k]] <- k

  } else col=col

  pch=pch

  cex=cex

   if(!is.null(filename)) {

    if(!is.null(figtype)) {

    par(mar=c(6,4,3,2))

    xlim=c(min(ld[,2]),max(ld[,2]))

    xlim[1]=xlim[1] - (max(ld[,2])- min(ld[,2]))*0.1

    xlim[2]=xlim[2] + (max(ld[,2])- min(ld[,2]))*0.1

    plot(ld[,2],ld[,3],cex=cex,col=col,pch=pch,xlim=xlim,main="PCA plot",xlab="2nd Component",ylab="3rd
component",cex.axis=1.2,cex.lab=1.2)

    text(ld[,2],ld[,3],labels,col=col)

    dev.off()

   }
```

Euclidean distance heatmaps were also generated in order to visualise the relative similarity of each sample, and highlight differences between groups. As with the PCA plot, the heatmaps were also built from p-value and logFC data. Euclidean distance heatmaps are useful in demonstrating overall group characteristics, as well as pairwise comparisons displaying how distinct each individual sample is.

```
corrheatmap <- function(count,sgroup,snumber=NULL,filename=NULL,figtype="png",useorder=NULL) {
```

```
if(is.null(filename)) x11() else {
        mtc=match(figtype,c("png","pdf","jpeg","bmp","tiff"))
        switch(mtc,
            png(file=paste0(filename,".png")),
            pdf(file=paste0(filename,".pdf")),
            jpeg(file=paste0(filename,".jpeg")),
            bmp(file=paste0(filename,".bmp")),
            tiff(file=paste0(filename,".tiff"))
        )}
    if(is.null(snumber)) {snumber=gsub("Sample_","",colnames(count)); snumber=gsub("sample_","",snumber)}; labels=paste(sgroup,snumber,sep="_")
    layout(matrix(c(1,2),1,2),widths=c(7,1))
    par(mar=c(7,7,5,3))
    cnd <- as.character(sgroup)
    if(is.null(useorder)) o <- order(cnd) else o <- useorder
    cnd <- cnd[o]
    ucnd <- unique(cnd)
    corr <- cor(count[,o])
    gsize <- NULL
    for(k in 1:length(ucnd)) gsize <- c(gsize,sum(cnd %in% ucnd[k]))
    cgsize <- cumsum(gsize)
    at0 <- c(0,cgsize[-length(ucnd)])
    at <- (cgsize + at0 +1)/2
    atl<- cgsize+0.5
    print(ucnd)
    grpbar <- NULL
    for(k in 1:length(ucnd)) grpbar <- c(grpbar,rep(ifelse(k %% 2 !=0,NA,1),gsize[k]))
    colbar <- seq(min(corr),max(corr),(max(corr)-min(corr))/255)
    labels <- format(c(min(corr),min(corr)/2+max(corr)/2,max(corr)),digits=3)
    corr <- rbind(corr,grpbar)
    image(1:(nrow(corr)),1:(ncol(corr)),corr,xlab="",ylab="",col=tim.colors(256),xaxt="n",yaxt="n",main="Sample correlation heatmap")
    abline(h=atl,lwd=2)
    abline(v=atl,lwd=2)
    if(1) {
    axis(side=1,at=1:(length(cnd)+1),labels=c(paste(cnd,snumber[o],sep="_"),"group"),las=2,cex.axis=1)
    axis(side=2,at=1:length(cnd),labels=paste(cnd,snumber[o],sep="_"),las=2,cex.axis=1)
    axis(side=4,at=at,labels=ucnd,cex.axis=1,las=2)
    } else {
```

```
    axis(side=1,at=1:(length(cnd)+1),labels=c(snumber[o],"group"),las=2,cex.axis=1.2)
    axis(side=2,at=1:length(cnd),labels=snumber[o],las=2,cex.axis=1.2)
    axis(side=4,at=at,labels=ucnd,cex.axis=1.2,las=3)
    }
    par(mar=c(14,1,10,6)/2)
    image(x=1,y=1:256,z=matrix(colbar,nrow=1),col=tim.colors(256),xaxt="n",yaxt="n",xlab="",ylab="")
    axis(side=4,at=c(1,128,255),labels=labels,cex.axis=1.2)
    if (!is.null(filename)) dev.off()
}
```

2.8 STATISTICS AND PLOTS IN R

2.8.1 SIMPLE REGRESSION AND T-TEST ANALYSIS

R Studio version 3.3.3 was used to perform basic regression analyses and Student’s T-tests on some of the experiment metrics and statistics reported by STAR, HTSeq-count etc. looking for relationships between different measures such as parasite burden and percentage of reads mapping to the *Leishmania* genome. Simple R scripts were used, such as the following, to perform regressions and T-tests:

```
library(ggplot2)

leish_regression = data.frame(
  LDU = c(12, 14, 21, 29, 34, 37, 37, 42),
  percent_L = c(0.24, 1.47, 2.55, 0.15, 0.12, 3.03, 1.6, 0.21)
)

ggplot(leish_regression, aes(x=LDU, y=percent_L)) +
  geom_point(shape=1) +
  geom_smooth(method=lm, se=FALSE)

fit <- lm(percent_L ~ LDU, data = leish_regression)
summary(fit)

host_LDU = data.frame(
  host = c(1, 1, 1, 1, 2, 2, 2, 2),
  LDU = c(34, 29, 42, 12, 14, 37, 21, 37)
)

host_percent_L = data.frame(
  host = c(1, 1, 1, 1, 2, 2, 2, 2),
  percent_L = c(0.12, 0.15, 0.21, 0.24, 1.47, 1.6, 2.55, 3.03)
)

t.test(LDU ~ host, host_LDU)
t.test(percent_L ~ host, host_percent_L)
```

2.8.2 GENE PANELS

In addition to basic regression analysis, R Studio was also used to generate heatmaps of filtered fold-change data drawn from significant ($p \leq 0.05$) mouse edgeR results, in order to compare expression of specific genes between samples. Three ‘panels’ of genes were selected, covering expression data from key cytokines, chemokines, and CD markers. Genes showing significant fold change in samples were included in the heatmap, which was generated according to the following script:

```
install.packages("gplots")
library(gplots)
install.packages("RColorBrewer")
library(RColorBrewer)
library(readr)

heatmap_test <- read.csv("cytokine_panel.csv")
map <- heatmap_test
attach(map)

maplabs<-map[,1]

map<-map[,2:4]

map_mat<-data.matrix(map)


my_palette <- colorRampPalette(c("blue", "white", "red"))(n=100)


pdf("plot.pdf")

heatmap.2(map_mat, Rowv=FALSE, na.color="grey", col=my_palette, labRow = maplabs, density.info =
"none", trace=c("none"), dendrogram = c("none"), scale = c("column"), cexCol = 0.7, cexRow = 0.6)

dev.off()
```

maplabs contains the names of the genes involved in the panel, separated from the values. map, the fold change values, was transformed into data matrix map_mat in order to be plotted into a heatmap using the heatmap2 function of the gplots package. Samples with no significant fold change for the gene in question were designated grey using NaN (not a number). Instead of displaying logFC values, the heatmap.2 function instead calculates Z-score, which scales and smooths the values of each row such that the changes are comparable.

2.9 GENE SET ENRICHMENT ANALYSIS

Gene set enrichment analysis (GSEA) was performed using the Broad Institute website (http://software.broadinstitute.org/gsea/msigdb/annotate.jsp), website version 6.1 and MSigDB version 6.1, last accessed 28/06/17 (Subramanian et al. 2005). GSEA is used to compare patterns of gene expression changes, for example, changes in expression typically associated with infection or cancer. One criticism of GSEA is that the datasets have a high degree of redundancy. For example, when calculating overlap between the different Hallmark sets, on average, 21% of genes overlap with another set; one set had over 75% of its genes overlap with another Hallmark set (see accompanying material A2 for full results). Although the curated datasets have a degree of overlap with each other, they are still useful for recognising similarities in expression changes.

Filtered lists of mouse genes were checked against Motif (C3), Immunological (C7), KEGG Pathway (CP:KEGG) and Hallmark (H) datasets for overlap. No gene sets were available for use with *Leishmania* data.

2.10 GENE ONTOLOGY ANALYSIS

Gene ontology (GO) analysis was performed, using the GOrilla website for filtered mouse data, last accessed 05/06/17 (http://cbl-gorilla.cs.technion.ac.il/, Eden et al. 2009), and for *Leishmania* genes, website TriTrypDB's (version 32) gene ontology function was used, also using filtered data, last accessed 05/06/17 (http://tritrypdb.org/tritrypdb/, Aslett et al. 2010).

GO analysis examines lists of differentially expressed genes, to determine if any particular category of gene, such as immune signalling genes or brain development genes, are expressed at a higher (over-enrichment) or lower (under-enrichment) level than expected. Hypergeometric statistical tests are performed to determine the likelihood that the enrichment is a coincidence, by comparing the gene set of interest against a standard background set (Eden et al. 2009). From this data it is possible to see what pathways, organs/tissues, and systems are involved in an organism's reaction to change.

To visualise the GO analysis results, the website REVIGO (http://revigo.irb.hr/) was used with default settings, using the GO database data from Jan 2017, UniProt DB data from March 2017, last accessed 27/06/17 (Supek et al. 2011).

2.11 PATHWAY VISUALISATION

In order to visualise fold changes in gene expression across multiple genes in a biochemical pathway, the Pathview website (https://pathview.uncc.edu/) was used (software version 1.14.0), last accessed 05/07/17 (Luo and Brouwer 2013). Insufficient KEGG data is available for use of Pathview with *Leishmania*, but mouse KEGG data is much more substantial, allowing for visualisation.

Initially pathway selection was set to 'auto', allowing the website to choose the most appropriate pathways for visualising the filtered input data, after which pathways of particular interest were selected manually, such as those suggested by gene ontology analysis, or immunity-related pathways. Default graphics and colour settings were used.

CHAPTER 3: RESULTS

3.1 EXPERIMENTAL DESIGN

In order to generate mouse and parasite RNA samples for comparing transcriptomes under different conditions, mouse strains of varied genetic background were chosen for infection with *L. donovani* (Figure 7). Black 6 mice were chosen for their aggressive immune response when faced with *Leishmania* infection; RAG2 KO variants were chosen as contrast, as a model of low immune pressure. Using transcriptome samples from these two genetic backgrounds and infection statuses, it was possible to determine differences in the baseline transcriptomes of the mice, identify genes involved in a healthy immune response, and determine the way RAG2 KO mice respond to infection.

Dual RNA-Seq approaches also allowed for the isolation of parasite transcriptome data from mixed mouse/parasite RNA samples. Additional libraries were prepared from an inoculum, to compare *Leishmania* transcriptomes over time during infection.

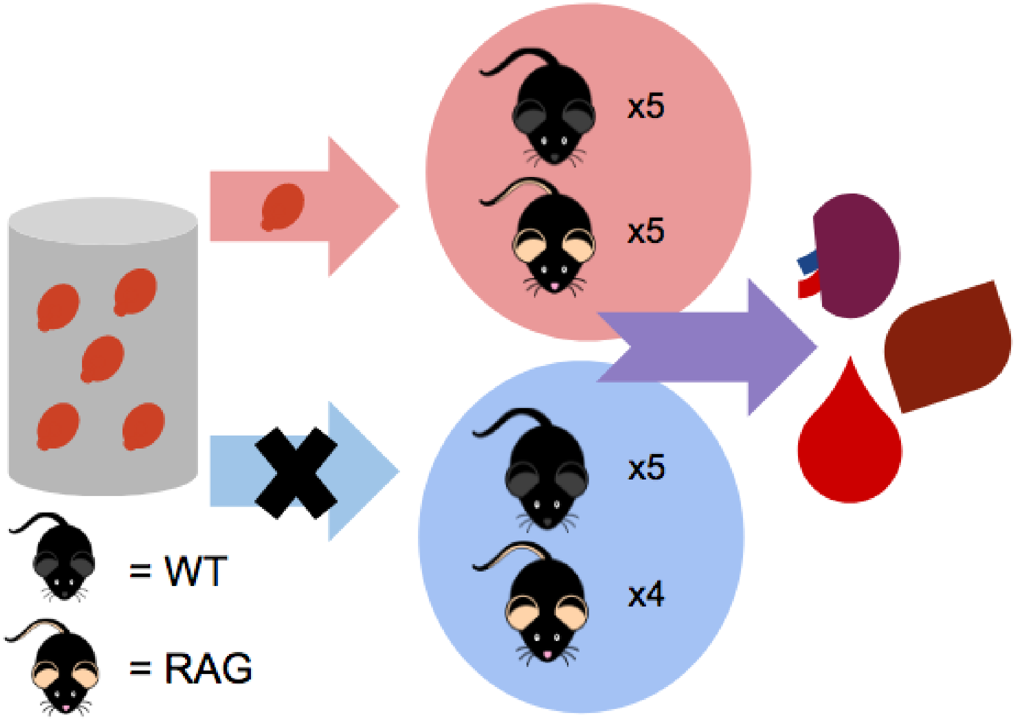


Figure 7. Experimental design overview. The *L. donovani* amastigote inoculum (grey) was used to infect 10 mice, 5 of each genetic background (pink). 9 mice (5 WT and 4 RAG) remained uninfected to serve as controls (blue). After 28 days, the spleen, liver and blood were harvested from euthanised mice. A minimum of 4 mice were used for each condition to serve as biological replicates. 3 samples were generated from the same inoculum RNA sample to serve as technical replicates.

Due to time constraints, cDNA libraries for Illumina sequencing were only generated from spleen and inoculum RNA samples. A minimum of three libraries were generated for each condition for statistical robustness (table 5).

Table 5. A summary of the RNA and cDNA samples prepared from the experiment.

CONDITION	NO. OF RNA SAMPLES GENERATED	NO. OF cDNA LIBRARIES GENERATED
B6 WT uninfected	5	3
B6 WT infected	5	4
B6 RAG2KO uninfected	4	3
B6 RAG2KO infected	5	4
Leishmania inoculum	3 (technical replicates)	3 (technical replicates)

3.2 PRE-CLEANING RNA QUALITY

3.2.1 SPLEEN RNA SAMPLES

RNA sample quality was checked after DNAse treatment, but before magnetic bead cleaning, in order to check for sample degradation that may have occurred during storage, and contamination with salts and other chemicals left over from the extraction and DNAse treatments. RNA Pico BioAnalyser chips were used for the spleen samples. Spleen total RNA samples show a mean RNA integrity number (RIN) score of 8.28, ranging from 6.00 – 9.40, with a mean score of 8.28 (table 6). All samples show a score above 6.00, the standard quality threshold. Sample concentration ranged from 2.156ng/µl to 7.178ng/µl, with an average of 4.874ng/µl. Concentration was measured to ensure that enough RNA was present in the sample to continue to the cleaning stage. 18S and 28S rRNA ratios are used as an indicator of degradation. For the spleen RNA samples, the rRNA ratio varied from 0.8 to 1.1, with a mean of 1.0. On other platforms, a standard ratio of 2.0 is typically used as a sign of good quality. However, on the BioAnalyser, it has been noted that even good quality RNA can struggle to meet the 2.0 standard (Imbeaud et al. 2010). Additionally, the mouse rRNA RIN score is affected by the presence of *Leishmania* rRNA, which instead comes in three peaks rather than two. The *Leishmania* peaks are obscured by the relative abundance of the mouse RNA.

Figures 8a and 8b show the electrophoresis gels produced by the BioAnalyser. The two bands correspond to mouse 18S and 28S rRNAs, which can be seen as the peaks in figure 8c. *Leishmania* rRNA peaks cannot be seen on the gel as the mouse rRNAs are massively more abundant and obscure the presence of the *Leishmania* peaks. Spleen samples 10 and 18 show mild degradation, indicated by grey banding, which was likely caused by over-vortexing.

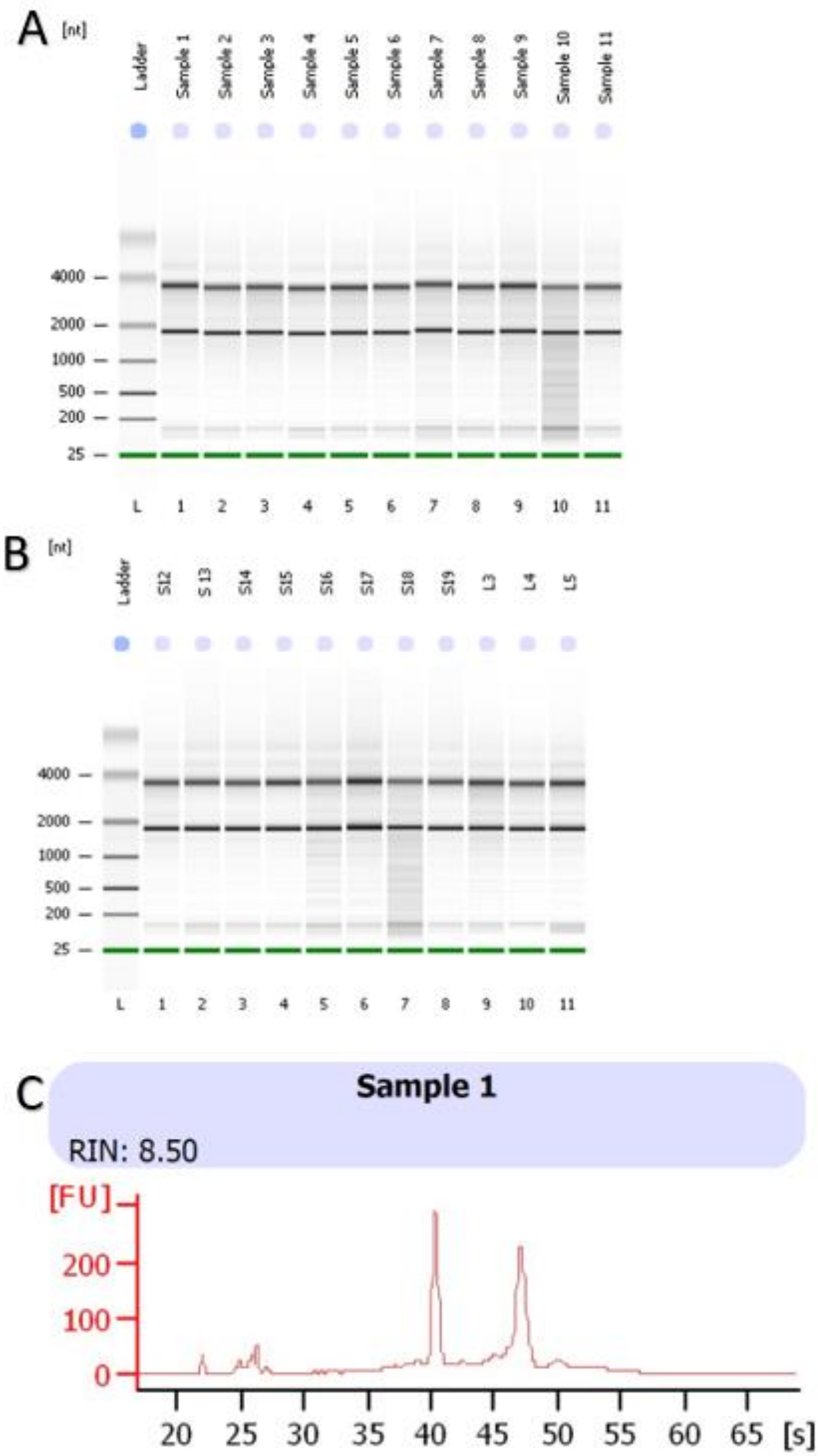


Figure 8a, b & c – Assessment of pre-cleaning RNA quality. Agilent RNA Pico BioAnalyser Chip Results, with vertical axis showing fragment size in nucleotides. 8a – spleen samples 1-11. 8b – spleen samples 12-19. 8c – sample 1 trace.

Table 6. Assessment of pre-cleaning RNA. Agilent RNA Pico BioAnalyser chip results of spleen RNA samples.

Sample background and no.	RIN score	Concentration (ng/µl)	rRNA ratio (28s/18s)
WT infected – 1	8.50	5.970	1.1
WT infected – 2	8.40	5.623	1.0
WT infected – 3	8.20	5.176	1.1
WT infected – 4	8.50	5.134	1.1
WT infected – 5	8.40	5.413	1.1
WT uninfected – 6	8.50	4.711	1.1
WT uninfected – 7	7.80	5.181	1.0
WT uninfected – 8	8.30	5.065	1.1
WT uninfected – 9	7.70	4.781	1.1
WT uninfected – 10	6.00	5.045	0.9
RAG infected – 11	8.30	4.290	1.0
RAG infected – 12	8.70	5.262	1.0
RAG infected – 13	9.20	6.007	0.9
RAG infected – 14	9.10	7.178	1.0
RAG infected – 15	9.40	4.198	1.0
RAG uninfected – 16	7.80	4.688	0.8
RAG uninfected – 17	8.70	4.355	1.0
RAG uninfected – 18	6.80	2.373	0.9
RAG uninfected – 19	9.10	2.156	0.9

3.2.2 INOCULUM RNA SAMPLES

Parasite total RNA, isolated from the inoculum, also underwent pre-cleaning assessment as storage issues caused the degradation of a number of samples. Two samples underwent assessment via BioAnalyser RNA pico chip, and one was chosen for further use, and split into three samples to act as technical replicates.

The parasite RNA samples show a lower average RIN score – 6.15 – but still pass the 6.00 quality standard (table 7). The rRNA ratio for the parasite RNA is not useful as a measure of quality as parasite rRNA does not show only two peaks (28S and 18S) as mammalian RNA does, reflecting the divergence between mammalian and kinetoplastid rRNA. Instead, the parasite RNA shows 3 peaks (figures 9a & 9b) representative of the smaller rRNAs used to assemble the ribosome (Zhang et al. 2016).

Given the slightly higher RIN score, and higher concentration, the first RNA sample was selected to be used as technical replicates.

Table 7 – Assessment of pre-cleaning parasite RNA. Agilent RNA Pico Chip BioAnalyser results of the parasite RNA.

Sample no.	RIN score	Concentration (ng/µl)	rRNA ratio (28s/18s)
1	6.20	6.027	0.0
2	6.10	4.474	0.0

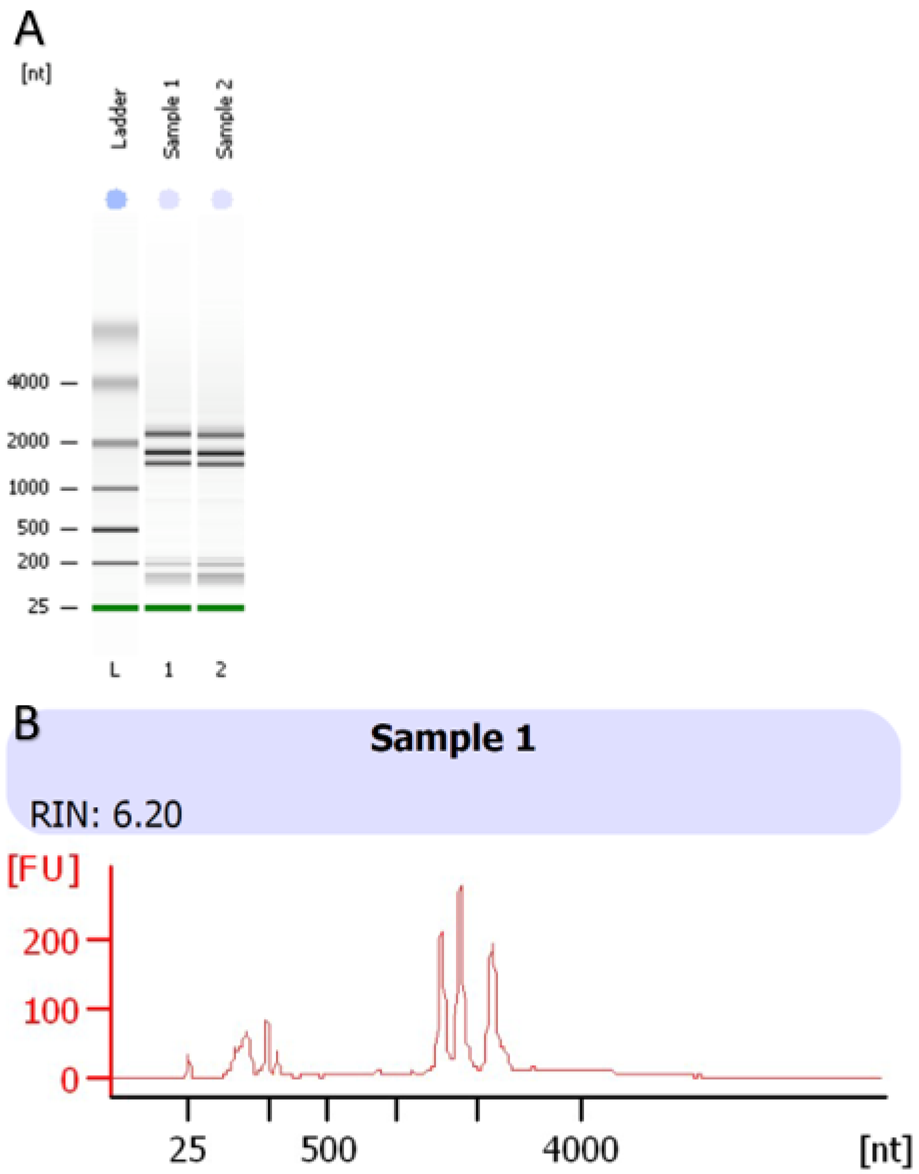


Figure 9a & 9b – Assessment of pre-cleaning parasite RNA. 9a - Agilent RNA Pico Chip BioAnalyser results of parasite RNA, showing the three bands corresponding to *Leishmania* rRNA peaks. 9b – inoculum sample 1 trace showing three rRNA peaks.

3.3 POST-CLEANING RNA QUALITY

3.3.1 SPLEEN RNA SAMPLES

RNA quality was checked again after magnetic bead cleaning, to remove small fragments of RNA that may have been produced by degradation. Spleen RNA samples show a range of RIN scores from 6.30-9.70, with an average of 8.73, showing that RIN score has been slightly increased by the magnetic cleaning procedure (table 8). Sample concentration has reduced, given that degraded RNA is being removed from the sample. The average concentration after cleaning was 3.215ng/µl, with a range of 1.934 to 4.420ng/µl. rRNA ratios improved slightly with cleaning, bringing the average up to 1.1, and ranging from 0.8 to 1.3.

Samples 10 and 18 still show minor degradation, but sample 10 has improved in RIN score from 6.00 to 6.30 (figure 10). Sample 18 appears to have a slightly lower RIN score, but is still above the quality threshold of 6.00.

Table 8. Assessment of post-cleaning RNA. Agilent RNA Pico BioAnalyser results for post-cleaning spleen RNA samples.

Sample background and no.	RIN score	Concentration (ng/µl)	rRNA ratio (28s/18s)
WT infected – 1	9.10	3.335	1.1
WT infected – 2	9.50	2.468	1.0
WT infected – 3	9.40	1.934	1.3
WT infected – 4	9.60	2.193	1.1
WT infected – 5	9.70	2.759	1.2
WT uninfected – 6	9.40	2.533	1.3
WT uninfected – 7	9.20	2.346	1.2
WT uninfected – 8	9.00	3.985	1.0
WT uninfected – 9	8.00	3.587	1.2
WT uninfected – 10	6.30	3.536	0.9
RAG infected – 11	8.90	3.187	1.0
RAG infected – 12	8.80	4.420	1.2
RAG infected – 13	9.30	3.508	1.1
RAG infected – 14	9.20	3.503	1.0
RAG infected – 15	9.30	3.197	1.0
RAG uninfected – 16	7.60	3.569	0.8
RAG uninfected – 17	8.10	3.907	0.9
RAG uninfected – 18	6.60	3.884	0.9
RAG uninfected – 19	8.90	3.246	0.9

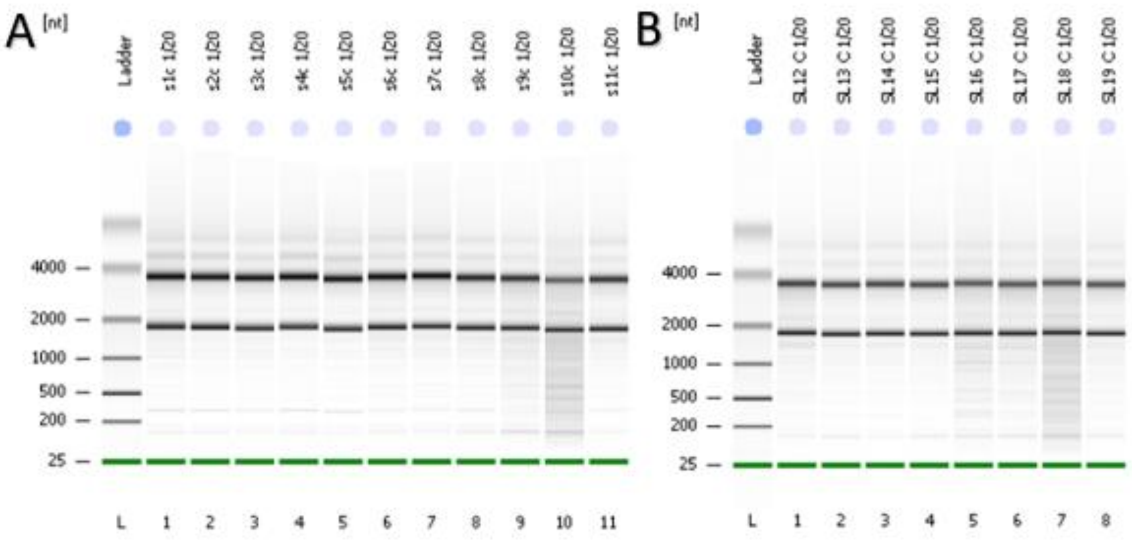


Figure 10. Assessment of post-cleaning RNA. Agilent RNA Pico BioAnalyser results, vertical axis showing fragment size in nucleotides. 10a – spleen samples 1-11. 10b – spleen samples 12-19.

After being run through the BioAnalyser, samples were additionally prepared for more precise concentration quantification using the Qubit (table 9). These samples were not diluted 1/20 for testing, as they were for the BioAnalyser, which is why there is a drastic difference in concentration. Qubit quantification allowed for appropriate dilution of the sample for cDNA library preparation.

Table 9. Assessment of post-cleaning RNA. Qubit results for post-cleaning spleen RNA samples.

Sample background and no.	In Original (ng/µl)	In Qubit tube (pg/ml)	Approximate 1/20 (ng/µl)
WT infected – 1	75.600	378000	3.780
WT infected – 2	36.000	180000	1.800
WT infected – 3	43.400	217000	2.170
WT infected – 4	45.200	226000	2.260
WT infected – 5	38.000	190000	1.900
WT uninfected – 6	29.000	145000	1.450
WT uninfected – 7	36.400	182000	1.820
WT uninfected – 8	29.000	145000	1.450
WT uninfected – 9	36.400	182000	1.820
WT uninfected – 10	33.200	166000	1.660
RAG infected – 11	27.800	139000	1.390
RAG infected – 12	25.200	126000	1.260
RAG infected – 13	34.000	170000	1.700
RAG infected – 14	34.800	174000	1.740
RAG infected – 15	34.800	174000	1.740
RAG uninfected – 16	31.400	157000	1.570
RAG uninfected – 17	15.700	78500	0.785
RAG uninfected – 18	32.400	162000	1.620
RAG uninfected – 19	79.800	399000	3.990

3.3.2 INOCULUM RNA SAMPLES

Parasite RNA was not run on a BioAnalyser chip after magnetic bead cleaning; instead, uncleaned RNA was cleaned, measured with the NanoDrop and Qubit, before being immediately used in cDNA library prep (tables 10 and 11). 260:280 and 260:230 ratios of 2 and 2.0 – 2.2 are used as indicators of sample purity, with significant deviations indicative of contamination. The nanodrop absorbance ratio results for the parasite RNA sample do not indicate contamination.

Table 10. Cleaned purified parasite RNA nanodrop results.

NanoDrop			
Sample no.	ng/µl	260:280	260:230
1	68.200	2.20	2.38

Table 11. Cleaned purified parasite RNA qubit results.

Qubit		
Sample no.	In Original (ng/µl)	In Qubit tube (pg/ml)
1	53.000	265000

3.4 PRE-SEQUENCING LIBRARY QUALITY

cDNA libraries were quality assessed using Agilent BioAnalyser HS DNA chips prior to sequencing. These results are shown in table 12. Average fragment size for the libraries is 381.5bp, with a range of 341 – 464bp. Library concentration showed a high level of variance, with a range from 0.71028 ng/µl to 343.02313 ng/µl. The low concentration is evident from the gel (11d). However, libraries are pooled in equimolar concentrations prior to sequencing.

Spleen sample 12 has two entries in the table, as two libraries were pooled together after the chip was run and submitted as one sample for sequencing. Figure 11c shows one of the sample 12 libraries with a significant gap in fragment size at around 320bp, indicative of a poor quality library. Such a gap is also present in sample 8 (figure 11a) but is not as pronounced. Parasite cDNA libraries (fig. 12) do not show substantial fragment gaps. Figure 13 shows examples of acceptable and poor quality cDNA libraries.

Table 12. Agilent BioAnalyser HS DNA results for cDNA libraries

Sample background and no.	Average size (bp)	Size distribution in CV (%)	Concentration (ng/µl)	Molarity (pmol/l)
Inoculum – 1	397	30.6	20.9296	88474
Inoculum – 2	438	33.0	11.60027	45432.1
Inoculum – 3	464	30.1	10.25117	37548.5
WT infected – 1	362	37.1	11.39127	54536.6
WT infected – 2	369	33.3	14.72119	67389.2
WT infected – 3	366	34.4	28.44802	132971.8
WT infected – 5	373	35.9	10.99224	50807.5
WT uninfected – 6	381	33.7	18.07308	80817
WT uninfected – 7	393	33.1	15.52196	67429.1
WT uninfected – 8	399	30.4	8.64519	36600
RAG infected – 11	341	22.7	2.94052	13873.2
RAG infected – 12	346	35.3	343.02313	1713187.6
RAG infected – 12	356	30.4	16.47886	76777.4
RAG infected – 13	348	31.8	13.34578	64171.8
RAG infected – 14	366	29.3	9.90055	44963.8
RAG uninfected – 16	403	29.6	4.54839	18972
RAG uninfected – 17	372	26.1	0.71028	3151.6
RAG uninfected – 18	402	30.4	6.32234	26638.7

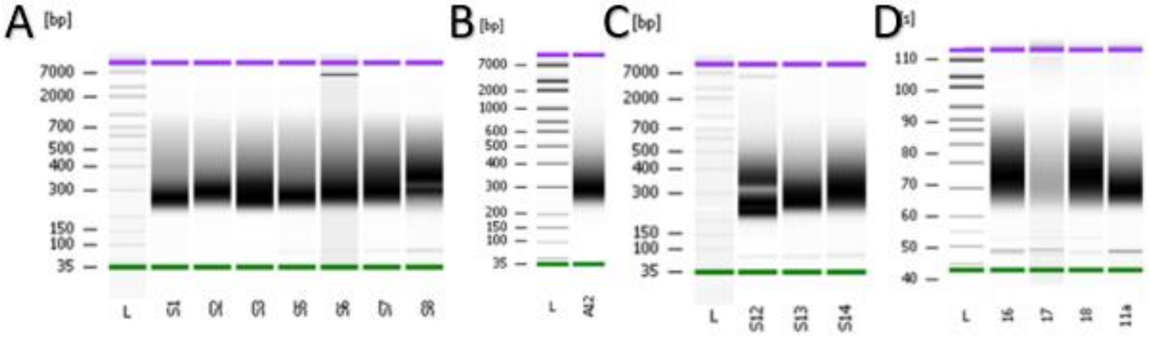


Figure 11. Assessment of cDNA library quality. Agilent BioAnalyser HS DNA results. a, b and c have vertical axes showing fragment size in bp. Due to a technical failure with the chip, d instead shows elution time in seconds, which cannot be converted to bp. 11a – Spleen samples 1-8. 11b – spleen sample 12. 11c – spleen samples 12-14. 11d – spleen samples 11a, 16-18.

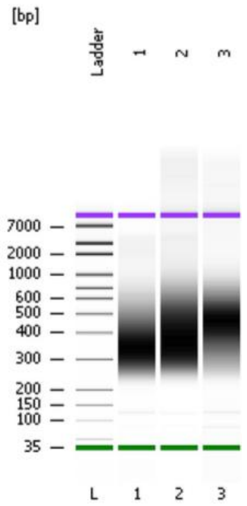
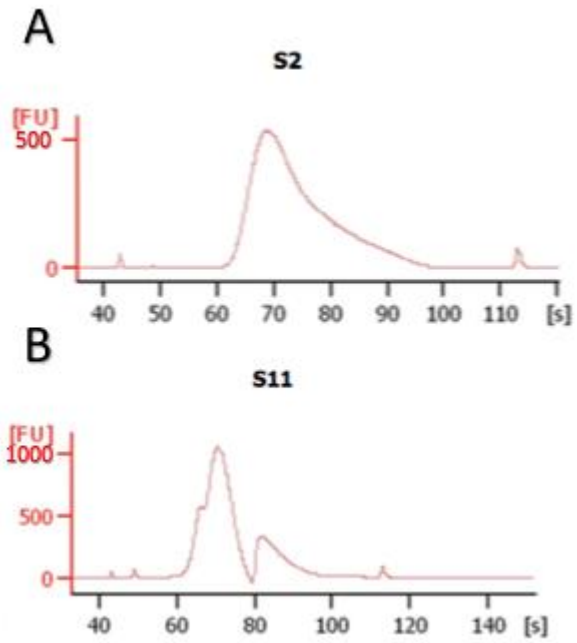


Figure 12. Assessment of pre-sequencing cDNA library quality. Agilent BioAnalyser HS DNA results for parasite inoculum libraries.



500

1000

Figure 13. Examples of good and poor quality cDNA library curves. Agilent BioAnalyser HS DNA curves for spleen cDNA libraries, with vertical axis showing fluorescent units. 13a shows a library with a good quality curve (WT infected 02). 13b shows a library with entire fragment fractions missing, signifying a poor quality library (RAG infected 11 – not sequenced – another cDNA library was generated and used instead).

3.5 POST-SEQUENCING LIBRARY QUALITY

17 cDNA libraries were sequenced by the University of Leeds NGS facility and returned in the form of pairs of FASTQ files, each containing forward (R1) and backward (R2) reads for each library (table 13). The total number of sequences returned ranges from 29,004,444 to 205,331,092, with an average of 88,662,474. Equimolar amounts of cDNA were used to produce the libraries, and as such the difference in the number of reads between each library is likely a result of the library content, and non-normal distribution causing preferential amplification.

FastQC was used to determine metrics, such as sequence length, GC content etc. Between sample variation exists, as aforementioned, the number of reads varies between samples, however the number of forward and backward reads of each library are identical, though some variation exists between the per base quality scores (Figure 14a) as a consequence of the Illumina sequencing process.

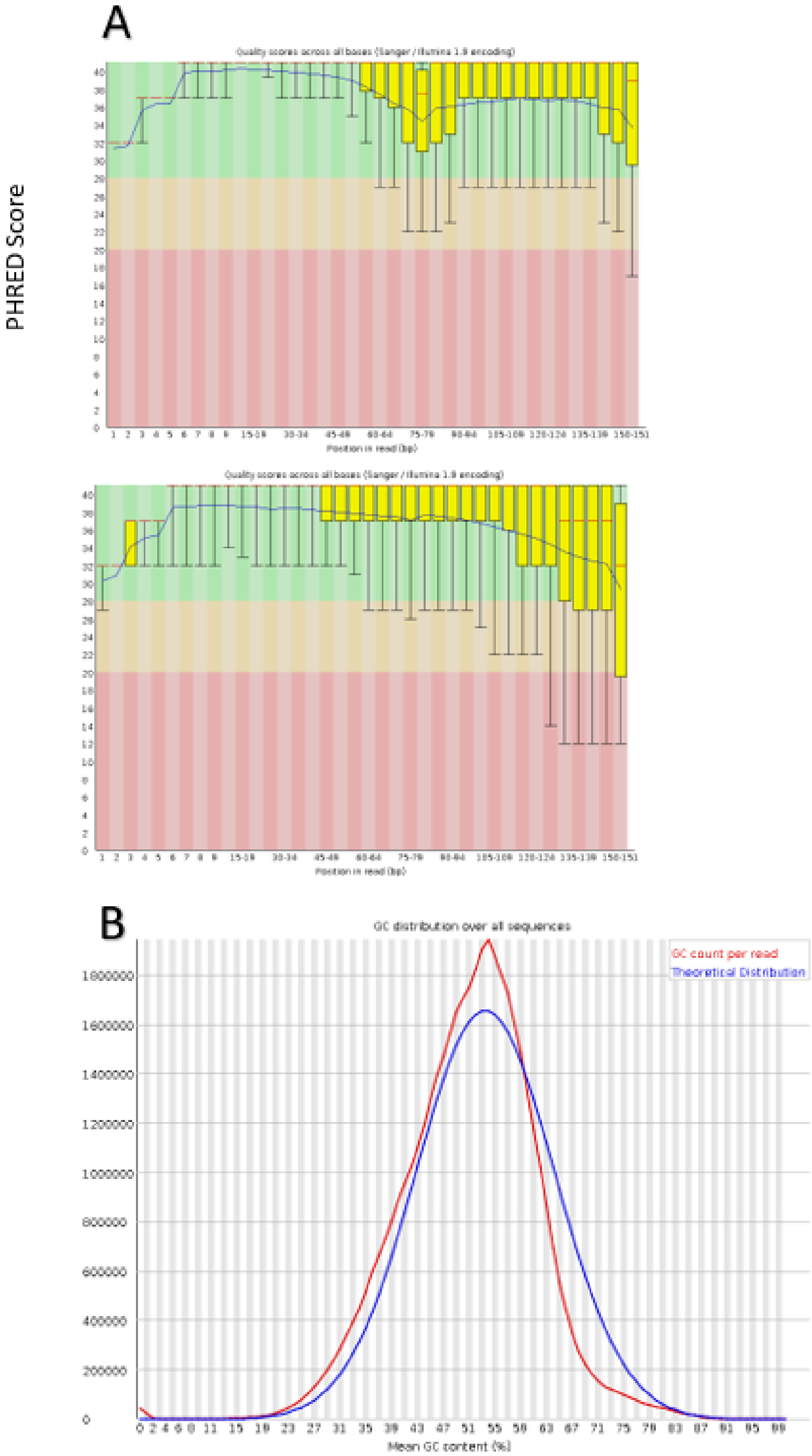


Figure 14. 14a - FastQC read quality profiles for forward (left) and reverse (right) reads of the WI01 library. The Y-axis is the PHRED score, which is a measure of the reliability of the base call. PHRED scores below 20 (shaded red) are considered too unreliable for use in the alignment process. Score of above 30 (shaded green) are considered good quality, which equates to a base calling error rate of 0.1%. The data are presented as whisker-and-box plots. The yellow boxes show inter-quartile range, with central red lines representative of the median value. The blue trace line represents mean quality. Upper and lower whiskers show the 10% and 90% points. 14b – FastQC GC distribution profile for the WC08 library. The blue curve shows the theoretical distribution, while the red curve shows the actual GC % per read. The Y-axis is number of reads, and the X is mean GC %.

Parasite libraries show a GC content reflective of the *L. donovani* genome content at around 59% (Zhou et al. 2004). The GC content of the mouse (and mixed) libraries is slightly lower, with an average of 49.5% (Figure 14b). Though the GC content of the mouse genome as a whole is around 42%, there is a marked difference in the content between introns and exons, at 46% and 53% respectively, which may explain the discrepancy in GC content between the whole genome and the transcriptome (Zhou et al. 2004).

Table 13. Initial FastQC statistics for unprocessed libraries.

GROUP	SAMPLE NUMBER	TOTAL READ COUNT (R1+R2)	GC CONTENT (%)
Parasite inoculum	1	116821560	58
Parasite inoculum	2	77002940	57
Parasite inoculum	3	83813960	59
WT infected	1	29004444	50
WT infected	2	40779458	50
WT infected	3	39964422	50
WT infected	5	118921726	50
WT uninfected	6	131594816	50
WT uninfected	7	86331978	49
WT uninfected	8	89671554	50
RAG infected	11	61641644	50
RAG infected	12	205331092	49
RAG infected	13	50464610	49
RAG infected	14	109784794	50
RAG uninfected	16	50765902	48
RAG uninfected	17	72667356	49
RAG uninfected	18	142699796	49

3.6 CUTADAPT AND SICKLE

3.6.1 CUTADAPT

CutAdapt was used to remove illumina universal sequencing adaptors. This trims the reads to remove any adaptor sequences. CutAdapt had no effect on the total number of sequences or GC content for each library, as it does not remove sequences; even those trimmed entirely are still listed with a length of 0. Instead of being a uniform 151bp long, after CutAdapt, the sequence length varied from 0bp, where the entire read had been dismissed (but still indexed), to 151, where the read was untouched by the CutAdapt process.

After CutAdapt processing, sample library statistics are identical to those in table 13, aside from the differences in sequence length discussed above.

3.6.2 SICKLE

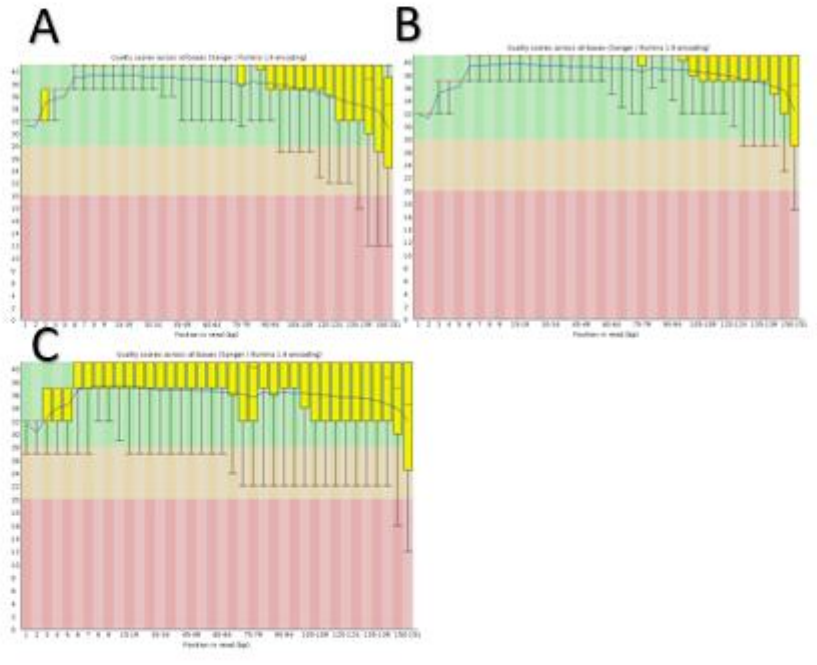
Sickle is another pre-processing program, used to remove unreliable (low quality) base calls from reads. Reads are trimmed until either the read has an overall score of 20 or higher, or the read reaches 20bp. Reads that reach 20bp without improving the score to above 20 are removed. The library GC content of some libraries was changed by 1% by this process.

Given that reads were removed from libraries by Sickle, the range and average of total sequences after CutAdapt and Sickle processing reduced slightly (table 14). The lowered range is 28,510,446 to 201,644,280, and the lowered average number of sequences is 87,164,197.

Sickle produces an additional output file to the forward (R1) and reverse (R2) input files – R0. The R0 file contains singlet discarded reads, which are not used in any further processing or analysis. Sickle is configured to make use of paired-end reads, so if one read in a pair fails quality filtering, both reads are removed. Figure 15 demonstrates the effects on library quality that processing with CutAdapt and Sickle have (15a & b), and also shows the kinds of read that are discarded as R0 (15c).

Table 14. Sample library FastQC statistics after processing with CutAdapt and Sickle.

GROUP	SAMPLE NUMBER	TOTAL READ COUNT (R1+R2)	GC CONTENT (%)
Parasite inoculum	1	114173872	58
Parasite inoculum	2	75575478	57
Parasite inoculum	3	82218296	60
WT infected	1	28510446	50
WT infected	2	40155420	50
WT infected	3	39350032	50
WT infected	5	117154558	50
WT uninfected	6	129431340	50
WT uninfected	7	84860046	49
WT uninfected	8	88157566	50
RAG infected	11	60828052	50
RAG infected	12	201644280	50
RAG infected	13	49458988	50
RAG infected	14	108044906	50
RAG uninfected	16	49931272	48
RAG uninfected	17	71666382	49
RAG uninfected	18	140630422	49



Figures 15a, b & c showing FastQC library profiles of RI11: a, before CutAdapt and Sickle treatment. b, after treatment. c, discarded R0 reads.

3.7 READ ALIGNMENT AND COUNTING

3.7.1 STAR ALIGNMENT

After trimming low-quality bases and reads from libraries, the reads were locally aligned to the reference genome, using the STAR aligner and the index generated from the concatenated genomes of *Mus musculus* GRCm38 (version 84) and *Leishmania donovani* BPK282a1 (version 28). The mouse genome contains 47,643 annotated genes, while the *Leishmania* genome contains 8,195. Reads that have matches to multiple places in the combined genome are flagged and excluded from the final output alignment. STAR counts paired-end mate pairs as one single read, which is why the numbers of input, mapping etc. reads appear half of what the original library contained (table 15). On average, 90.69% of reads mapped uniquely, with 8.49% mapping to multiple places across the genomes (table 15). Reads can map to multiple locations if there is a true extra copy of a gene, such as a duplication, but also if a read has a strong match to more than one location in the genome, and the aligner is unable to determine which position is correct.

Not all reads will have mapped to genomic features. Some will map to introns or repeats, which is why unique and non-unique read percentages do not add up to 100% (table 15). No library had more than 1.35% of reads fail to map to a genomic feature.

Table 15. STAR alignment reports showing proportions of uniquely and non-uniquely mapping reads. Read pairs are treated by STAR as a single read, which is why numbers appear smaller than the original library size.

Sample	Input reads	Uniquely mapping reads	Unique reads (%)	Non-uniquely mapping reads	Non-unique reads (%)	Non-feature mapping reads (%)
Parasite inoculum - 1	57086936	52310083	91.63	4169449	7.30	1.07
Parasite inoculum - 2	37787739	34714858	91.87	2633327	6.97	1.16
Parasite inoculum - 3	41109148	37705204	91.72	2920239	7.10	1.18
WT infected - 01	24956536	22998916	92.12	1713134	6.86	1.02
WT infected - 01	35833191	31179239	87.01	4169317	11.64	1.35
WT infected - 03	70315211	64064282	91.11	5773473	8.21	0.68
WT infected - 05	30414026	26437018	86.92	3745225	12.31	0.77
WT uninfected - 06	100822140	91223063	90.48	8760840	8.69	0.83
WT uninfected - 07	24718494	21556421	87.17	2871928	11.61	1.22
WT uninfected - 08	54022453	48312335	89.43	4989471	9.24	1.33
RAG infected - 11	64715670	59532096	91.99	4905506	7.58	0.43
RAG infected - 12	42430023	39025800	91.98	3185858	7.51	0.51
RAG infected - 13	44078783	40958089	92.92	2920523	6.63	0.45
RAG infected - 14	14255223	13115511	92.00	1078764	7.57	0.43
RAG uninfected - 16	20077710	18187255	90.58	1814232	9.04	0.38
RAG uninfected - 17	19675016	18246089	92.74	1353235	6.88	0.38
RAG uninfected - 18	58577279	52807365	90.15	5455071	9.31	0.54

3.7.2 HT-SEQ COUNT

After alignment, HT-Seq count was used to count the number of reads at each genomic feature, which are coding regions such as genes and RNA genes. Similarly to STAR, HT-Seq count counts pairs of mated reads as one read, so alignment statistics appear halved when compared with the total library size.

In infected mouse samples, an average 98.82% of reads mapped to the mouse genome (table 16). In the uninfected mouse samples, up to 2.94% of reads aligned to the *Leishmania* genome despite the absence of the parasite in the host. This indicates that a small proportion of reads are being attributed to the *Leishmania* genome through random chance. Though this is not ideal, it is possible that this is a consequence of the increased error tolerance settings that were changed in order to compensate for the presence of splice leader sequences in *Leishmania* reads.

Similarly, despite procedures in place to remove mouse cells from the parasite inoculum sample, up to 28.53% of reads still aligned to the mouse genome. The three parasite inoculum technical replicates were all generated from the same RNA sample, stored in the same manner, and processed in a batch together on the same day using the same kit. Despite this, one sample shows considerably less contamination (only 0.4%) compared to the other two technical replicates (19.28% and 28.53%). No explanation for this difference could be established, but the transcriptome results for the replicates were extremely similar.

Given that the data for mouse and *Leishmania* were analysed separately, i.e the lists of mouse and *Leishmania* genes were separated, the issues discussed above have likely not significantly affected analysis.

Table 16. HT-Seq count results showing read alignment statistics.

Sample	Total reads aligned	Reads aligning to mouse	Reads aligning to mouse (%)	Reads aligning to *Leishmania*	Reads aligning to *Leishmania* (%)
PI01	16,736,482	3226871	19.28	13509611	80.72
PI02	11,611,158	3312428	28.53	8298730	71.47
PI03	11,753,093	46997	0.40	11706096	99.60
WI01	5,895,142	5888259	99.88	6883	0.12
WI02	8,251,608	8234004	99.79	17604	0.21
WI03	8,238,564	8225928	99.85	12636	0.15
WI05	23,961,993	23904382	99.76	57611	0.24
WC06	26,321,499	26313533	99.97	7966	0.03
WC07	16,986,980	16981707	99.97	5273	0.03
WC08	17,952,881	17947556	99.97	5325	0.03
RI11	11,728,813	11429324	97.45	299489	2.55
RI12	39,928,136	38717814	96.97	1210322	3.03
RI13	9,676,481	9533990	98.53	142491	1.47
RI14	21,292,528	20949294	98.40	343234	1.60
RC16	10,197,422	10188113	99.99	9309	0.01
RC17	13,881,742	13863407	99.88	18335	0.12
RC18	28,174,339	27347320	97.06	827019	2.94

3.8 SAMPLE METRIC REGRESSION AND CORRELATION ANALYSES

A number of simple regression analyses and t-tests were performed in order to determine significant correlations or relationships between sample metrics such as organ weight, host genetic background, percentage of reads mapping to each genome, and percentage of uniquely mapping reads.

3.8.1 EFFECT OF MOUSE GENETIC BACKGROUND AND PARASITE BURDEN

Table 17 displays data collected from mice at the time of euthanasia. The parasite burden and organ/body weight data was analysed in combination with sequencing and mapping information to determine if any relationships were present.

Table 17. Mouse LDU, body weight and organ weight statistics. Uninfected mice do not have LDU values, as LDU is a measure of parasite burden.

Sample	Body weight (g)	Spleen LDU	Spleen weight (g)	Liver LDU	Liver weight (g)
WI01	21.48	34	0.49	106	1.43
WI02	20.4	42	0.49	100	1.67
WI03	20.53	29	0.47	112	1.36
WI05	21.09	12	0.67	72	1.54
WC06	15.93	N/A	0.07	N/A	0.72
WC07	18.03	N/A	0.06	N/A	0.75
WC08	17.86	N/A	0.06	N/A	0.85
RI11	17.09	21	0.03	1458	0.68
RI12	21.33	37	0.03	827	0.68
RI13	18.65	14	0.03	1267	0.91
RI14	21.22	37	0.04	1717	1.01
RC16	18.58	N/A	0.03	N/A	0.86
RC17	18.93	N/A	0.03	N/A	0.89
RC18	18.67	N/A	0.04	N/A	0.68

No statistically significant difference was observed in the spleen LDU between WT Black 6 mice and RAG2KO mice (table 18). This was to be expected as the mice were euthanised specifically on day 28, when the parasite burdens are expected to be similar despite the presence/absence of an immune system.

Table 18. Welch’s two-sample t-test results

Hypothesis	t-value	Degrees of freedom	p-value	95% confidence interval
LDU and host genetic background	0.23256	5.954	0.829	-19.08267, 23.0867
Percentage of reads mapping to the *Leishmania* genome and host genetic background	-5.2553	3.0318	0.01307	-3.1759404, -0.7890596

The relationships between spleen LDU and spleen weight, liver weight, and total mouse body weight were also investigated and no significant correlation was found (table 19). The spleen and liver weights of WT mice increased considerably during infection as white blood cells infiltrated the tissues. This process was absent in RAG2KO mice. However, despite the differences in organ weight, the LDU remained the same. RAG2KO mice were no different in body weight to WT mice.

Table 19. Regression analysis statistics for parasite burden and host background metrics.

Factor 1	Factor 2	F-statistic	P-value	Significance
Spleen LDU	Spleen weight	0.00283	0.9589	Not significant
Spleen LDU	Liver weight	0.1028	0.7567	Not significant
Spleen LDU	Body weight	1.795	0.2171	Not significant

After alignment of reads to the reference genomes, further regression analyses were performed to determine a relationship between parasite burden and proportion of *Leishmania* reads, if any.

Table 20. Regression analysis statistics for parasite burden and alignment metrics.

Factor 1	Factor 2	F-statistic	P-value	Significance
Spleen LDU	% mapping to *Leish*	0.00053	0.9824	Not significant
Spleen LDU	% mapping to mouse	0.00071	0.9796	Not significant
Spleen weight	% mapping to *Leish*	20.7	0.0039	Significant

Given that the parasite burden was not statistically different between WT and RAG2KO mice, it is unsurprising that no significant relationship was found between spleen LDU and the percentage of reads mapping to either the mouse or *L. donovani* genome, when the LDU was fitted as a single dataset instead of being separated by background (table 20 and figure 16). However, the difference in spleen weights between the two sets of mice was found to show a significant relationship to the percentage of reads mapping to the *Leishmania* genome. A lower proportion of the tissue weight was attributable to mouse cells in RAG2KO samples. In WT samples, most of the tissue weight can be found in the infiltrated white cells. A further t-test confirms that there was a significant (p = 0.0131) difference between the percentage of reads mapping to the *Leishmania* genome for each mouse background (Table 18).

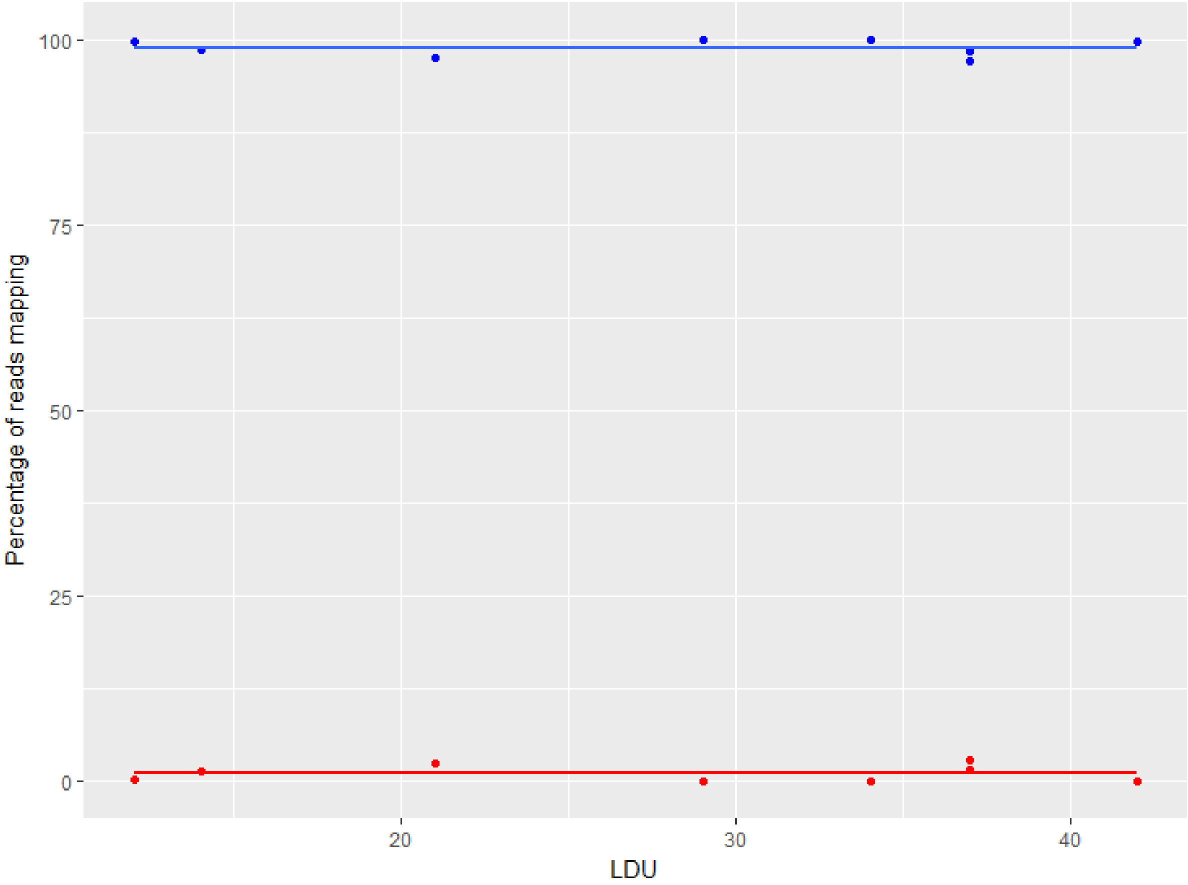


Figure 16. Plot showing the relationship between spleen LDU and the percentage of reads mapping to the *Leishmania* genome. Each library is represented by two points – a blue point showing the percentage of reads mapping to the mouse genome, and a red point showing the percentage of reads mapping to the *Leishmania* genome. The lines show the average percentage for each mouse and *Leishmania* data across LDU.

3.8.2 SEQUENCING COVERAGE STATISTICS

Due to the selective nature of RNA sequencing, statistics on read depth and average coverage are not as useful an indicator for how accurate a representation of the genome – or in this case, transcriptome – is. Instead, comparing the number of genes found in the sample with a known number of genes or annotations can relay how thoroughly represented a transcriptome is.

Table 21. Mouse gene expression statistics. Genes considered “expressed” have a minimum read count of 1 across all samples. Only genes with a non-zero read count were considered for calculating the median.

Sample	No. of mouse genes expressed	No. of mouse genes not expressed	Median no. of reads per gene	Percentage of mouse genes expressed
Inoculum – 01	20777	26866	40	43.6
Inoculum – 02	20674	26969	41	43.4
Inoculum – 03	10292	37351	2	21.6
WT infected - 01	20570	27073	137	43.2
WT infected – 02	19848	27795	193	41.7
WT infected – 03	24052	23591	190	50.5
WT infected - 05	21149	26494	112	44.4
WT uninfected - 06	26383	21260	162	55.4
WT uninfected – 07	22387	25256	79	47.0
WT uninfected – 08	24756	22887	122	52.0
RAG infected – 11	25245	22398	137	53.0
RAG infected – 12	24417	23226	108	51.2
RAG infected – 13	24275	23368	114	51.0
RAG infected – 14	20994	26649	60	44.1
RAG uninfected – 16	21499	26144	70	45.1
RAG uninfected – 17	21813	25830	71	45.8
RAG uninfected - 18	21813	25830	119	45.8

Table 21 shows that for both inoculum and mixed RNA samples, mouse genes are present. 21.6% of mouse genes are still represented in the ‘purest’ inoculum sample. For mixed RNA samples, between 41.7% and 55.4% of mouse genes are found to be expressed, with an average of 47.9%. Mammalian cells would not be expected to express each gene concurrently so an average such as this is not unusual. For expressed genes, the average number of reads is 103, with the median for each sample ranging from 2 – 193. The inoculum samples all have a median of below 41, with the ‘purest’ inoculum sample only having a median of 2 reads per mouse gene.

Table 22. *Leishmania* gene expression statistics. Genes considered "expressed" have a minimum read count of 1 across all samples. Only genes were with a non-zero considered for calculating the median.

Sample	No. of Leish genes expressed	No. of Leish genes not expressed	Median no. of reads per gene	Percentage of Leish genes expressed
Inoculum – 01	8041	154	1151	98.1
Inoculum – 02	8031	164	711	98.0
Inoculum – 03	8028	167	1002.5	98.0
WT infected - 01	3416	4779	1	41.7
WT infected – 02	4649	3546	2	56.7
WT infected – 03	7935	260	72	96.8
WT infected - 05	7715	480	27	94.1
WT uninfected - 06	7991	204	105	97.5
WT uninfected – 07	7812	383	13	95.3
WT uninfected – 08	7902	293	29	96.4
RAG infected – 11	3883	4312	1	47.4
RAG infected – 12	3070	5125	1	37.5
RAG infected – 13	2997	5198	1	36.6
RAG infected – 14	3412	4783	1	41.6
RAG uninfected – 16	5444	2751	2	66.4
RAG uninfected – 17	4812	3383	2	58.7
RAG uninfected - 18	7400	795	5	90.3

Gene expression statistics for *Leishmania* genes can be seen in table 22. For both the inoculum and RAG infected samples, almost 100% *Leishmania* genes can be detected in the sample. The average for the WT infected samples is much lower at 71.8%. However, even in uninfected mouse samples, up to 96.8% of *Leishmania* genes are still being detected with at least 1 read in the sample. The average number of reads per gene in the inoculum samples is 98. For infected mouse samples, the average is 80.1. For uninfected samples, the average is considerably lower, at 52.8 reads per gene, similar to the number of mouse reads in the inoculum samples. The presence of reads aligning to the *Leishmania* genome in uninfected samples is indicative of an underlying technical issue, as few reads – if any – should align to the parasite genome if no parasite is present in the host tissues.

Only 135 of the total 8,224 annotated *Leishmania* genes are without a single read detected in any of the inoculum samples (Data available in accompanying material A1). Given in this analysis, the threshold for expression is the alignment of a single read, very few genes are uniquely expressed in each inoculum sample; only 14 in PI01, 7 in PI02, and 6 in PI03 (figure 17). Similarly, only a few genes are found expressed in two inoculum samples but not the third.

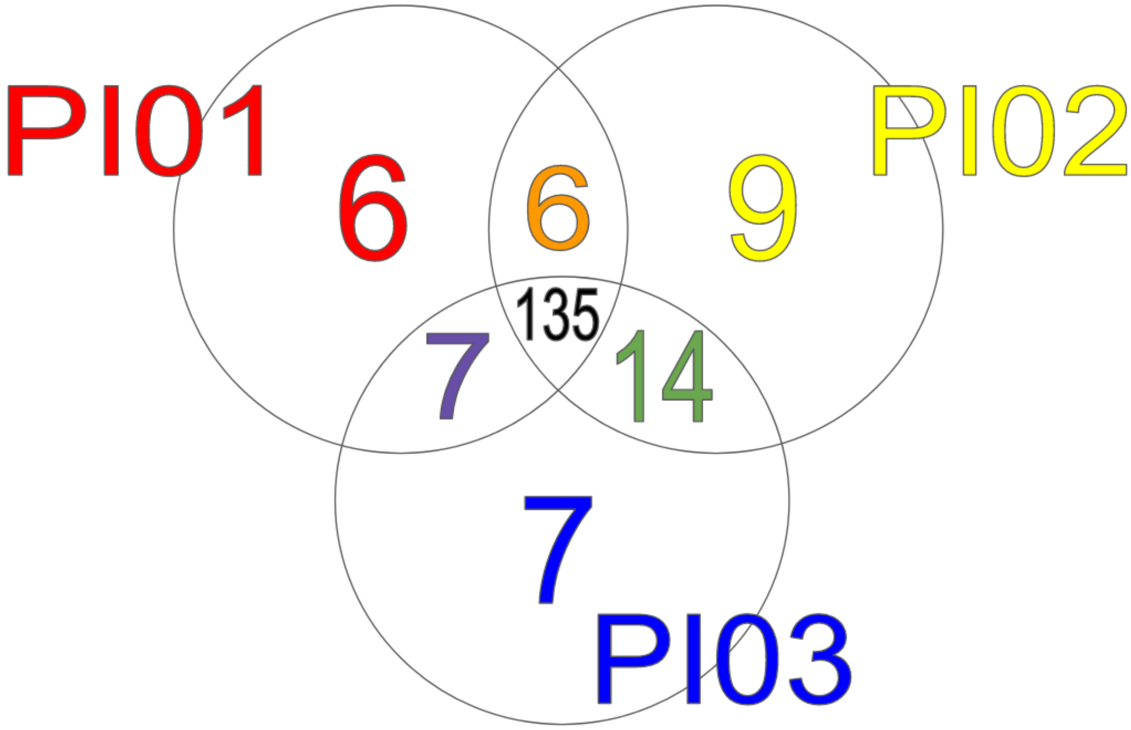


Figure 17. A Venn diagram showing the number of genes *not* expressed in each of the inoculum samples, where expressed means at least one read aligns. For example, 135 are absent from all three samples (black), 9 genes (yellow) are not found in PI02, while 14 genes (green) are missing from PI02 and PI03.

3.9 WT UNINFECTED VS. RAG UNINFECTED

Gene count data produced by HT-Seq count was transformed appropriately for use with edgeR. For comparisons of gene expression between mouse samples, all reads mapping to the *Leishmania* genome were filtered out of the analysis. Similarly, for *Leishmania* analysis, reads aligned to the mouse genome were excluded.

In the following analyses, the read count data interpreted by edgeR was used for comparing patterns of gene expression between samples of different backgrounds. This differential expression data was further processed in order to analyse enrichment and pathway patterns to discover overall trends in the differences and responses between each condition (see figure 6).

This analysis compares the gene expression of uninfected, resting profiles of mice from the two different genetic backgrounds – the WT black 6 mice and the RAG2 KO mice.

3.9.1 EDGER ANALYSIS

3.9.1.1 DIFFERENTIAL GENE EXPRESSION

EdgeR (3.18.1) was used to identify differentially expressed genes through pairwise comparisons between wild type and RAG background mice, and uninfected and infected mice. Genes that were differentially expressed were annotated using gene coordinates from annotations in the TriTrypDB database. Of the 30,976 genes with non-zero read counts, edgeR found that 6,439 were significantly ($p \leq 0.05$) differentially expressed between the uninfected WT and uninfected RAG mice (20.78% of non-zero count annotated genes). Differentially expressed genes were then filtered using a log CPM cut off of 1, and a log2FC of +/- 1.5, the number of differentially expressed genes was reduced to 1,223 (3.95%, 81% failed the logFC/logCPM cutoff) (table 23).

Few of the most significantly differentially expressed genes are directly related to the immune system. A handful – Ifi44, Mt2 and Oas1g – have antiviral connotations, but are more highly expressed in the RAG samples (Kitamura et al. 1994; Kumar et al. 2000; Ghoshal et al. 2001). Clec4a2, a C-type lectin receptor, is known to act as an immune signal receptor, but is also found to be more highly upregulated in RAG samples (Kanazawa et al. 2001). However, Ebf1, a transcription factor associated with B-lymphocyte differentiation, is expressed much more highly in the WT mice (Lin and Grosschedl 1995).

Table 23. The 20 most significantly differentially expressed genes, comparing uninfected WT and RAG samples, ordered by FDR q-value. Full results table in accompanying material A1.

Gene name	logCPM	logFC	FDR	Gene description
Mmrn1	4.85	-2.79	2.78E-26	multimerin 1
Chil3	6.35	-3.10	8.44E-26	chitinase-like 3
Ifi44	5.26	-2.91	1.43E-25	interferon-induced protein 44
Mt2	4.96	-3.11	5.95E-20	metallothionein 2
Npl	4.51	-2.32	2.13E-19	N-acetylneuraminate pyruvate lyase
Oas1g	4.02	-2.49	2.13E-19	2'-5' oligoadenylate synthetase 1G
Gm15675	4.84	2.75	3.12E-19	predicted gene 15675
Fpr2	4.94	-2.59	2.22E-18	formyl peptide receptor 2
Cmah	6.61	2.55	2.86E-18	cytidine monophospho-N-acetylneuraminic acid hydroxylase
Alox12	6.06	-2.10	5.81E-18	arachidonate 12-lipoxygenase
Sbk1	5.89	3.42	9.75E-18	SH3-binding kinase 1
Clec4a2	5.29	-2.26	1.52E-17	C-type lectin domain family 4, member a2
Aff3	5.29	2.85	1.52E-17	AF4/FMR2 family, member 3
Ebf1	5.37	4.30	1.71E-17	early B cell factor 1
Rhof	5.55	1.82	1.88E-17	ras homolog family member F (in filopodia)
Thbs1	7.36	-2.25	2.36E-17	thrombospondin 1
Pde5a	5.16	-2.39	2.36E-17	phosphodiesterase 5A, cGMP-specific
Gm9025	3.23	-3.28	3.27E-17	predicted gene 9025
Rgs18	5.36	-2.50	3.43E-17	regulator of G-protein signaling 18
Ttpal	6.04	1.69	3.77E-17	tocopherol (alpha) transfer protein-like

3.9.1.2 MA PLOT

MA plots are used to demonstrate the differences in fold change and average gene expression between samples or groups of samples. Figure 18 shows the MA plot generated from unfiltered edgeR results. Both the WT and RAG samples show a number of genes that are only expressed in one sample or the other (orange), of which 119 are expressed only in RAG samples, and 468 are expressed only in WT. Slightly more of the differentially expressed genes (red) appear to be expressed in the RAG samples; 6728 genes are expressed more highly in RAG samples, while 6149 are expressed more highly in WT samples. Most genes are expressed between +/- 5 logFC, though expression ranges from -7.5 logFC to +8.6 logFC, with 330 genes being expressed outside the +/- 5 logFC metric.

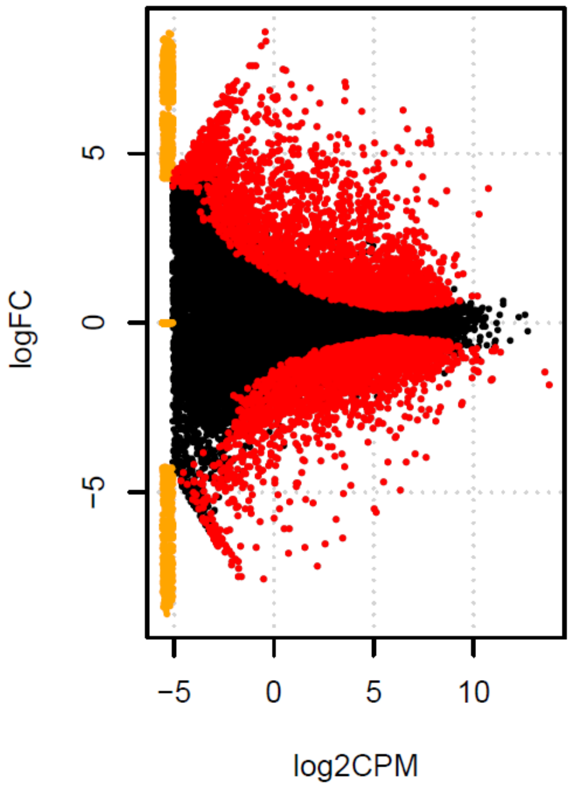


Figure 18. MA plot comparing uninfected WT and RAG samples. Each dot is representative of a gene. Genes with a positive logFC value are more highly expressed in the WT samples; those with a negative logFC value are more highly expressed in the RAG samples. Black dots are not considered differentially expressed, red dots are considered differentially expressed, and orange is representative of genes that are exclusively expressed in one group.

3.9.2 ENRICHMENT ANALYSIS

3.9.2.1 GENE SET ENRICHMENT ANALYSIS

Gene Set Enrichment Analysis was performed via the Broad Institute web interface (version 6.1, last accessed 28/06/17). Lists of differentially expressed genes are submitted to the database and compared with motif (C3), immunological (C7), KEGG Pathway (CP:KEGG) and hallmark (H) sets. Hallmark sets are built by collecting data from the same strictly defined biological process or state; DNA motif sets are built from sequences containing highly conserved cis-regulatory regions.

Significant overlap between gene lists and the defined sets are returned, along with ratio of overlapping genes to set size, p-value, FDR, and number of overlapping genes. By comparing the data with that of the GSEA database, insight can be gained in what kind of immunological, metabolic or other response the host is undergoing, and inidcate what types of cells are being recruited.

Overlaps were detected between the differentially expressed gene list and 42 gene sets in the database (table 24). Of these overlaps, 20 matches were for immunological gene sets, 20 for hallmark gene sets, and 2 for DNA motif sets.

The majority of overlaps were related to immune cells. A wide variety of cell types were listed, including CD4 and CD8 T-cells, B-cells, monocytes, myeloid cells, myeloid and plasmacytoid DCs, macrophages and NK cells. IL-10, IL-6 and IL-2 were listed as notable stimulating factors. Several immune-related overlaps were detected for Hallmark sets, such as complement response, inflammatory response, and interferon gamma response.

Table 24. Gene Set Enrichment Analysis results for genes differentially expressed between uninfected RAG2 KO and uninfected WT mice, ordered by FDR/q-value. P-value and overlap size have been omitted. Only the top 20 overlaps are reported when using the Broad Institute web interface. Full results table in accompanying material A2.

Gene Set Description	Genes in set	Overlapping genes / set size	FDR q-value
Genes having at least one occurrence of the highly conserved motif M92 TGCTGAY in the region spanning up to 4 kb around their transcription start sites. The motif does not match any known transcription factor binding site.	1085	0.0857	1.39E-36
Genes up-regulated in comparison of healthy B cells versus healthy myeloid cells.	200	0.23	3.35E-36
Genes down-regulated in T cells: CD8A versus CD8A CD8B.	200	0.23	3.35E-36
Genes up-regulated in comparison of naive CD8 T cells versus day 0 monocytes.	200	0.215	1.74E-32
Genes up-regulated in comparison of naive CD4 CD8 T cells versus monocytes cultured for 0 days.	200	0.21	1.79E-31
Genes up-regulated in comparison of B cells from influenza vaccinee at day 7 versus monocytes from influenza vaccinee at day 7.	200	0.21	1.79E-31
Genes down-regulated in bone marrow-derived macrophages with IL10 knockout and 45 min of stimulation by: LPS versus LPS and IL10.	200	0.21	1.79E-31
Genes up-regulated during transplant rejection.	200	0.2	4.58E-29
Genes up-regulated in comparison of systemic lupus erythematosus CD4 T cells versus systemic lupus erythematosus myeloid cells.	200	0.195	4.97E-28
Genes up-regulated in comparison of B cells from influenza vaccinee at day 7 post-vaccination versus myeloid dendritic cells (mDC) at day 7 post-vaccination.	200	0.195	4.97E-28
Genes down-regulated in NKT cells versus CD8A T cells.	200	0.195	4.97E-28
Genes down-regulated in monocyte-derived dendritic cells: control versus treated with LGALS1.	200	0.195	4.97E-28
Genes down-regulated in untreated spleen: DUSP1 knockout versus wildtype.	177	0.209	1.03E-27
Genes down-regulated in B lymphocytes: expressing IgM BCR fusion and untreated versus expressing IgMG BCR fusion and treated by anti-HEL.	166	0.2169	1.45E-27
Genes up-regulated in comparison of naive CD4 T cells versus day 0 monocytes.	200	0.19	5.89E-27
Genes up-regulated in comparison of B cells versus monocytes.	200	0.19	5.89E-27
Genes having at least one occurrence of the highly conserved motif M7 TGANTCA in the region spanning up to 4 kb around their transcription start sites.	2485	0.0487	7.23E-26
Genes down-regulated in B lymphocytes treated by anti-HEL and expressing BCR fusions with: IgM versus IgMG.	170	0.2	8.14E-25
Genes up-regulated in comparison of naive B cells versus day 0 monocytes.	200	0.18	1.01E-24
Genes up-regulated in comparison of B cells from influenza vaccinee at day 7 post-vaccination versus plasmacytoid dendritic cells (pDC) at day 7 post-vaccination.	200	0.18	1.01E-24

3.9.2.2 GENE ONTOLOGY ANALYSIS

Similarly to GSEA, gene ontology (GO) analysis examines lists of differentially expressed genes for enrichment of particular biological categories, such as metabolisms, enzyme activity, or pathways. The categories are divided into biological process, which checks for biological sub-systems such as the immune system; molecular function, which looks at less high-level processes such as protein activity; and cellular component, which finds patterns in the cellular localisation of genes.

A total of 59 GO terms were found to be enriched in the DEG list provided to GOrilla (last accessed 05/06/17); 44 biological processes, 3 molecular functions and 12 cellular components (table 25). The average p-value was 3.22E-04, ranging from 9.98E-04 to 3.61E-09.

Although many GO categories are very broad, such as “biological regulation”, “localisation” or “binding”, enrichment of more narrow categories can be insightful. For this DEG list, multiple GO terms involving peptidase, nitrogen compound metabolism and hydrolase activity are mentioned (figure 19, table 25). Nitrogen compound metabolism is also shown with a large bubble, indicating several related genes are involved.

For hypergeometric tests such as those performed in Gene Ontology Analysis, a threshold p-value of 0.05 is fairly relaxed. However, the application of more stringent p-values, such as p = 0.01 or 0.001 to our results considerably reduces the number of informative GO terms; when p = 0.001 only 4, very general GO terms are reported: “cellular process”, “single-organism process”, “single-organism cellular process”, and “positive regulation of biological process”. Similarly, with p = 0.01, the only standout term listed was “positive regulation of hydrolase activity”.

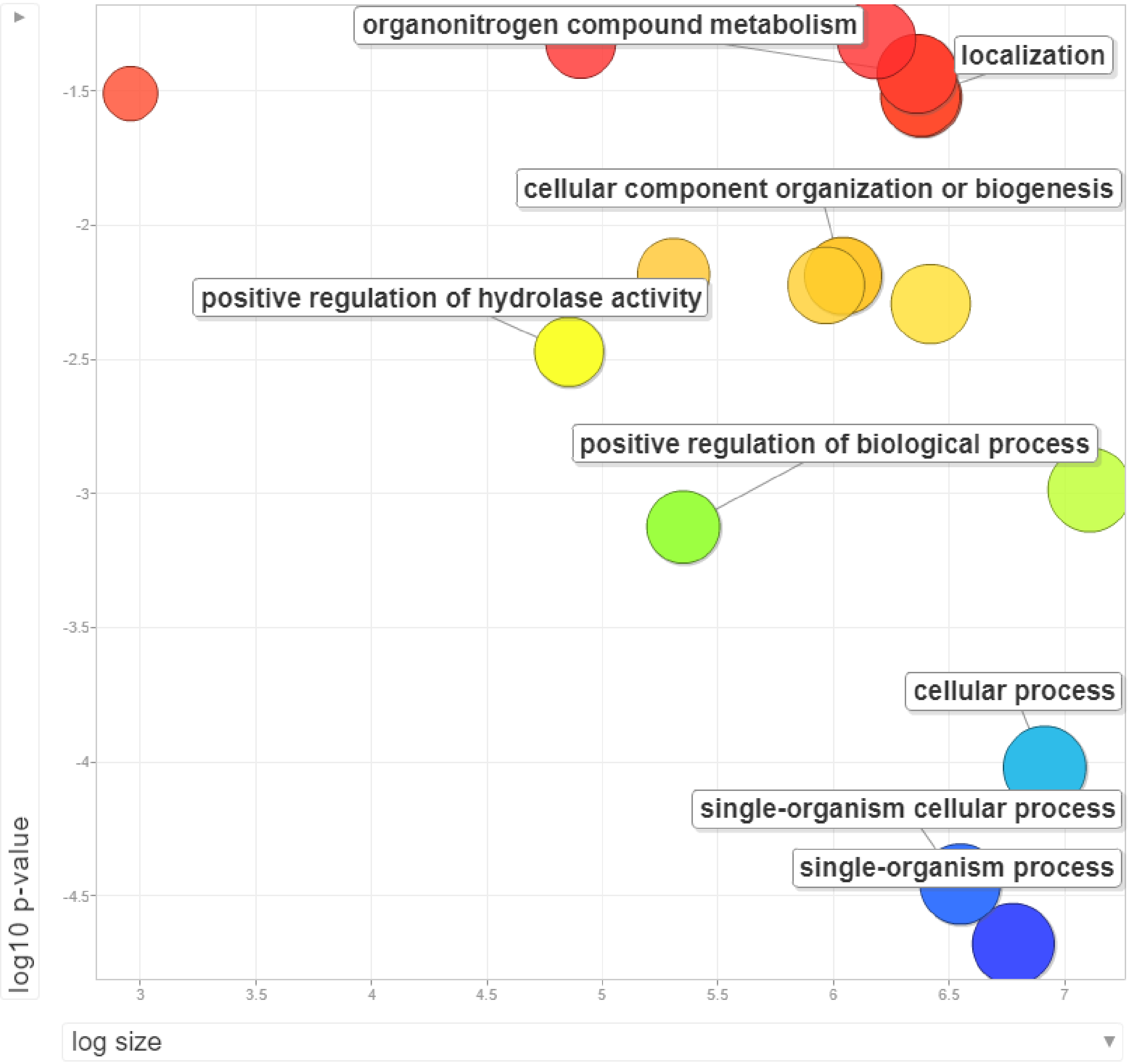


Figure 19. A REVIGO plot visualising the most enriched terms (p ≤ 0.05) in the uninfected WT vs uninfected RAG differentially expressed gene list. Significance is shown on the y-axis in log10 p-value. The number of genes and terms collapsed into the category are displayed by both the x-axis and the bubble size.

Table 25. Top 20 Gene Ontology results for the genes differentially expressed between uninfected Black 6 and uninfected RAG2 KO mice. Process = biological process, function = molecular function, component = cellular component. Full results table in accompanying material A3.

Type	Description	p-value
Process	single-organism process	3.61E-09
Process	single-organism cellular process	1.21E-08
Process	cellular process	4.93E-08
Process	positive regulation of biological process	5.19E-07
Process	biological_process	8.91E-07
Process	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	1.15E-06
Process	positive regulation of peptidase activity	2.43E-06
Process	positive regulation of endopeptidase activity	2.43E-06
Process	positive regulation of cysteine-type endopeptidase activity	2.43E-06
Process	positive regulation of hydrolase activity	5.84E-06
Process	biological regulation	9.68E-06
Process	cellular component organization or biogenesis	1.34E-05
Process	cellular component organization	1.34E-05
Process	positive regulation of cellular process	1.59E-05
Function	protein binding	9.21E-06
Function	binding	9.95E-06
Component	cell part	8.01E-07
Component	intracellular organelle	7.52E-06
Component	intracellular part	1.02E-05
Component	intracellular vesicle	2.13E-05

3.9.3 PATHVIEW

PathView (1.14.0) was used to visualise differentially expressed genes across whole pathways, based on KEGG IDs. The web interface can be used to automatically select appropriate pathways to draw from the data based on enrichment significance, or manual pathways can be selected for print out. Both automatically selected pathways and manually chosen immunology and infection-related pathways were chosen for visualisation.

Pathways automatically selected for the uninfected WT vs. uninfected RAG data are ‘neuroactive ligand-receptor interaction’, ‘phagosome’ and ‘NK-cell mediated cytotoxicity’. Figure 20 displays the pathway for NK-cell mediated cytotoxicity. Genes such as Fas Ligand and Granzyme are upregulated in the RAG sample, whereas genes such as Linker for Activation of T-cells (LAT) and ZAP70 are upregulated in the WT sample. The absence of a colour gradient – i.e only red or green – is indicative of extremes of expression, likely a result of the filtering process in which genes with small fold-change differences were removed from the analysis.

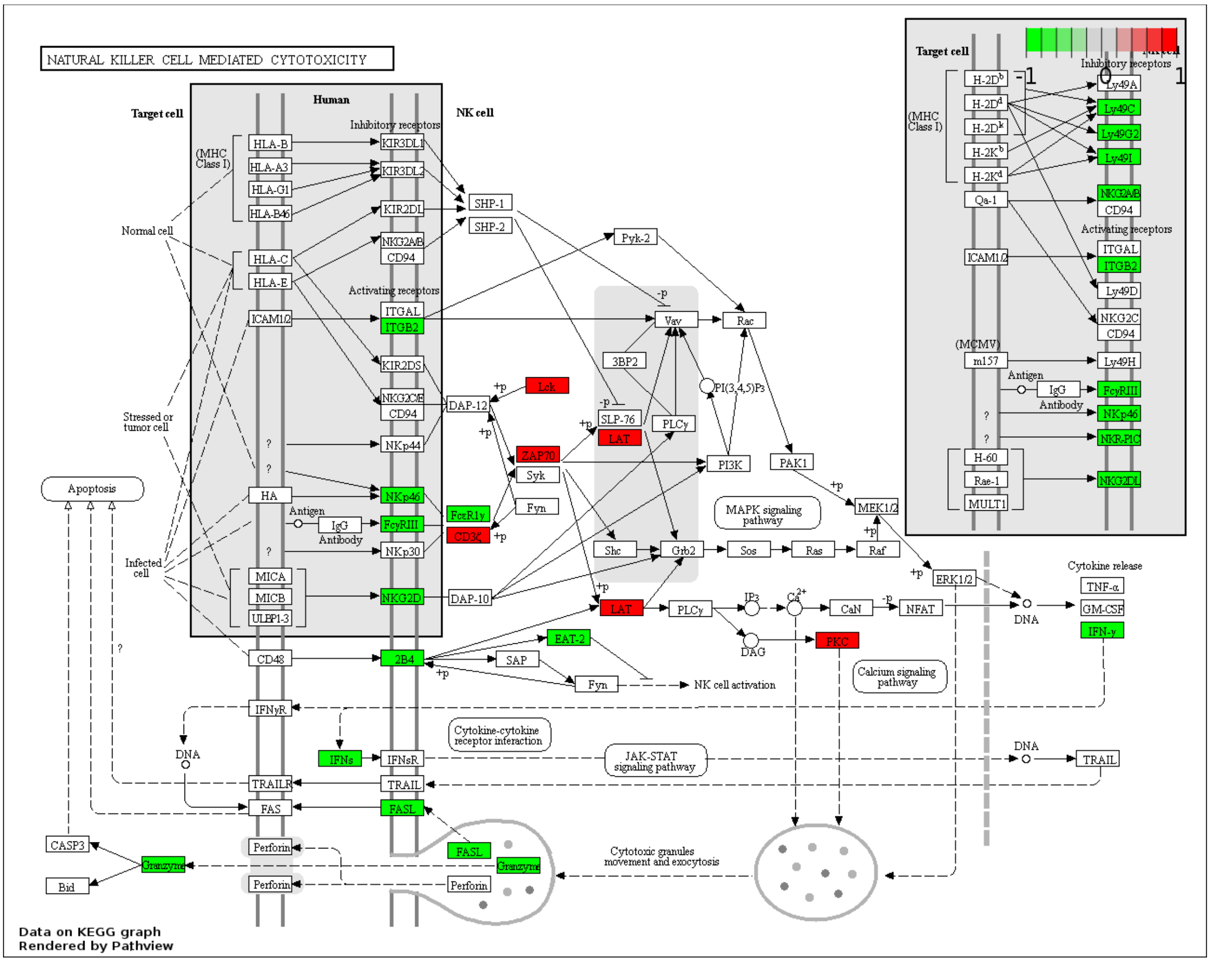


Figure 20. Natural Killer Cell Mediated Cytotoxicity. Each box is a gene in the pathway; Genes displayed in red are upregulated in the WT sample, and green genes are upregulated in the RAG sample. Genes displayed in white are not significantly differentially expressed.

Additionally, pathways suggested by GOrilla enrichment data and immunity-related pathways were also generated. Figure 21, shows the T-cell receptor signalling pathway, the majority of genes (18) in which are significantly upregulated in the WT sample, compared to a single gene more highly upregulated in the RAG samples. The B-cell receptor signalling pathway (not shown) has an almost identical pattern, in that all genes in the pathway that fold change data is present for show upregulation in the WT sample.

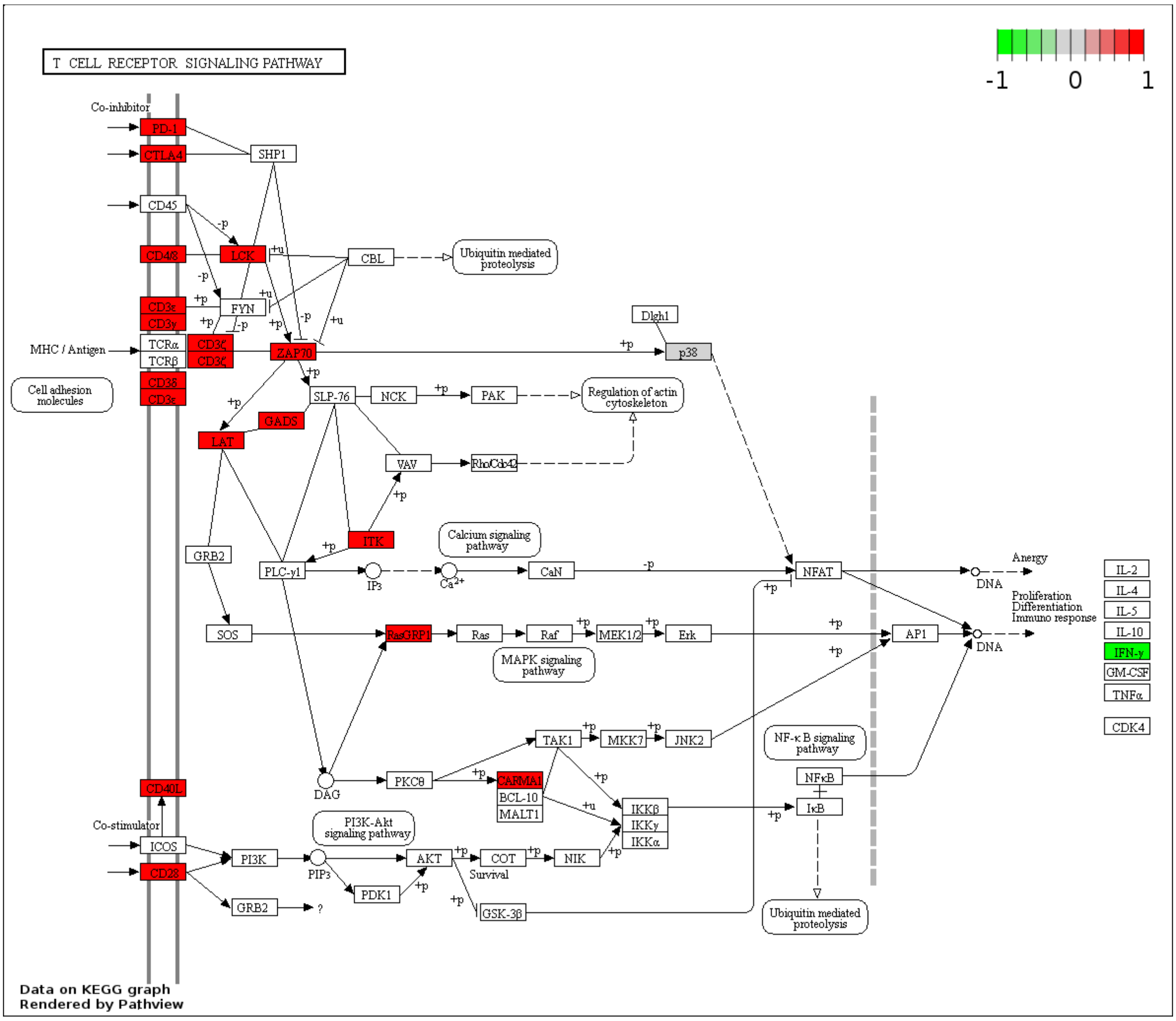


Figure 21. T-cell Receptor Signalling Pathway. Each box is a gene in the pathway; Genes displayed in red are upregulated in the WT sample, and green genes are upregulated in the RAG sample. White genes are not significantly differentially expressed.

3.10 WT INFECTED VS. RAG INFECTED

This section reports the results of comparing gene expression data from infected mice of the two different genetic backgrounds, WT B6 and RAG2 KO.

3.10.1 EDGER ANALYSIS

3.10.1.1 DIFFERENTIAL GENE EXPRESSION

EdgeR found 3,725 genes differentially expressed between the infected WT and infected RAG groups, 12.05% of the 30,976 non-zero read count genes listed. After application of the stringency logCPM and logFC criteria, 1,069 were found to be differentially expressed (3.45%, filtering out the 71.3% of genes which were below the logFC/logCPM threshold) (table 26).

A number of killer-lectin related genes are found to be significantly more expressed in RAG mice, such as Klra13-ps, Klrb1c, Klrd1, Klrd2 and Klri2. Klra13-ps is a pseudogene, but may be expressed at low levels as a transcript, primarily in bladder and spleen cells (NCBI gene database, last accessed 20/04/18, https://www.ncbi.nlm.nih.gov/gene/16631).

Other killer-lectin related genes are CD markers, for example Klrb1c, which in humans marks IL-17 producing T-cells (Maggi et al. 2010). Klri2 is associated with NK cell development (Saether et al. 2005). Two genes associated with antiviral response are also more significantly upregulated in RAG mice – Prf1 and Ncr1 (Gazit et al. 2006; van Dommelen et al.2006).

Table 26. The 20 most significantly differentially expressed genes, comparing infected WT and RAG samples, ordered by FDR q-value. Full results table in accompanying material A1.

Gene name	logCPM	logFC	FDR	Gene description
Klra13-ps	5.69	-4.09	3.02E-34	killer cell lectin-like receptor subfamily A, member 13, pseudogene
Adamts14	5.09	-3.11	1.34E-30	a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 14
Klrb1c	4.36	-3.70	2.24E-29	killer cell lectin-like receptor subfamily B member 1C
Prf1	5.27	-2.71	3.05E-29	perforin 1 (pore forming protein)
Klrd1	5.58	-2.41	7.51E-29	killer cell lectin-like receptor, subfamily D, member 1
Klrc2	4.85	-3.60	2.34E-27	killer cell lectin-like receptor subfamily C, member 2
Cd209a	4.08	-4.57	1.04E-26	CD209a antigen
Btnl9	4.34	-3.13	1.40E-25	butyrophilin-like 9
Ncr1	5.27	-4.72	5.55E-25	natural cytotoxicity triggering receptor 1
Ltbp4	5.99	-2.96	5.24E-24	latent transforming growth factor beta binding protein 4
Klra4	4.49	-5.11	1.05E-23	killer cell lectin-like receptor, subfamily A, member 4
Gm9025	3.23	-3.63	1.17E-23	predicted gene 9025
Igfbp3	6.71	-2.28	1.42E-23	insulin-like growth factor binding protein 3
Klri2	4.78	-3.64	1.96E-23	killer cell lectin-like receptor family I member 2
Cxcl12	9.20	-3.26	1.96E-23	chemokine (C-X-C motif) ligand 12
Gzma	6.78	-3.26	3.45E-22	granzyme A
Abcc3	7.03	-1.86	4.32E-22	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
Fbln5	5.88	-1.93	8.95E-22	fibulin 5
Chil1	6.87	2.64	2.23E-21	chitinase-like 1
Camk2n1	5.23	-2.29	1.69E-20	calcium/calmodulin-dependent protein kinase II inhibitor 1

3.10.1.2 MA PLOT

Figure 22 shows an MA plot comparing expression in infected WT and RAG samples. 5595 genes are more highly upregulated in the WT infected sample, compared to 7283 genes more highly expressed in the RAG infected samples. The range of fold change is from -10.2 logFC to 6.4 logFC, though 96 only genes are expressed +/- 5 logFC. Both the WT and RAG samples show a large number of uniquely expressed genes (orange), of which 98 are unique to the WT mice and 304 are unique to the RAG mice.

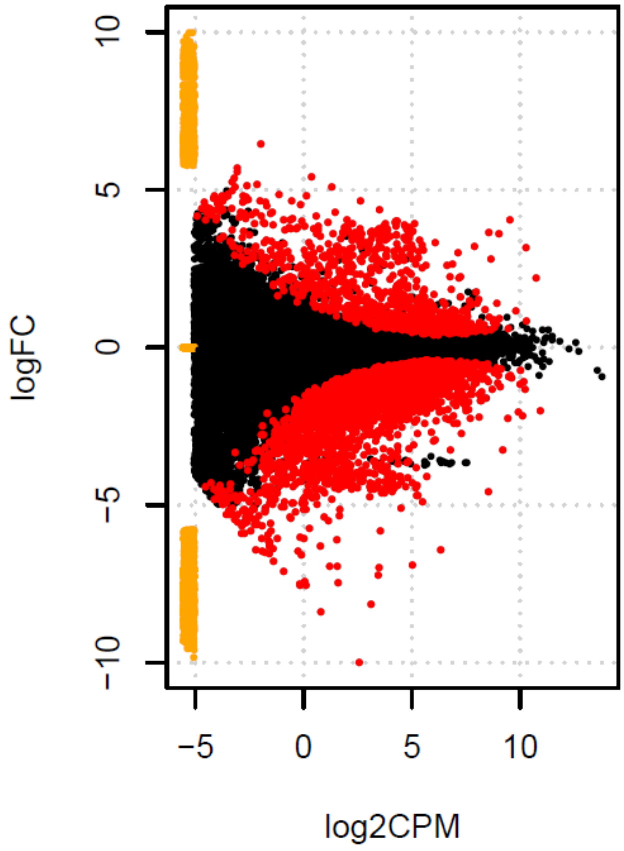


Figure 22. MA plot comparing infected WT and RAG samples. Each dot is representative of a gene. Genes with a positive logFC value are more highly expressed in the WT samples; those with a negative logFC value are more highly expressed in the RAG samples. Black dots are not considered differentially expressed, red dots are considered differentially expressed, and orange is representative of genes that are exclusively expressed in one group.

3.10.2 ENRICHMENT ANALYSIS

3.10.2.1 GENE SET ENRICHMENT ANALYSIS

52 overlaps were detected between the DEG list and the GSEA database sets (table 27). 12 of these overlaps were for DNA motifs, 20 for hallmark gene sets, and 20 for immunologic gene sets. The number of overlapping genes ranged from 8 to 141, with an average of 38. The average p-value was 2.28E-06, with a range of 1.04E-04 to 1.21E-40.

As with the uninfected comparison of the two mouse backgrounds, the majority of GSEA overlaps were for immune cells and immune response. Cell types flagged include myeloid, B-cells, memory B-cells, CD4 and CD8 T-cells, helper T-cells, macrophages, T-regs, and DCs. Cytokines such as IL-6, IL-2, IL-10 and TGFβ were found to be associated with the overlapping gene sets. Hallmark sets contained responses to infection, such as overlap with the complement pathway.

Table 27. Gene Set Enrichment Analysis results for genes differentially expressed between infected and uninfected RAG2 KO mice, ordered by FDR/q-value. P-value and overlap size have been omitted. Only the top 20 overlaps are reported when using the Broad Institute web interface. Full results table in accompanying material A2.

Gene Set Description	Genes in set	Overlapping genes / set size	FDR/q-value
Genes having at least one occurrence of the highly conserved motif M7 TGANTCA in the region spanning up to 4 kb around their transcription start sites.	2485	0.0567	7.17E-37
Genes having at least one occurrence of the highly conserved motif M127 WWTAAGGC in the region spanning up to 4 kb around their transcription start sites.	1896	0.0628	8.76E-35
Genes encoding proteins involved in processing of drugs and other xenobiotics.	200	0.205	6.49E-30
Genes having at least one occurrence of the highly conserved motif M169 TTTNNANAGCYR in the region spanning up to 4 kb around their transcription start sites.	2274	0.0536	1.76E-29
Genes having at least one occurrence of the highly conserved motif M55 TGGAAA in the region spanning up to 4 kb around their transcription start sites.	2061	0.0543	2.52E-27
Genes having at least one occurrence of the highly conserved motif M39 TCCCRNNRTGC in the region spanning up to 4 kb around their transcription start sites.	1121	0.0696	6.46E-25
Genes encoding components of blood coagulation system; also up-regulated in platelets.	138	0.2174	1.62E-22
Genes having at least one occurrence of the highly conserved motif M172 TTGCWCAAY in the region spanning up to 4 kb around their transcription start sites.	1972	0.0507	5.75E-22
Genes having at least one occurrence of the highly conserved motif M114 YTCCCRNNAGGY in the region spanning up to 4 kb around their transcription start sites. The motif does not match any known transcription factor binding site.	1296	0.0602	4.20E-21
Genes down-regulated in T reg (FOXP3+) cells from B6 mice: Foxp3-Fusion-GFP versus Foxp3-ires-GFP.	172	0.1802	7.35E-21
Genes having at least one occurrence of the highly conserved motif M46 WTTGKCTG in the region spanning up to 4 kb around their transcription start sites.	722	0.0762	5.22E-19
Genes having at least one occurrence of the highly conserved motif M139 AAAYWAACM in the region spanning up to 4 kb around their transcription start sites.	1890	0.0481	2.04E-18
Genes up-regulated in CD4 follicular helper T cells (Tfh) with SH2D1A knockout versus wildtype Tfh cells.	200	0.15	6.77E-18
Genes down-regulated in macrophages: wildtype versus MYD88 knockout.	200	0.15	6.77E-18
Genes having at least one occurrence of the highly conserved motif M41 TGACAGNY in the region spanning up to 4 kb around their transcription start sites.	1524	0.0512	4.21E-17
Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	200	0.145	6.75E-17
Genes having at least one occurrence of the highly conserved motif M161 TTANWNANTGGM in the region spanning up to 4 kb around their transcription start sites.	738	0.0678	3.35E-15
Genes down-regulated in CD4 T cells treated with IL6: STAT3 knockout versus wildtype.	200	0.135	6.89E-15
Genes up-regulated in HMC-1 (mast leukemia) cells: untreated versus incubated with the peptide ALL1 followed by stimulation with T cell membranes.	200	0.13	6.56E-14
Genes having at least one occurrence of the highly conserved motif M27 TGGNNNNNNKCCAR in the region spanning up to 4 kb around their transcription start sites.	919	0.0577	2.60E-13

3.10.2.2 GENE ONTOLOGY ANALYSIS

Gene ontology analysis with GOrilla found 69 significantly enriched terms, of which 63 were biological processes, 4 molecular functions, and 2 cellular components (table 28). The average p-value was 2.98E-04, ranging from 9.38E-04 to 1.14E-06.

GO terms involving cell motility, cell migration and cell movement are particularly abundant. Terms associated with phosphorus metabolism and the negative regulation of phosphorylation also appear several times. Additionally, terms related to nitrogen compounds and their metabolism appear multiple times in the analysis results (figure 23). Use of a stricter p-value reduces significant biologically meaningful GO terms; for p = 0.01, only “nitrogen compound metabolic process” and “regulation of cell migration”, and at 0.001, only “regulation of locomotion” is considered significant.

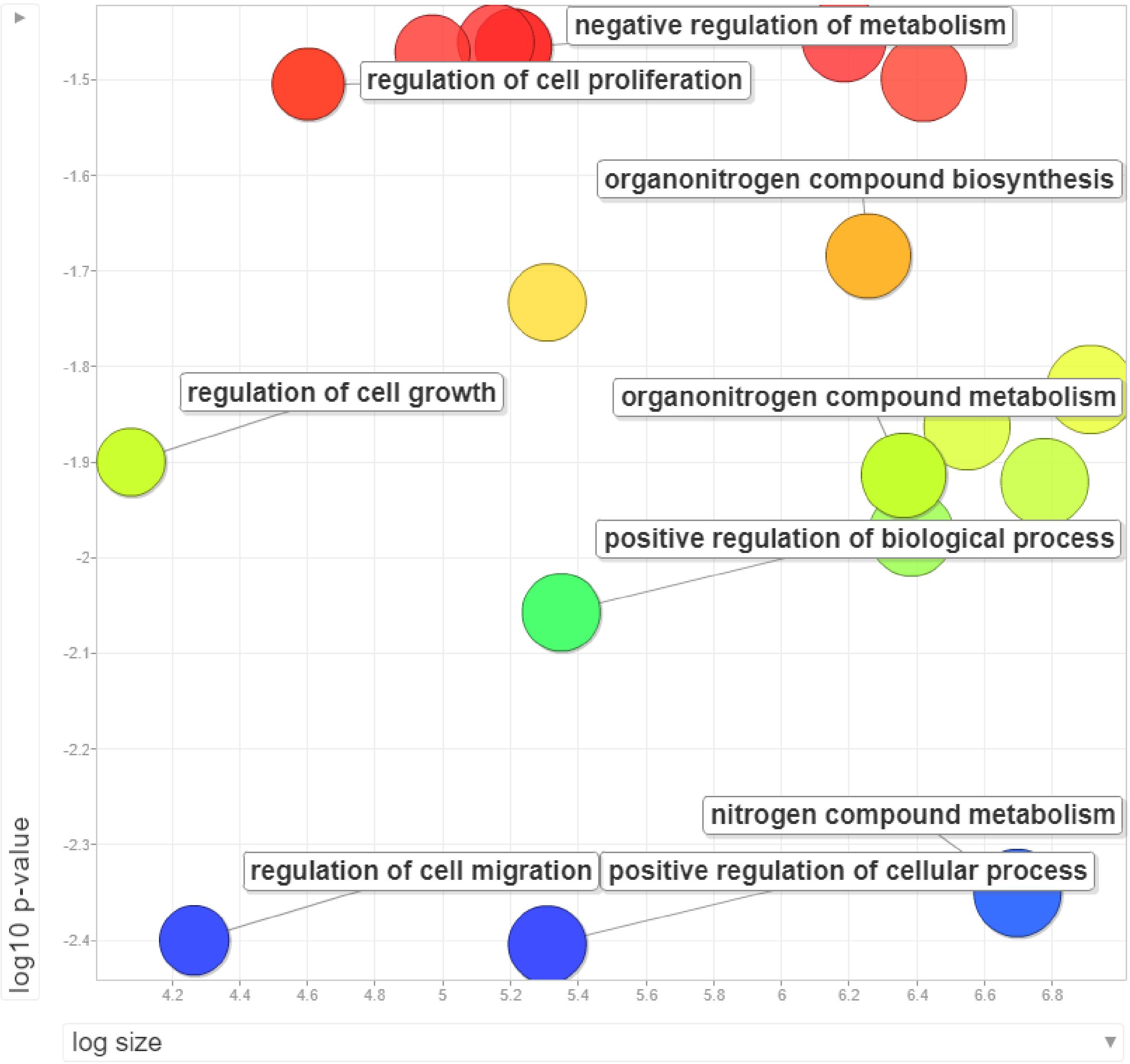


Figure 23. A REVIGO plot visualising the most enriched terms ($p \leq 0.05$) in the infected WT vs infected RAG differentially expressed gene list. Significance is shown on the y-axis in log10 p-value. The number of genes and terms collapsed into the category are displayed by both the x-axis and the bubble size.

Table 28. Top 20 Gene ontology analysis results for genes differentially expressed between infected Black 6 mice and infected RAG2 KO mice, ordered by p-value. Process = biological process, function = molecular function, component = cellular component. Full results table in accompanying material A3.

Type	Description	p-value
Process	regulation of locomotion	1.14E-06
Process	regulation of cell motility	1.21E-06
Process	positive regulation of cellular process	1.97E-06
Process	regulation of cellular component movement	2.67E-06
Process	regulation of cell migration	3.32E-06
Process	nitrogen compound metabolic process	4.46E-06
Process	positive regulation of biological process	1.02E-05
Process	regulation of cellular process	1.42E-05
Process	regulation of cell growth	1.89E-05
Process	single-organism process	2.00E-05
Process	organonitrogen compound metabolic process	2.24E-05
Process	single-organism cellular process	2.74E-05
Process	cellular process	3.25E-05
Process	regulation of cellular component organization	4.31E-05
Process	organonitrogen compound biosynthetic process	5.19E-05
Process	regulation of biological process	6.72E-05
Process	negative regulation of metabolic process	9.71E-05
Process	positive regulation of cell motility	1.05E-04
Component	envelope	1.01E-04
Component	organelle envelope	1.01E-04

3.10.3 PATHVIEW

Pathways selected by enrichment data for the genes differentially expressed between infected WT and infected RAG mice include ‘Steroid hormone biosynthesis’, ‘ppar signalling pathway’ and ‘coagulation and complement cascades’. Figure 24 shows the coagulation and complement pathway. The majority of genes, in this case 21, are upregulated in the RAG infected sample, with the sole exception of EPCR, endothelial protein C receptor. As with the comparison of the uninfected WT and RAG samples, genes showing fold change differences between 1.5 and -1.5 logFC were filtered out of the analysis, resulting in Pathview displaying extremes of relative expression – only red and green.

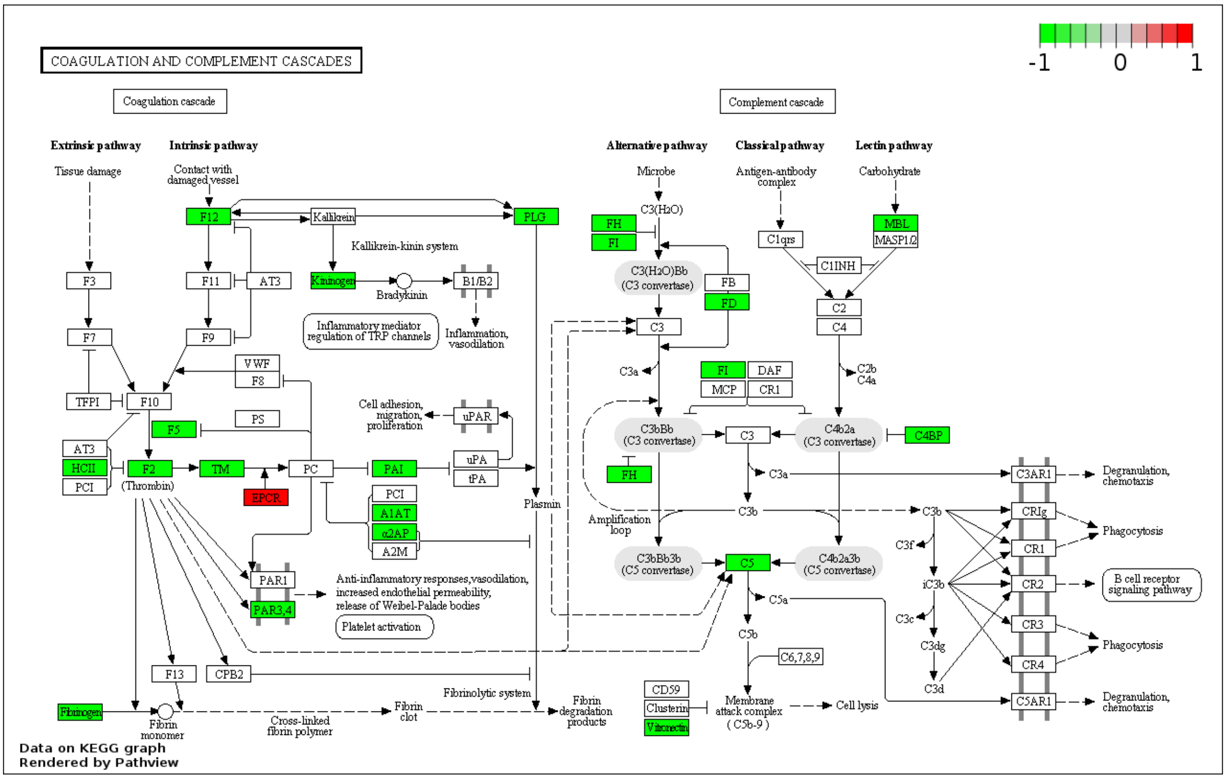


Figure 24. Coagulation and Complement Cascades. Each box is a gene in the pathway; Genes displayed in red are upregulated in the WT sample, and green genes are upregulated in the RAG sample. White genes are not significantly differentially expressed.

Immunity-related pathways were also generated. T-cell and B-cell receptor pathways (not shown) very closely resembled those of the uninfected samples, i.e. the majority of genes in the pathway were highly upregulated in the WT samples. Expression data was available for 3 genes in the 'Leishmaniasis' pathway (figure 25). Interferon-γ, COX2 and iNOS show upregulation in the WT.

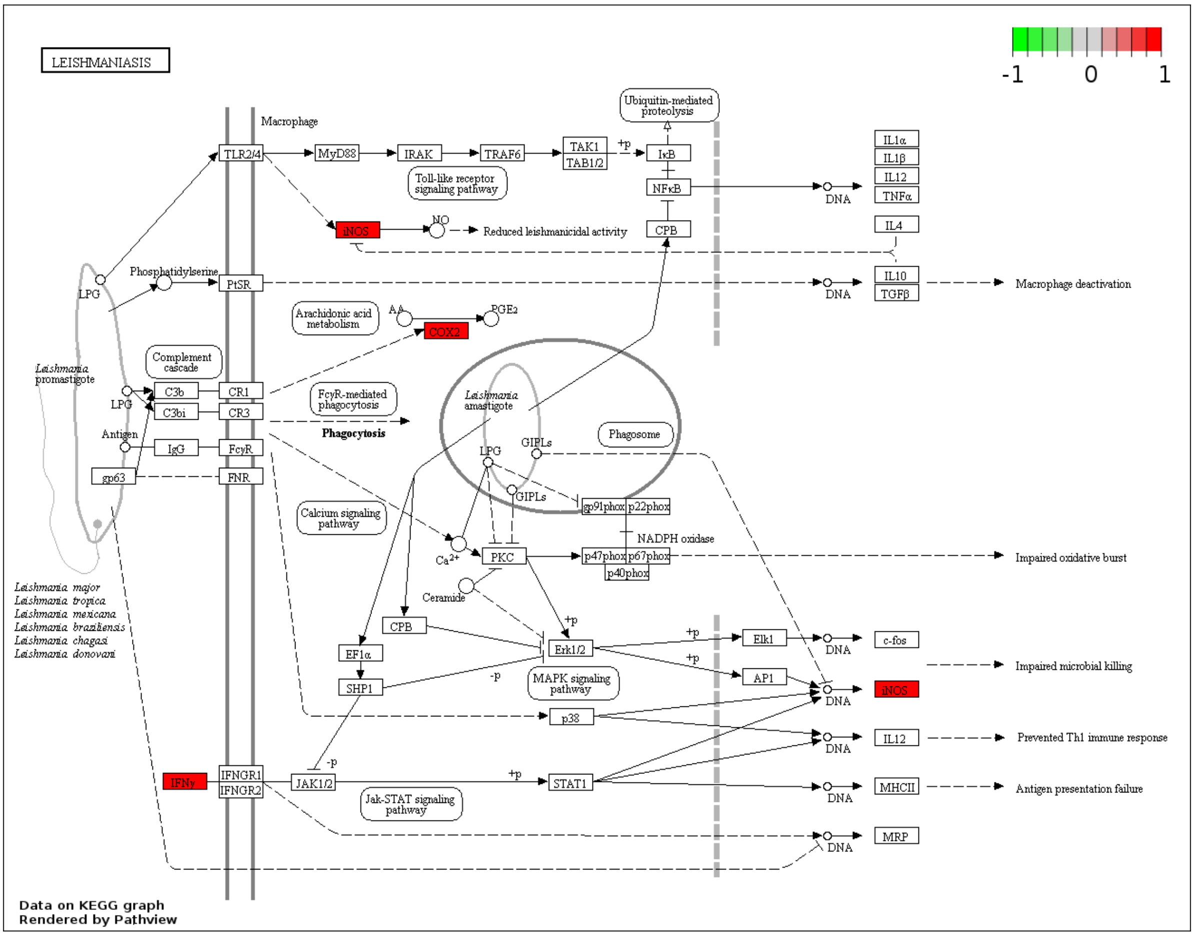


Figure 25. Leishmaniasis Pathway. Each box is a gene in the pathway; Genes displayed in red are upregulated in the WT sample, and green genes are upregulated in the RAG sample. White genes are not significantly differentially expressed.

3.11 WT UNINFECTED VS. WT INFECTED

These results compare gene expression data from WT B6 mice in infected and uninfected states.

3.11.1 EDGER ANALYSIS

3.11.1.1 DIFFERENTIAL GENE EXPRESSION

After application of the negative binomial distribution and generalised linear model, edgeR produced a list of 6,553 genes differentially expressed between infected and uninfected WT samples, which is 21.15% of the 30,976 total. After application of logCPM and logFC criteria, the number of genes considered differentially expressed reduced to 1,421, 4.59% of the total, filtering out the 78.32% of genes which failed the logFC/logCPM criteria (table 29).

Two genes involved in the production of reactive oxide species – Fprl1 and Nos2 – are highly upregulated in the infected mice, most likely in response to infection (Bylund et al. 2002). Clec4e and Lilr4b are both immunoreceptors, associated with macrophages and dendritic cells respectively, are both more highly expressed in infected mice (Cao et al. 2006; Wells et al. 2008).

Table 29. The 20 most significantly differentially expressed genes, comparing infected and uninfected WT samples, ordered by FDR q-value. Full results table in accompanying material A1.

Gene name	logCPM	logFC	FDR	Gene description
Chil1	6.87	4.50	1.01E-40	chitinase-like 1
Igfbp3	6.71	-3.17	2.14E-40	insulin-like growth factor binding protein 3
Apol11a	6.62	7.21	1.25E-34	apolipoprotein L 11a
Fpr2	4.94	3.41	5.91E-33	formyl peptide receptor 2
Nos2	5.32	4.86	4.12E-28	nitric oxide synthase 2, inducible
Hk3	7.35	2.57	8.13E-27	hexokinase 3
Abca13	3.16	4.62	1.25E-26	ATP-binding cassette, sub-family A (ABC1), member 13
F10	4.66	3.31	4.64E-26	coagulation factor X
Mt2	4.96	3.43	7.99E-26	metallothionein 2
Clec4e	4.63	4.26	1.53E-25	C-type lectin domain family 4, member e
Gbp2b	6.08	6.29	3.24E-25	guanylate binding protein 2b
Upp1	3.44	3.55	1.31E-24	uridine phosphorylase 1
Lilr4b	5.73	2.43	1.31E-24	leukocyte immunoglobulin-like receptor, subfamily B, member 4B
Gm4841	4.39	4.73	1.74E-24	predicted gene 4841
Fcgr4	6.68	3.20	2.02E-23	Fc receptor, IgG, low affinity IV
Acod1	4.70	6.45	1.84E-22	aconitate decarboxylase 1
Ms4a6d	4.41	3.06	2.28E-22	membrane-spanning 4-domains, subfamily A, member 6D
Gm45837	6.32	-1.92	3.67E-22	predicted gene 45837
Igtp	7.57	2.91	9.19E-22	interferon gamma induced GTPase
Gbp7	7.55	2.62	3.15E-21	guanylate binding protein 7

3.11.1.2 MA PLOT

Considerably more genes are found to be expressed more highly in the uninfected samples, with 8318, while only 4540 are expressed more highly in the infected samples (figure 26). Many genes are uniquely expressed in one group or the other (orange); 100 genes are uniquely expressed in the infected samples, while 284 are unique to the uninfected samples. The range of LogFC runs from -7.5 and 8.6 logFC, however, only 81 genes are expressed +/- 5 logFC.

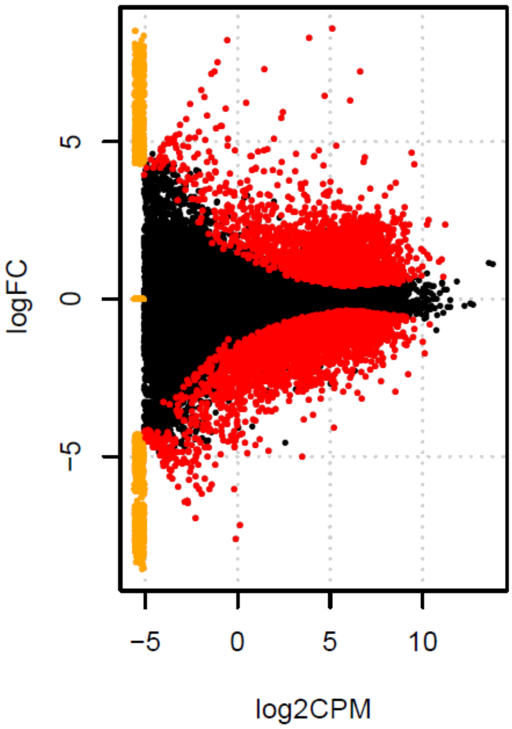


Figure 26. MA plot comparing infected and uninfected WT samples. Each dot is representative of a gene. Genes with a positive logFC value are more highly expressed in the infected samples; those with a negative logFC value are more highly expressed in the uninfected samples. Black dots are not considered differentially expressed, red dots are considered differentially expressed, and orange is representative of genes that are exclusively expressed in one group.

3.11.2 ENRICHMENT ANALYSIS

3.11.2.1 GENE SET ENRICHMENT ANALYSIS

GSEA found 40 overlaps between the database sets and the DEG list (table 30). Hallmark sets accounted for 20 of the overlaps, with the other 20 overlaps being of immunologic gene sets; no matches were detected for DNA motif sets. The average number of overlapping genes was 54, with a range of 14 to 118. The p-value average was 3.82E-09, ranging from 4.78E-08 to the extreme 2.32E-140.

A number of overlapping gene sets indicated an immune response: for example, a number of sets involving naïve vs. effector or memory cells, as well as several gene sets associated with infection responses or immunisation. Overlaps were also detected with hallmark sets involved with inflammatory response.

Table 30. Gene Set Enrichment Analysis results for genes differentially expressed between infected and uninfected Black 6 mice, ordered by FDR/q-value. P-value and set size ratios have been omitted. Only the top 20 overlaps are reported when using the Broad Institute web interface. Full results table in accompanying material A2.

Gene set description	Genes in set	k/K	FDR/q-value
Genes up-regulated in comparison of wild type CD8 effector T cells at day 6 versus those from mice deficient for TRAF6 at day 10.	200	0.59	1.38E-136
Genes up-regulated in comparison of wild type CD8 effector T cells at day 6 versus those at day 10.	200	0.535	3.15E-117
Genes up-regulated in B lymphocytes: control versus stimulated by anti-IgM for 12h.	182	0.5055	2.41E-97
Genes down-regulated in CD8 T cells after immunization: day 3 versus day 6.	200	0.47	1.00E-95
Genes up-regulated in comparison of splenic primary CD8 effector T cells at day 8 post-acute infection versus splenic secondary CD8 effector T cells at day 8 post-acute infection.	199	0.4472	4.30E-88
Genes down-regulated in natural T reg versus T conv.	180	0.4722	1.14E-86
Genes up-regulated in comparison of effector CD8 T cells versus memory CD8 T cells.	200	0.395	9.80E-73
Genes up-regulated in B lymphocytes: control versus stimulated by anti-IgM for 2h.	180	0.4222	1.21E-72
Genes down-regulated in comparison of naïve CD8 T cells versus effector CD8 T cells.	200	0.38	1.89E-68
Genes down-regulated in T reg: induced versus natural.	178	0.3989	1.18E-65
Genes up-regulated in B lymphocytes stimulated by anti-IgM for 8h: wildtype versus NFATC1 knockout.	200	0.36	8.17E-63
Genes up-regulated after poly(IC) injection: CD8A dendritic cells versus NK cells.	200	0.355	1.91E-61
Genes encoding cell cycle related targets of E2F transcription factors.	200	0.35	4.39E-60
Genes up-regulated in B lymphocytes stimulated by anti-IgM: 2h versus 12h.	172	0.3837	1.07E-59
Genes down-regulated in polymorphonuclear leukocytes (9h): S. aureus infection versus control.	200	0.345	9.28E-59
Genes up-regulated in CD4 T conv: control versus over-expression of IKZF4	173	0.3757	4.13E-58
Genes up-regulated in polarizing CD4 Th17 cells: wildtype versus RORC knockout.	161	0.3913	1.24E-57
Genes up-regulated in comparison of untreated CD25+ T effector cells at day 7 versus untreated CD25- T cells at day 7.	200	0.34	1.84E-57
Genes up-regulated in induced T reg versus T conv.	178	0.3652	3.13E-57
Genes down-regulated in comparison of untreated CD4 memory T cells from young donors versus those treated with TSST at 40 h.	200	0.335	3.50E-56

3.11.2.2 GENE ONTOLOGY ANALYSIS

GOrilla analysis of the DEG list found significantly 99 enriched GO terms, including 65 biological processes, 20 molecular functions, and 14 cellular components (table 31). P-value ranged from 1.00E-03 to 1.60E-11, with an average of 3.22E-04.

Many of the enriched GO categories listed involved the metabolism of nitrogen compounds. Additionally, terms related to nucleic acid metabolism, DNA processing and DNA replication were also common (figure 27). Use of lower p-values such as 0.01 and 0.001 found that “organinitrogen compound metabolic process” and “phosphate-containing compound metabolic process”, and related terms, were both still significant at the respective values.

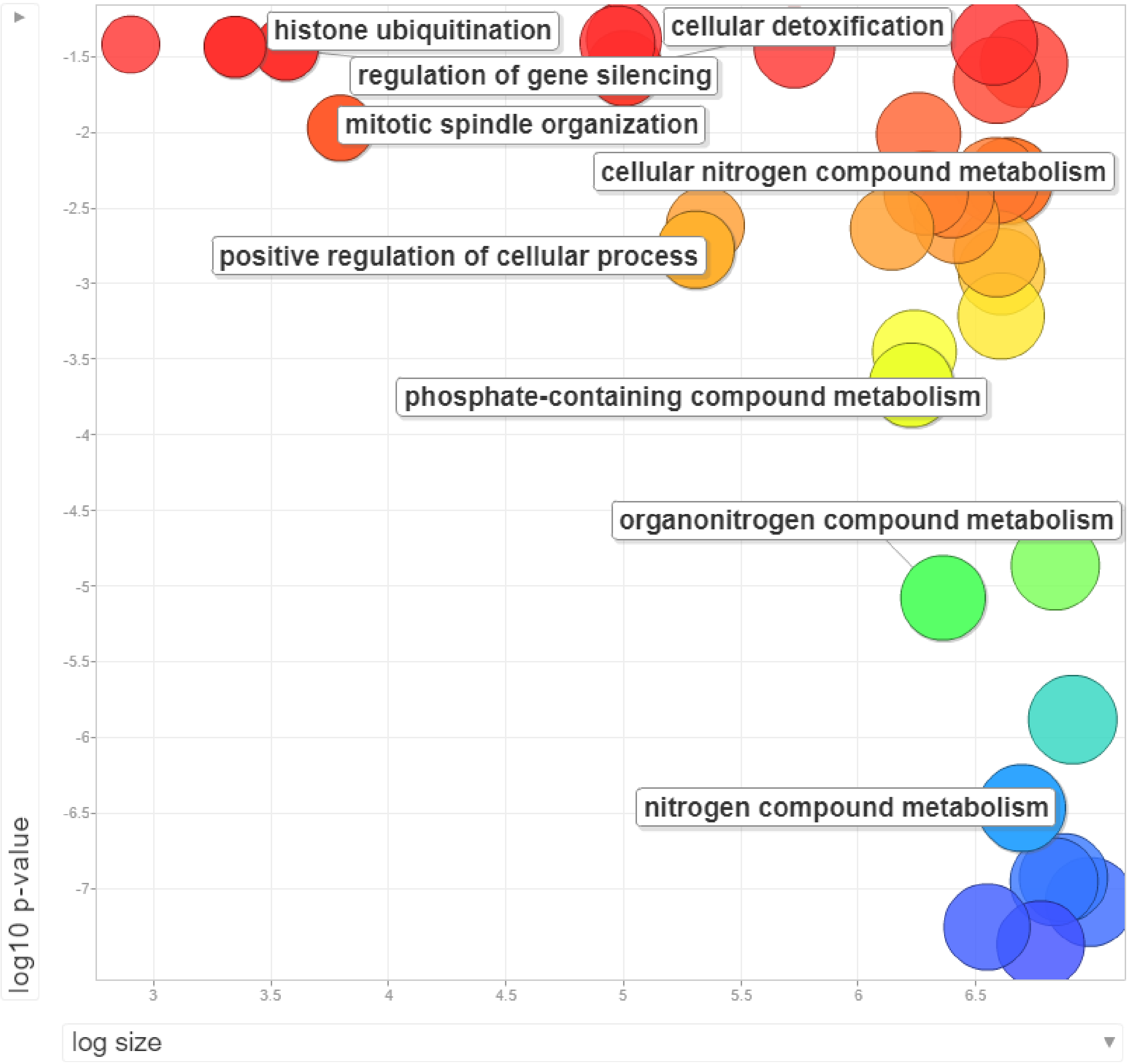


Figure 27. A REVIGO plot visualising the most enriched terms (p ≤ 0.05) in the uninfected WT vs infected WT differentially expressed gene list. Significance is shown on the y-axis in log10 p-value. The number of genes and terms collapsed into the category are displayed by both the x-axis and the bubble size.

Table 31. Gene ontology analysis results for genes differentially expressed between infected and uninfected Black 6 mice, ordered by p-value. Process = biological process, function = molecular function, component = cellular component. Full results table in accompanying material A3.

Type	Description	p-value
Process	cellular metabolic process	1.60E-11
Process	metabolic process	2.36E-11
Process	single-organism cellular process	2.42E-11
Process	single-organism process	2.46E-11
Process	organic substance metabolic process	8.51E-11
Process	nitrogen compound metabolic process	2.95E-10
Process	cellular process	1.32E-09
Process	organonitrogen compound metabolic process	9.64E-09
Process	primary metabolic process	1.77E-08
Process	phosphate-containing compound metabolic process	3.09E-07
Process	phosphorus metabolic process	5.62E-07
Process	biosynthetic process	1.05E-06
Process	single-organism metabolic process	2.24E-06
Process	organic substance biosynthetic process	3.19E-06
Process	positive regulation of cellular process	3.62E-06
Function	ion binding	6.82E-07
Function	catalytic activity	7.25E-07
Function	transferase activity	2.22E-06
Component	membrane-bounded organelle	3.84E-07
Component	organelle	5.34E-07

3.11.3 PATHVIEW

Enrichment-based PathView figures generated very closely match the GOrilla data, producing figures for categories such as ‘pyramidine metabolism’, ‘DNA replication’, and ‘cell cycle’. The coagulation and complement cascade pathway was also generated automatically (figure 28). The data resembles that of the infected WT vs. infected RAG coagulation and complement pathway. Most genes (11 of 15) are upregulated in the uninfected sample. However, in this case, in addition to the EPCR upregulated in the infected WT sample, F10 (Factor 10), CR3 and CR4 (complement receptors) are also upregulated. Relative expression of samples appears highly polarised as genes with smaller differences in fold change have been filtered out of the dataset.

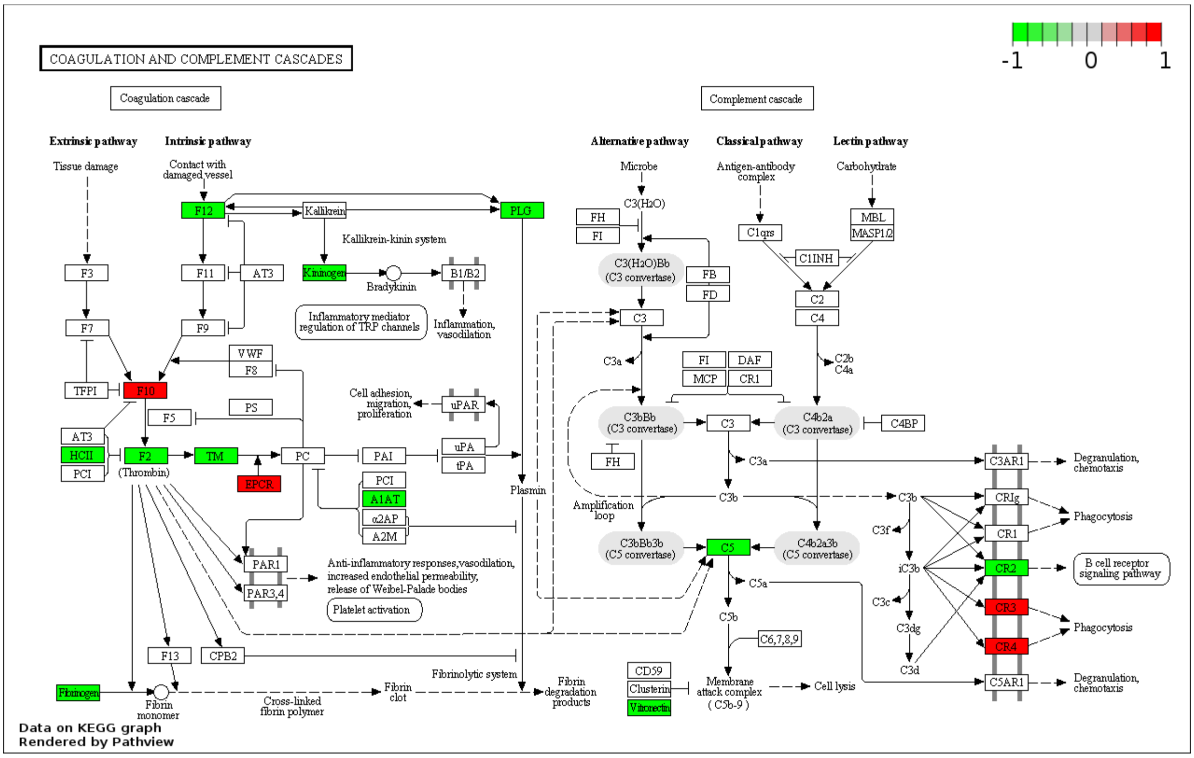


Figure 28. Coagulation and complement cascade Pathway. Each box is a gene in the pathway; Genes displayed in red are upregulated in WT infected, and green genes are upregulated in WT uninfected. White genes are not significantly differentially expressed.

Immunity related pathways reveal that unlike the T-cell and B-cell receptor pathways for the previous comparisons, very little differences in expression are detected between the WT infected and WT uninfected samples (Figures 29 and 30), with only interferon gamma being upregulated in the infected sample, and CD4/8, ITK, p38, CD19, CD21, Igα and Igβ being upregulated in the uninfected sample.

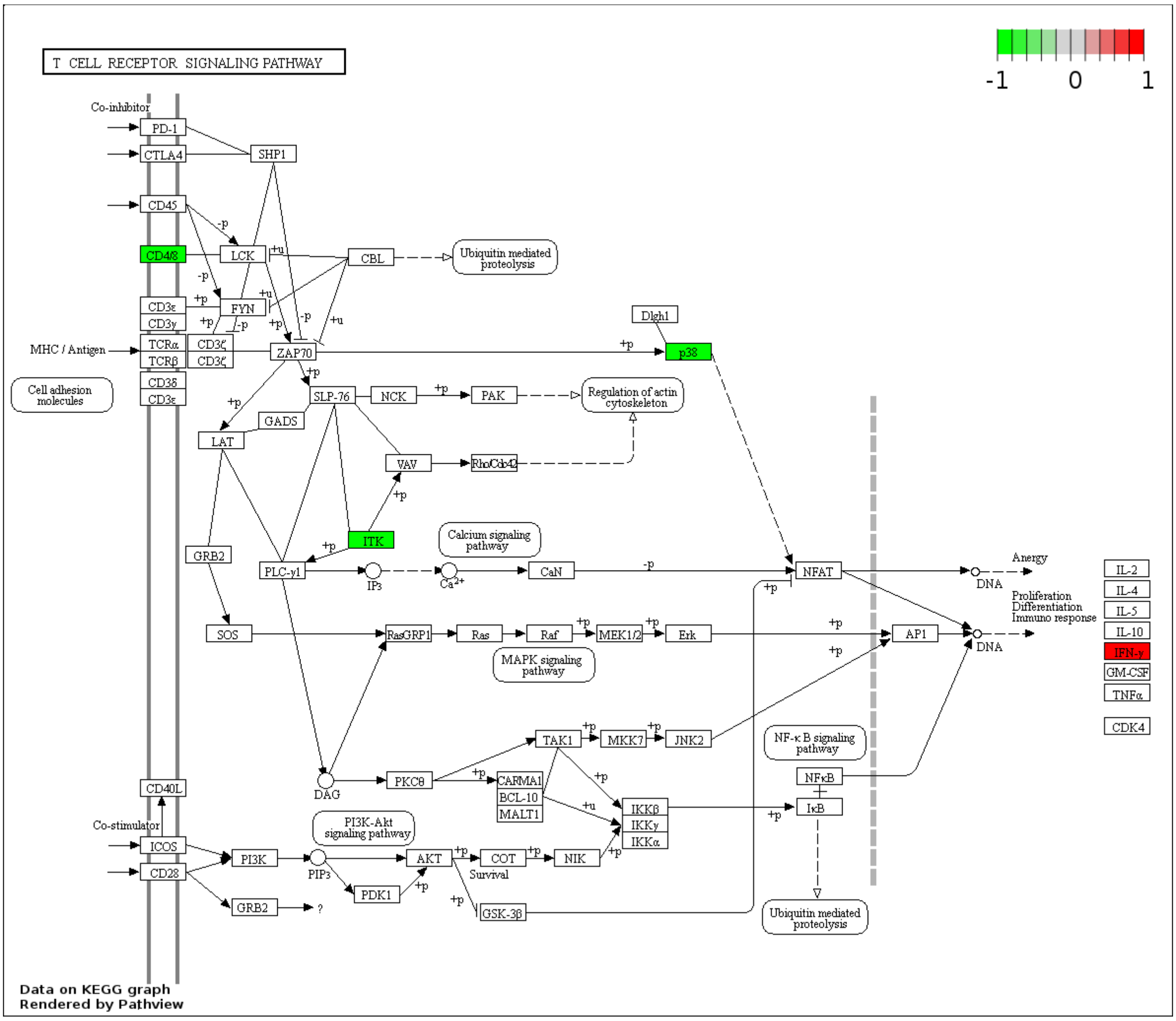


Figure 29. T-cell Receptor Signalling Pathway. Each box is a gene in the pathway; Genes displayed in red are upregulated in the infected sample, and green genes are upregulated in the uninfected sample. White genes are not significantly differentially expressed.

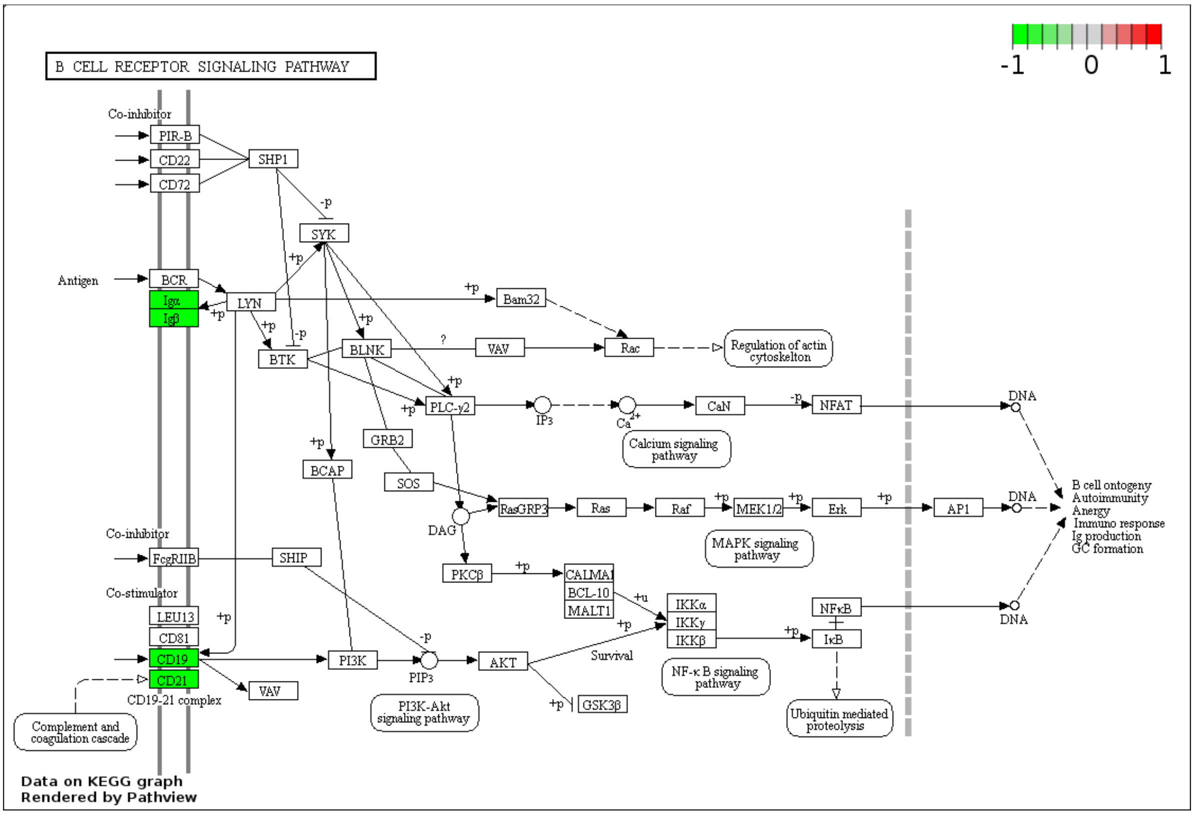


Figure 30. B-cell Receptor Signalling Pathway. Each box is a gene in the pathway; Genes displayed in red are upregulated in the infected sample, and green genes are upregulated in the uninfected sample. White genes are not significantly differentially expressed.

3.12 RAG UNINFECTED VS. RAG INFECTED

The results reported here are generated from gene expression data from RAG2 KO mice, comparing infected and uninfected.

3.12.1 EDGER ANALYSIS

3.12.1.1 DIFFERENTIAL GENE EXPRESSION

EdgeR found that 423 genes were differentially expressed between infected and uninfected RAG samples, 4.37% of the 30,976 total non-zero count genes. After logCPM and logFC criteria were applied, only 87 genes (0.28%, 79.43% were filtered out) were listed as differentially expressed (table 32).

Few of the significantly differentially expressed genes are related to immunology and infection – 5 are predicted genes, for example. Klrb1a, which is more highly expressed in the uninfected RAG mice, is associated with IL-17 producing T-cells in humans (Maggi et al. 2010).

Table 32. Top 20 significantly differentially expressed genomic features, comparing infected and uninfected RAG samples, ordered by FDR q-value. Full results table in accompanying material A1.

Gene name	logCPM	logFC	FDR	Gene description
Gm37376	8.29	-1.52	1.37E-03	predicted gene, 37376
A130071D04Rik	3.29	-1.68	1.37E-03	RIKEN cDNA A130071D04 gene
Gm16278	6.15	1.65	4.54E-03	predicted gene 16278
Nanos1	2.21	2.37	7.39E-03	nanos homolog 1 (Drosophila)
Gm11724	4.45	1.55	7.39E-03	predicted gene 11724
Gm20490	5.72	1.69	7.39E-03	predicted gene 20490
Klrb1a	2.10	-2.07	7.75E-03	killer cell lectin-like receptor subfamily B member 1A
Pif1	4.91	1.66	7.77E-03	PIF1 5'-to-3' DNA helicase
Kif18b	5.82	1.63	8.30E-03	kinesin family member 18B
Gm42456	1.45	2.66	8.30E-03	predicted gene 42456
Rbm38	7.95	1.63	9.22E-03	RNA binding motif protein 38
Troap	4.17	1.55	9.22E-03	trophinin associated protein
E2f8	6.81	1.55	1.11E-02	E2F transcription factor 8
Gfap	3.74	2.18	1.12E-02	glial fibrillary acidic protein
Cit	6.44	1.54	1.12E-02	citron
Ankrd9	4.16	1.95	1.12E-02	ankyrin repeat domain 9
Gch1	5.69	1.79	1.30E-02	GTP cyclohydrolase 1
Tspan32os	1.97	1.98	1.30E-02	tetraspanin 32, opposite strand
Slc6a9	6.30	2.00	1.44E-02	solute carrier family 6 (neurotransmitter transporter, glycine), member 9
Ccdc92b	4.34	2.45	1.48E-02	coiled-coil domain containing 92B

3.12.1.2 MA PLOT

Relatively few features are differentially expressed (red) between the infected and uninfected RAG samples (figure 31). Overall, 6454 genes are more highly expressed in the uninfected samples, while 6385 are more highly expressed in the infected samples. Most of those considered differentially expressed are of higher log2CPM rather than high logFC. Differences in logFC range from -6.8 to 6.2, with only 46 genes falling outside of +/-5 logFC.

As with other comparisons, a considerable number of genes are uniquely expressed in each group (orange). 430 genes are uniquely expressed in the infected sample, while only 59 are unique to the uninfected sample.

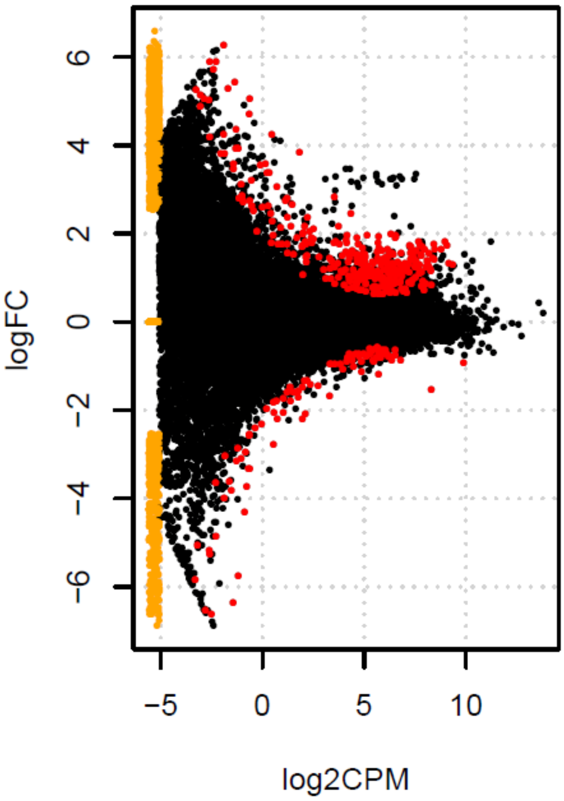


Figure 31. MA plot comparing infected and uninfected RAG samples. Each dot is representative of a gene. Genes with a positive logFC value are more highly expressed in the infected samples; those with a negative logFC value are more highly expressed in the uninfected samples. Black dots are not considered differentially expressed, red dots are considered differentially expressed, and orange is representative of genes that are exclusively expressed in one group.

3.12.2 ENRICHMENT ANALYSIS

3.12.2.1 GENE SET ENRICHMENT ANALYSIS

29 overlaps were found between the GSEA database sets and the DEG list (table 33). 2 overlaps matched hallmark gene sets, with 20 matches to immunologic gene sets and 7 matches to DNA motif sets. An average of 6 genes overlapped, ranging from 4 to 15.

The most significant overlap was for the haeme metabolism hallmark set. A number of immunologic sets were associated with infection and immune response, including cytokines IL-2, IL-6 and IL-10, as well as mentions of “acute infection” and RSV infection. Several immune cell types were also listed, including T-regs, T-convs, T-helpers, B lymphocytes, CD4 and CD8 cells, NKs, DCs, and PBMCs.

Table 33. Gene Set Enrichment Analysis results for genes differentially expressed between infected and uninfected RAG2 KO mice, ordered by FDR/q-value. P-value and overlap size have been omitted. Only the top 20 overlaps are reported when using the Broad Institute web interface. Full results table in accompanying material A2.

Gene set description	Gene in set	overlapping genes / set size	FDR/q-value
Genes involved in metabolism of haeme (a cofactor consisting of iron and porphyrin) and erythroblast differentiation.	200	0.065	1.36E-16
Genes down-regulated in CD4 T conv: control versus over-expression of GATA1 and FOXP3	145	0.0483	5.28E-07
Genes having at least one occurrence of the highly conserved motif M169 TTTNNANAGCYR in the region spanning up to 4 kb around their transcription start sites.	2274	0.0066	9.59E-06
Genes having at least one occurrence of the highly conserved motif M96 YGCANTGCR in the region spanning up to 4 kb around their transcription start sites.	1294	0.0093	9.59E-06
Genes down-regulated in CD8 T cells: control versus primary acute viral infection.	196	0.0306	5.64E-05
Genes up-regulated in macrophages in response to LPS: naïve versus tolerant.	199	0.0302	5.64E-05
Genes up-regulated in double positive thymocytes stimulated by anti-CD3: ELK4 knockout versus ELK1 and ELK4 knockout.	200	0.03	5.64E-05
Genes up-regulated in induced T reg versus T conv.	178	0.0281	9.16E-04
Genes up-regulated in comparison of wild type CD8 effector T cells at day 6 versus those from mice deficient for TRAF6 at day 10.	200	0.025	1.18E-03
Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) from healthy donors versus PBMCs from infant with acute RSV infection.	200	0.025	1.18E-03
Genes down-regulated in CD8 T cells 3 days after immunization: control versus IL2 treatment.	200	0.025	1.18E-03
Genes having at least one occurrence of the transcription factor binding site V$E2F_Q3 (v7.4 TRANSFAC) in the regions spanning up to 4 kb around their transcription starting sites.	227	0.022	2.01E-03
Genes having at least one occurrence of the highly conserved motif M174 WTGAAAT in the region spanning up to 4 kb around their transcription start sites.	924	0.0087	2.19E-03
Genes having at least one occurrence of the transcription factor binding site V$E2F1_Q3 (v7.4 TRANSFAC) in the regions spanning up to 4 kb around their transcription starting sites.	244	0.0205	2.44E-03
Genes up-regulated in polarizing CD4 Th17 cells: wildtype versus RORC knockout.	161	0.0248	9.37E-03
Genes having at least one occurrence of the highly conserved motif M72 TTCYRGAA in the region spanning up to 4 kb around their transcription start sites.	1232	0.0065	9.37E-03
Genes down-regulated in B lymphocytes stimulated by anti-IgM: 2h versus 12h.	181	0.0221	9.37E-03
Genes up-regulated in IL10 knockout macrophages stimulated by LPS versus those also stimulated by IL10	181	0.0221	9.37E-03
Genes having at least one occurrence of the highly conserved motif M7 TGANTCA in the region spanning up to 4 kb around their transcription start sites	2485	0.0044	9.37E-03
Genes up-regulated in B lymphocytes: control versus stimulated by anti-IgM for 12h.	182	0.022	9.37E-03

3.12.2.2 GENE ONTOLOGY ANALYSIS

The differentially expressed gene list was run through GOrilla to check for term enrichment. No significant ($p < 0.05$) GO term enrichment was detected for biological process, molecular function, or cellular component.

Due to the lack of enriched GO terms, REVIGO plots could not be generated.

3.12.3 PATHVIEW

Insufficient data was available to generate PathView figures. No output was produced for any of the GOrilla-suggested pathways, nor any immunity-related pathway. Figures generated from enrichment data only contained information about a single gene.

3.13 INOCULUM PARASITES VS. RAG-DERIVED PARASITES

3.13.1 EDGER ANALYSIS

3.13.1.1 DIFFERENTIAL GENE EXPRESSION

When comparing the gene expression of *Leishmania* in the inoculum and those in RAG mice, edgeR found that 3,813 genes were differentially expressed, 47.27% of a possible 8,066. 89 genes were found to be differentially expressed after the application of logCPM and logFC criteria, only 1.1% (97.67% failed to pass the logCPM and logFC thresholds) of the total (table 34). Of the 3,813 genes originally listed as differentially expressed, 2,911 were filtered by the logFC criteria alone.

Over three quarters (77.5%) of the differentially expressed genes are either entirely hypothetical or putative (Supplementary data A1). Of the 89 DE genes, 41 have putative function (46.1%) and 28 are labelled as hypothetical (31.5%). The majority of those labelled as hypothetical (75%) had no non-hypothetical match in either the NCBI BLAST database nor the TriTrypDB. The remaining 7 hypothetical genes had matches in either other *Leishmania* species or other trypanosomatids in the TriTrypDB.

Amastin-related genes are generally more highly expressed in RAG-derived parasites. Two histone proteins, H2B and H3, are also more highly upregulated in the RAG-derived parasites, contrasting with the upregulation of two nucleases in inoculum parasites (see table 34 and accompanying material A1).

Table 34. Top 20 significantly differentially expressed genes, comparing the inoculum parasites and parasites derived from RAG mice, ordered by FDR q-value. Full results table in accompanying material A1.

Gene ID	logCPM	logFC	FDR	Gene Description
LdBPK_150660.1	8.31	-2.11	5.73E-119	hypothetical protein, unknown function
LdBPK_341150.1	9.45	-1.67	1.03E-113	amastin-like surface protein, putative
LdBPK_251160.1	8.27	-1.89	9.90E-94	aldehyde dehydrogenase, mitochondrial precursor
LdBPK_201670.1	8.24	-1.53	4.61E-79	hypothetical protein, conserved
LdBPK_251970.1	7.48	-1.85	1.69E-75	hypothetical protein, conserved
LdBPK_342660.1	8.23	-1.97	2.47E-73	amastin-like surface protein, putative
LdBPK_282530.1	7.96	-1.63	7.63E-71	serine hydroxymethyltransferase (SHMT-L)
LdBPK_120350.1	7.37	1.91	2.05E-68	3'-nucleotidase/nuclease, putative
LdBPK_171320.1	8.16	-1.70	1.02E-67	histone H2B
LdBPK_080720.1	10.61	-2.47	6.70E-61	amastin-like protein
LdBPK_292010.1	6.67	-2.02	1.48E-54	DnaJ domain containing protein, putative
LdBPK_241150.1	7.61	-1.51	1.77E-49	hypothetical protein, conserved
LdBPK_101070.1	8.61	-1.61	9.26E-49	histone H3
LdBPK_360500.1	7.32	-1.50	1.42E-48	dihydrouridine synthase domain protein-like protein
LdBPK_366980.1	7.20	-1.57	6.80E-48	cytochrome b5-like Heme/Steroid binding domain containing protein, putative
LdBPK_030370.1	6.76	-2.16	2.31E-47	hypothetical protein, conserved
LdBPK_260040.1	7.25	-1.93	4.59E-45	glycine dehydrogenase, putative
LdBPK_333390.1	6.95	-1.86	1.09E-42	h1 histone-like protein
LdBPK_343780.1	5.98	3.41	3.87E-39	hypothetical protein, conserved
LdBPK_312380.1	6.14	2.34	1.47E-38	3'-nucleotidase/nuclease precursor, putative

3.13.1.2 MA PLOTS

The MA plot showing the expression differences between parasites from the inoculum and RAG-derived parasites (figure 32) displays relatively fewer genes when compared with mouse samples, given that the genome of *Leishmania* is smaller than that of the mouse, and additionally the genome is less well-annotated. In total, 1893 genes are expressed more highly in the inoculum, and 1920 more highly in the RAG-derived samples. Only 3 genes are uniquely expressed in the inoculum, while 6 are unique to the RAG-derived parasites. Only 38 genes are expressed more extremely than +/- 2 logFC, but overall expression ranges from -6.2 to 7.3 logFC. The ‘stripes’ of genes that appear between -5 and 0 log2CPM occur when one sample has a very low read count, but the other does not. The fold change data from these genes are not necessarily invalid, but exact values must be treated with scepticism as limited reads in one sample may cause irregular read count scaling and uneven normalisation.

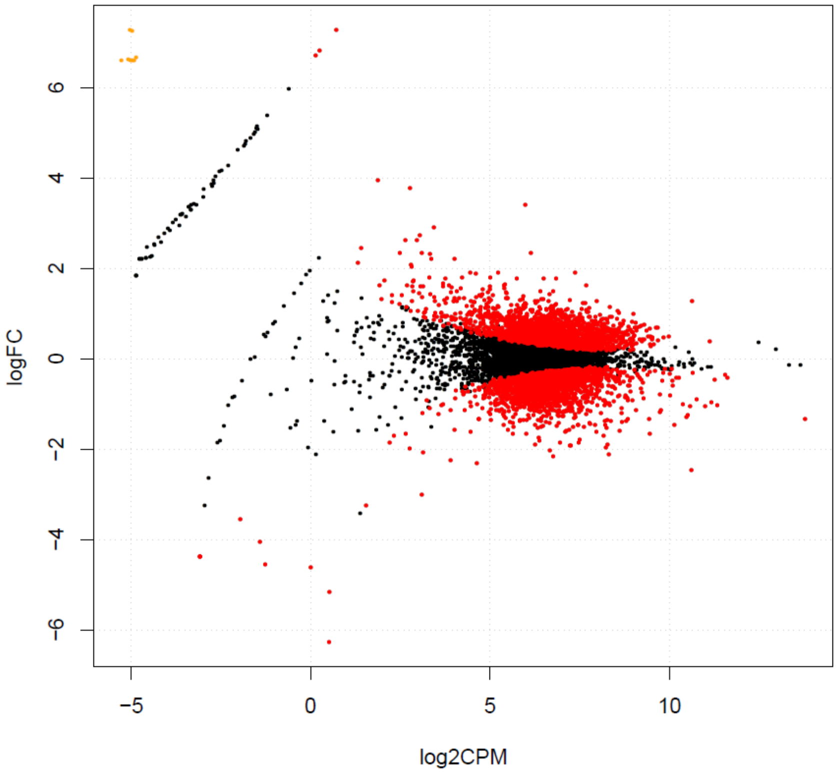


Figure 32. MA plot comparing inoculum and RAG mouse-derived parasites. Each dot is representative of a gene. Genes with a positive logFC value are more highly expressed in the inoculum samples; those with a negative logFC value are more highly expressed in the RAG samples. Black dots are not considered differentially expressed, red dots are considered differentially expressed, and orange is representative of genes that are exclusively expressed in one group.

3.13.2 ENRICHMENT ANALYSIS

3.13.2.1 GENE SET ENRICHMENT ANALYSIS

The gene sets within the GSEA database are built on human and mouse data, and therefore lists of *Leishmania* genes are inappropriate for GSEA.

3.13.2.2 GENE ONTOLOGY ANALYSIS

Instead of GOrilla, which is unable to generated GO terms from *Leishmania* data, gene ontology analysis was instead performed using the gene ontology function of TriTrypDB. A total of 70 terms were found to be significantly enriched (table 35); 15 cellular components, 23 molecular functions and 32 biological processes. The average p-value was 1.91E-02, ranging from 4.23E-27 to 4.79E-02.

Nucleobase-related, DNA packaging-related and nucleic acid metabolism-related categories appear particularly abundant (figure 33). Additionally, sucrose and disaccharide metabolism, hydrolase and endonuclease activity, and nitrogen compound metabolism categories are also prevalent. Use of 0.01 as a p-value cut off resulted in a single GO term: “nucleobase-containing compound catabolism”. “Cellular component” was the only term reported with a threshold of 0.001 or less.

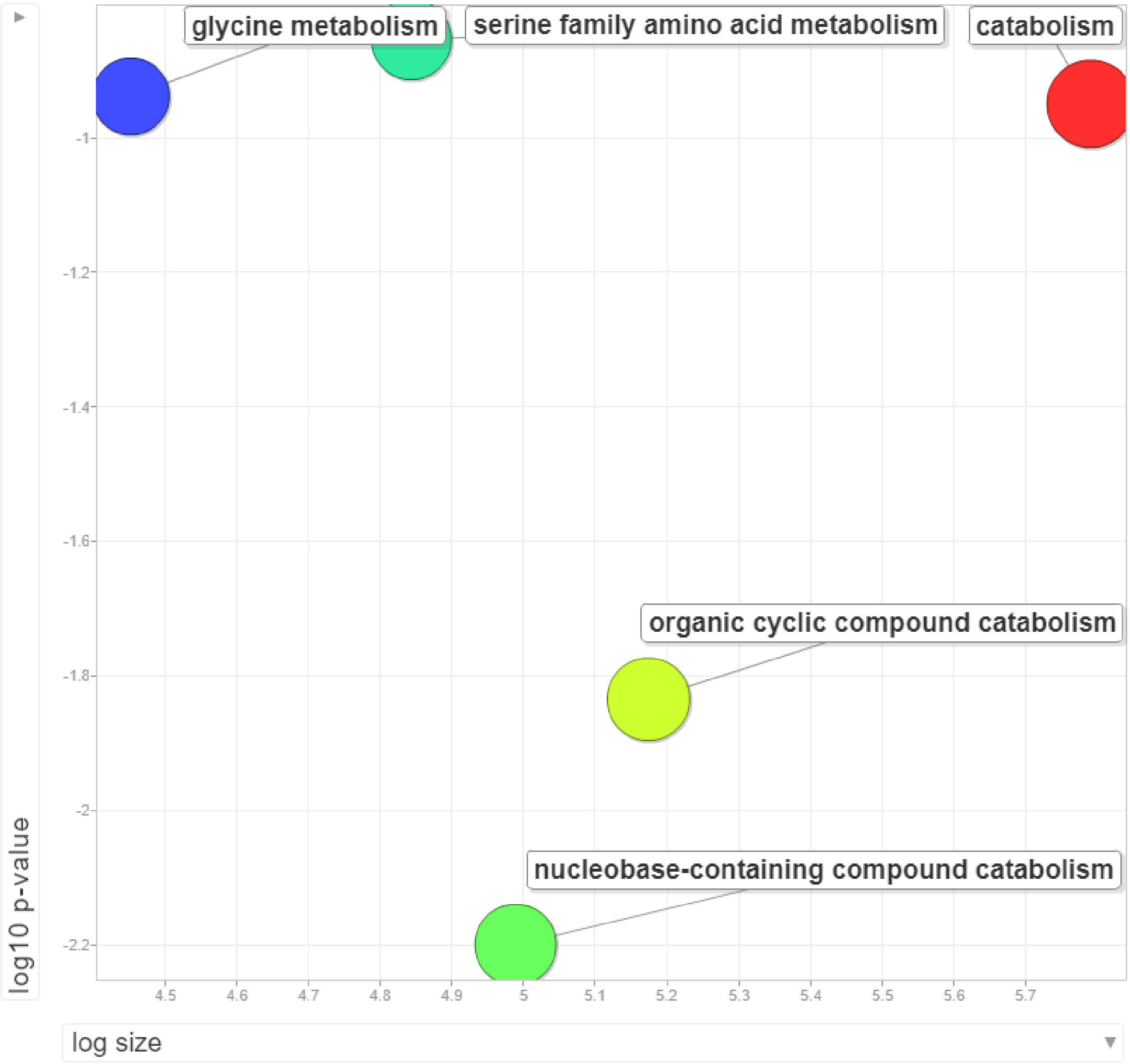


Figure 33. A REVIGO plot visualising the most enriched terms (p ≤ 0.05) in the inoculum parasites vs RAG parasites differentially expressed gene list. Significance is shown on the y-axis in log10 p-value. The number of genes and terms collapsed into the category are displayed by both the x-axis and the bubble size.

Table 35. Top 20 Gene Ontology results for the genes differentially expressed between inoculum parasites and RAG parasites. Process = biological process, function = molecular function, component = cellular component. Full results table in accompanying material A2.

Type	Description	p-value
Process	nucleobase-containing compound catabolic process	0.0001971
Process	organic cyclic compound catabolic process	0.0004568
Process	cellular nitrogen compound catabolic process	0.0004568
Process	aromatic compound catabolic process	0.0004568
Process	heterocycle catabolic process	0.0004568
Process	cellular catabolic process	0.0012186
Process	DNA catabolic process	0.0017098
Process	organic substance catabolic process	0.0032488
Process	catabolic process	0.0035108
Process	glycine metabolic process	0.0036003
Function	protein heterodimerization activity	0.000871
Function	nucleic acid binding	0.0014595
Function	protein dimerization activity	0.002052
Component	cellular component	4.23E-27
Component	chromosomal part	0.0002664
Component	nucleosome	0.0006427
Component	chromosome	0.0006509
Component	protein-DNA complex	0.0007514
Component	chromatin	0.0007514
Component	DNA packaging complex	0.0007514

3.13.3 PATHVIEW

PathView builds figures from KEGG data. However, relatively limited KEGG data is available for *Leishmania* and as a result, no figures were able to be generated.

3.14 SAMPLE CROSS-COMPARISON AND EXPRESSION PLOTS

3.14.1 MOUSE TRANSCRIPTOME DATA

3.14.1.1 SAMPLE SIMILARITY

Differential gene expression data produced by edgeR was transformed for the generation of correlation heatmaps and PCA plots.

Figure 34 shows a sample correlation matrix produced from mouse transcriptome data. Both the wild type infected and the wild type uninfected (control) samples appear highly similar (dark red/dark orange) to other samples from the same background and infection status. WT infected samples were between 97.8% and 99.0% similar, with an average of 98.5%; WT uninfected samples were all found to be 99.0% similar to each other. RAG sample transcriptomes appear less closely related (pale orange/yellow) than the wild type, with infected RAG samples ranging from 96.6% to 98.2% in similarity, and uninfected RAG samples ranging from 96.1% to 97.2% . The most dissimilar transcriptomes are displayed in yellow, showing that the wild type infected transcriptome is least similar to the RAG control transcriptome (average similarity 95.5%).

However, the scale of the difference between transcriptomes is fairly minor, as the least similar transcriptomes still show 0.942 similarity out of a possible 1.000.

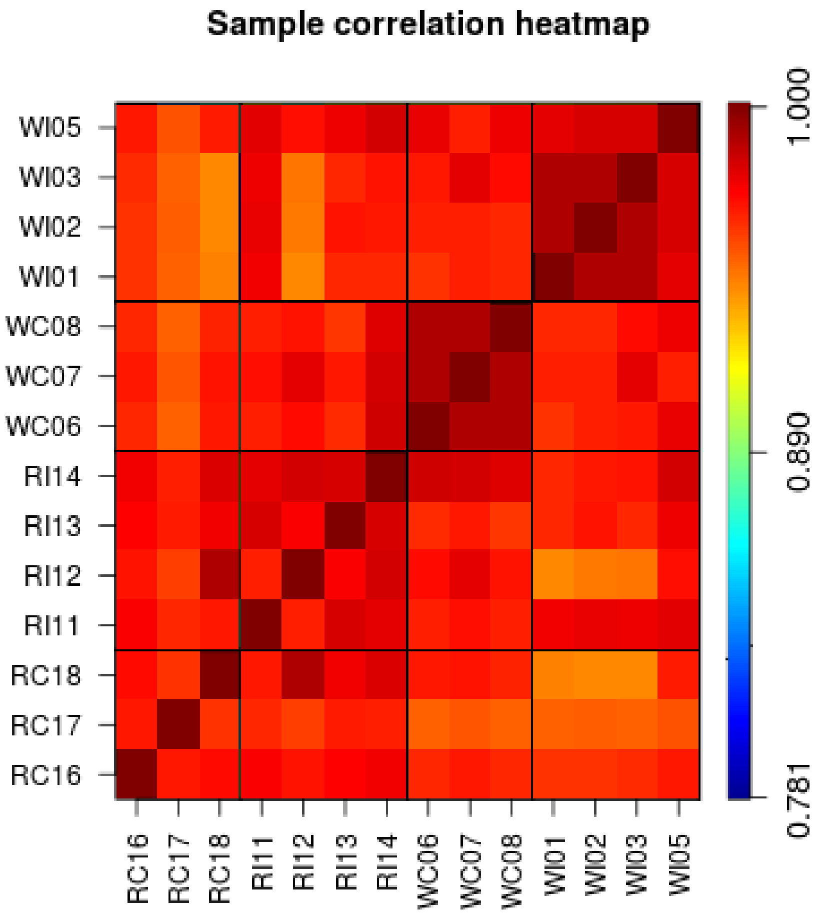


Figure 34. Mouse transcriptome Euclidean distance heatmap. Samples that are more similar are indicated with dark red; less similar samples are inicated with dark blue.

Two PCA plots were also generated from the mouse transcriptome data, comparing the variance of the samples the 1st, 2nd and 3rd components. The 1st and 2nd components, as seen in figure 35, do not appear to differentiate the sample clusters much, indicating that overall gene expression in the samples is similar. However, figure 36 shows that the 3rd component accounts for a large amount of variance between samples, clearly distinguishing the WT infected and uninfected samples from each other and the RAG samples. RAG infected and uninfected groups are not so easily differentiated, but this is somewhat expected given that, relative to the WT, few genes were found to be differentially expressed between the groups.

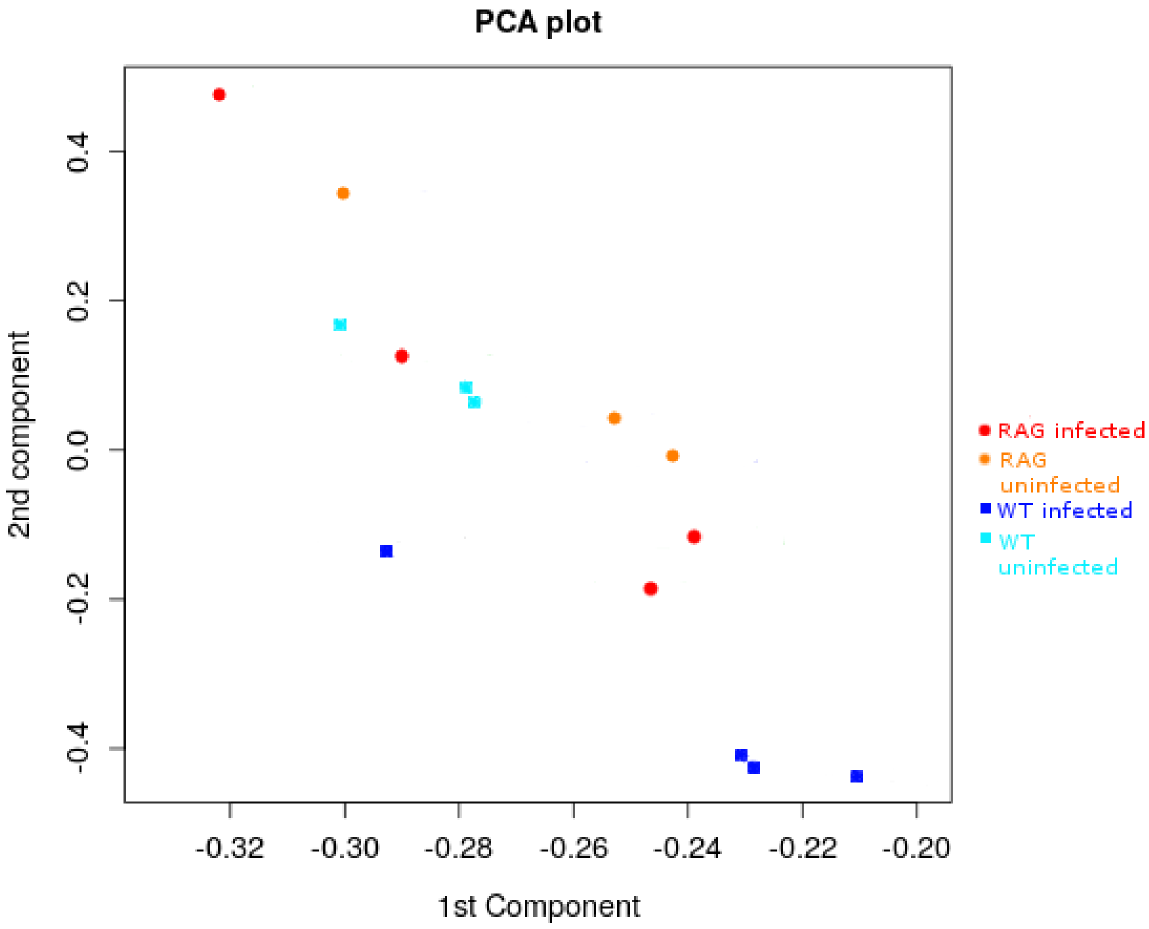


Figure 35. 1st & 2nd PCA components for mouse transcriptome data. WT uninfected samples are represented in cyan, WT infected samples are represented in blue, RAG uninfected samples are represented in orange and RAG infected samples are in red.

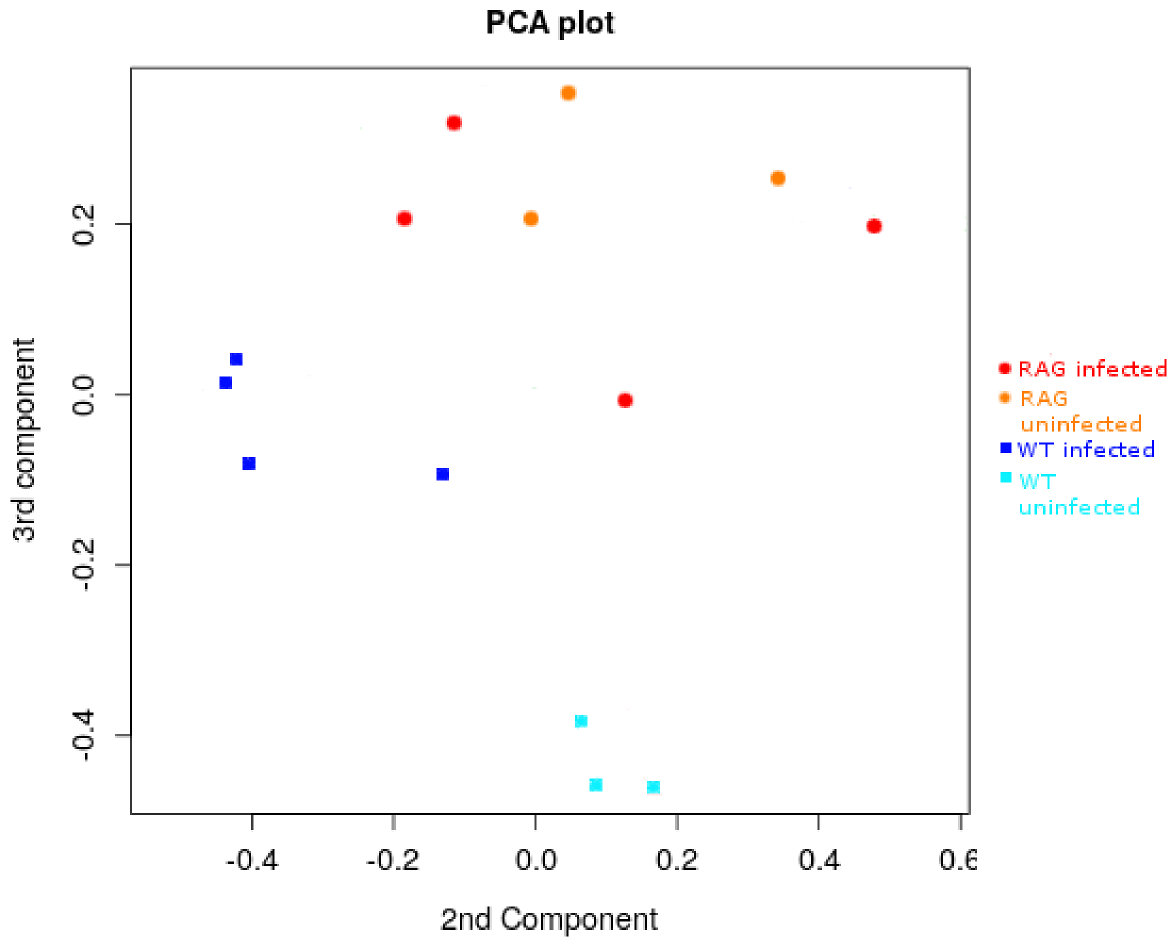


Figure 36. 2nd & 3rd PCA components for mouse transcriptome data. WT uninfected samples are represented in cyan, WT infected samples are represented in blue, RAG uninfected samples are represented in orange and RAG infected samples are in red.

3.14.1.2 GENE PANELS

Heatmaps displaying fold-changes in expression of key immunology-related genes between mouse samples were generated from edgeR data; all genes listed were significantly differentially expressed between samples ($p \leq 0.05$). Panels of genes were produced for CD (cluster of differentiation) markers genes, cytokines, and chemokines. logFC values are transformed into Z-score by the heatmap.2 function, which smooths and scales the logFC values so that they can more easily be compared between rows.

CD marker genes are used to indicate cell lineage and cell type, for example, distinguishing different types of T-cells or B-cells. Figure 37 shows the panel produced for CD marker genes. Genes related to CD 209 appear significantly more upregulated in all samples when compared with WT infected samples. Colmenares et al. 2002, found that CD 209 acts as a receptor for *Leishmania* amastigotes in human dendritic cells. CD5, CD6, CD79 and CD3-related genes all appear more highly expressed in WT infected samples relative to both RAG infected and uninfected mice. Depletion of CD5+ B-cells was found to have no effect on the outcome of *L. major* infections in different mice strains (Babai et al. 1999). CD79+ B-lymphocytes were found to be differentially trafficked to the brain in dogs infected with *Leishmania chagasi* when compared with uninfected dogs (Melo et al. 2009).

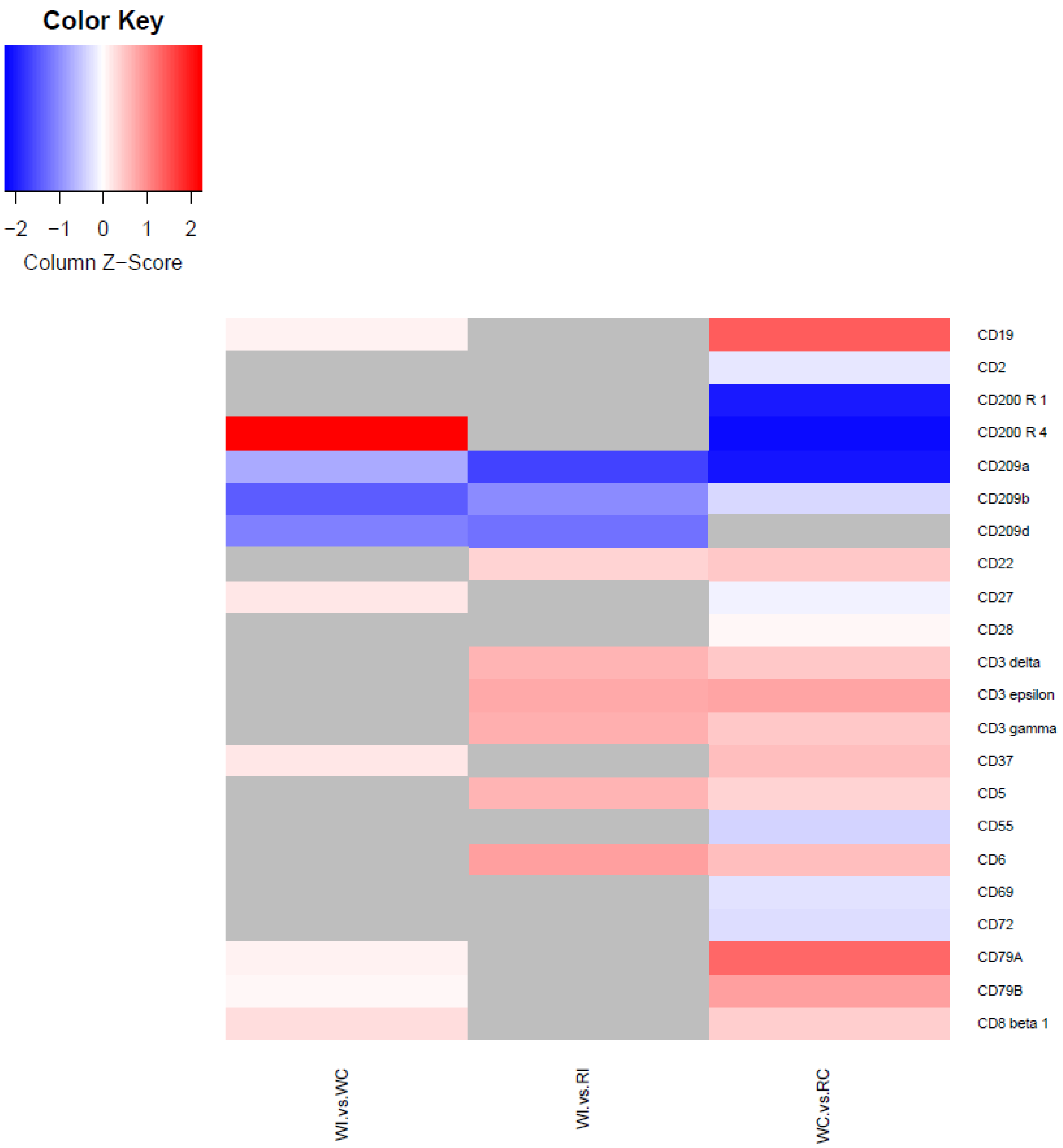


Figure 37. A gene panel displaying the fold changes in key CD marker genes between samples. Genes highly upregulated in the first listed sample (L-R: WT infected, WT infected, WT uninfected) are represented in red, while those highly upregulated in the second listed sample (L-R: WT uninfected, RAG infected, RAG uninfected) are represented in blue. Comparisons for which the genes are not differentially expressed are represented with dark grey.

Differences in chemokine expression are displayed below in figure 38. Both RAG infected and uninfected samples appear to highly express C-C receptor 3 when compared with WT infected mice. C-X-C receptor 5 is more highly expressed in WT infected samples than in either RAG sample. A number of C-C and C-X-C receptors, such as C-X-C receptors 5 and 2, and C-C receptors 9, 7, and 5, show differing expression between WT infected and uninfected samples, with no overall pattern to expression. Sato et al. 1999 found that mice deficient in C-C R 2 had a reduced interferon gamma response when infected with *L. donovani*. Additional work on relative *Trypanosoma cruzi* found that C-C R 5 is essential for the control of parasite replication and tissue inflammation (Hardison et al. 2006). C-C R 7 was found to be upregulated in dendritic cells in upon interaction with *Leishmania major* (Steigerwald and Moll 2005).

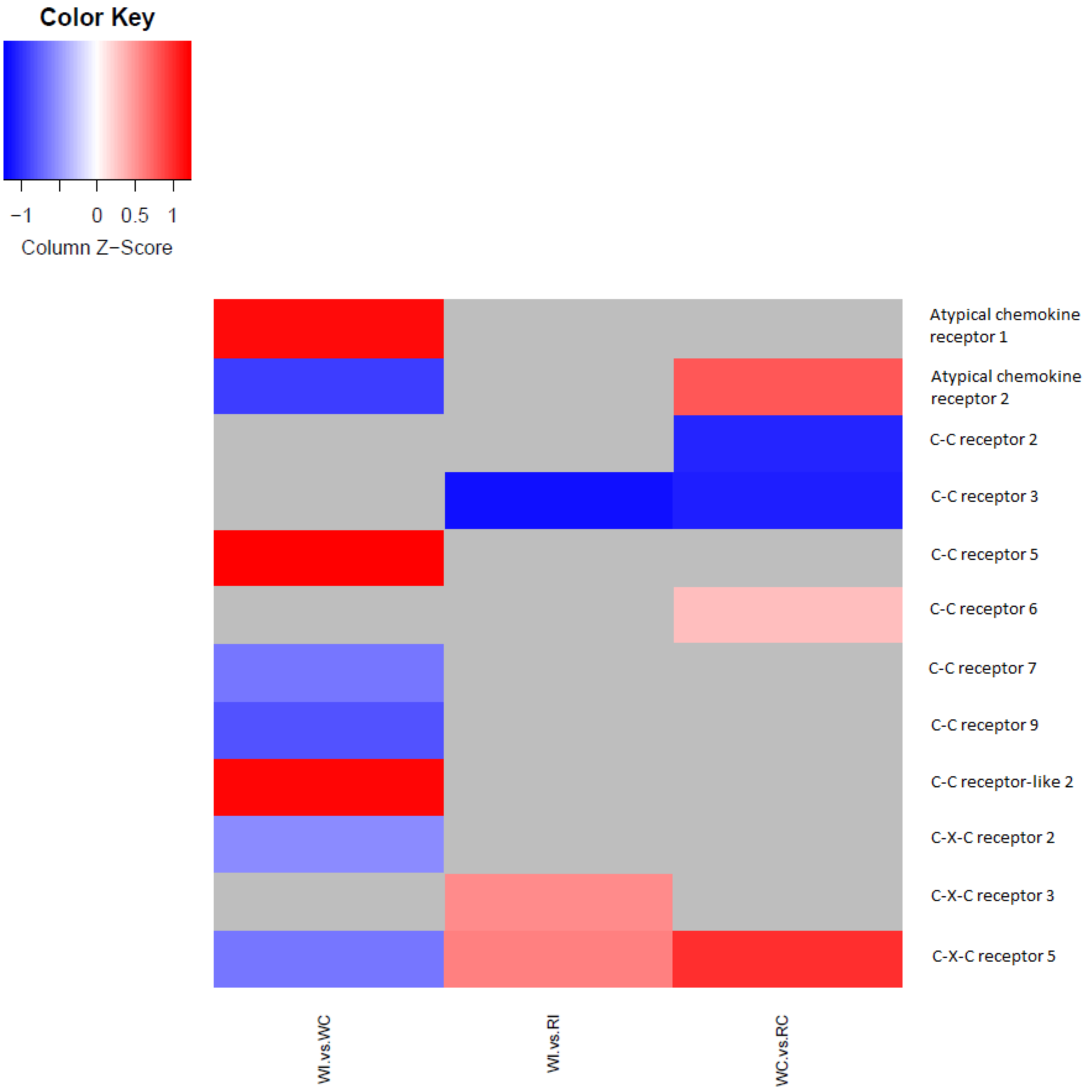


Figure 38. A gene panel displaying the fold changes in key chemokine genes between samples. Genes highly upregulated in the first listed sample (L-R: WT infected, WT infected, WT uninfected) are represented in red, while those highly upregulated in the second listed sample (L-R: WT uninfected, RAG infected, RAG uninfected) are represented in blue. Comparisons for which the genes are not differentially expressed are represented with dark grey.

The relative differences in expression of cytokines are shown in figure 39 (below). Receptors of the TNF-alpha superfamily are highly expressed in WT samples when compared with RAG, and in WT infected samples when compared with WT uninfected samples. In *L. donovani* infections, TNF-alpha is known to be critically important in parasite control, due to its involvement in leukocyte recruitment to the liver (Engwerda et al. 2004). Interferon gamma is more highly expressed in WT infected samples than in RAG infected samples, and in WT infected samples than WT uninfected samples. The role of interferon gamma in leishmaniasis is complex; early formative studies in mice, using *L. major*, found that parasite control was correlated with the presence of interferon gamma, and that the absence was associated with disease progression (Kima and Soong 2013).

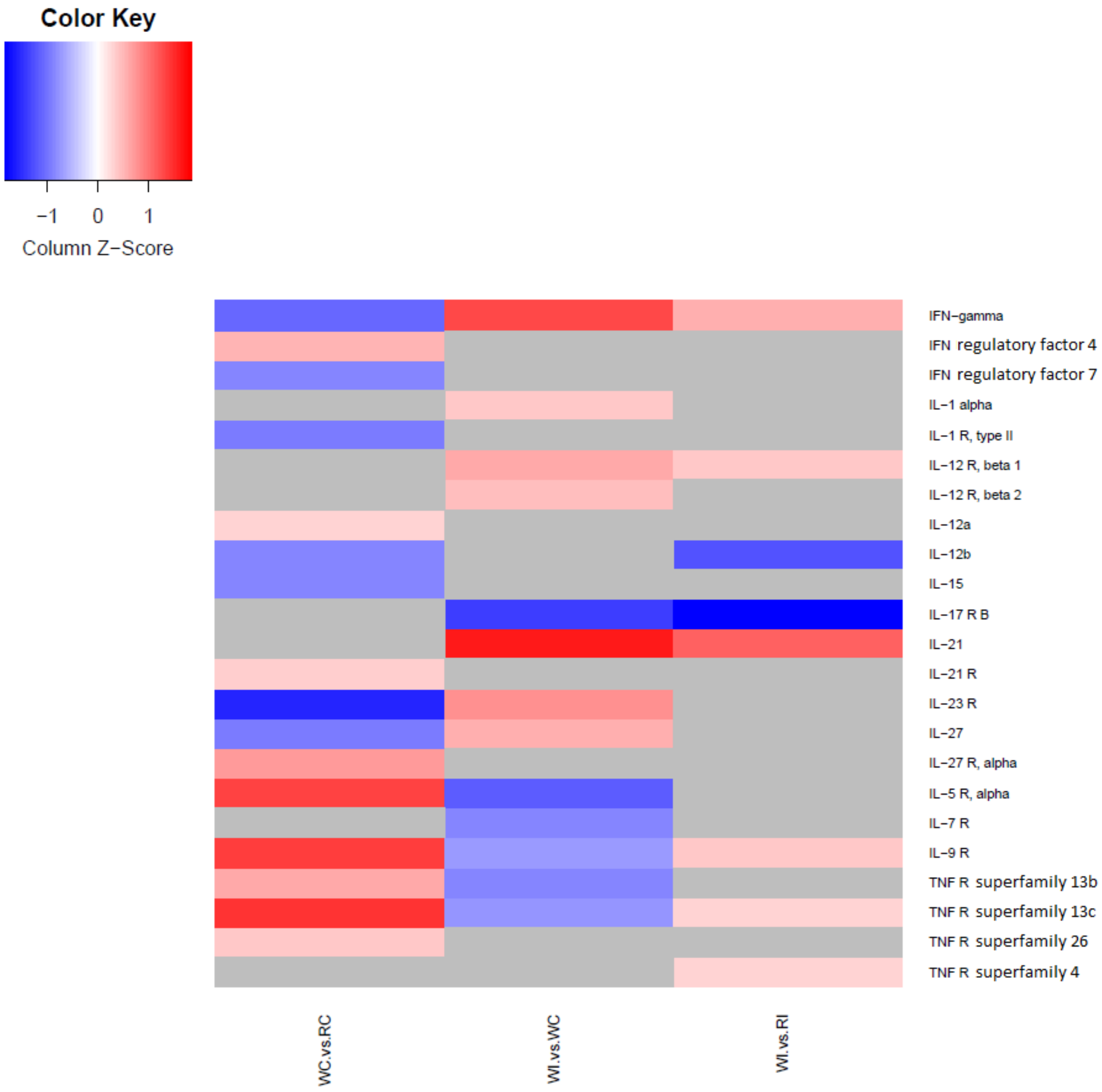


Figure 39. A gene panel displaying the fold changes in key cytokine genes between samples. Genes highly upregulated in the first listed sample (L-R: WT infected, WT infected, WT uninfected) are represented in red, while those highly upregulated in the second listed sample (L-R: WT uninfected, RAG infected, RAG uninfected) are represented in blue. Comparisons for which the genes are not differentially expressed are represented with dark grey.

3.14.2 *LEISHMANIA* TRANSCRIPTOME DATA

3.14.2.1 SAMPLE SIMILARITY

The correlation heatmap of *Leishmania* transcriptome data can be seen in figure 40, comparing transcriptomes from inoculum parasites and RAG-derived parasites. Reassuringly, the technical replicates (PI01 – PI03) are almost identical, showing an extreme degree of similarity, ranging from 97.9% to 99.5%. RAG-derived parasites show less similarity to each other than the technical replicates (average similarity 87.0%), but appear on average more similar to each other than to inoculum samples (average similarity 84.2%). The *Leishmania* transcriptomes are considerably less similar to each other than the mouse samples are to other mouse samples, with the lowest Euclidean distance being 0.781 compared with the mouse distance of 0.942. One potential source of variation between each of the RAG-derived samples is that the parasites may be affected by differences in the kinetics of disease progression from mouse to mouse.

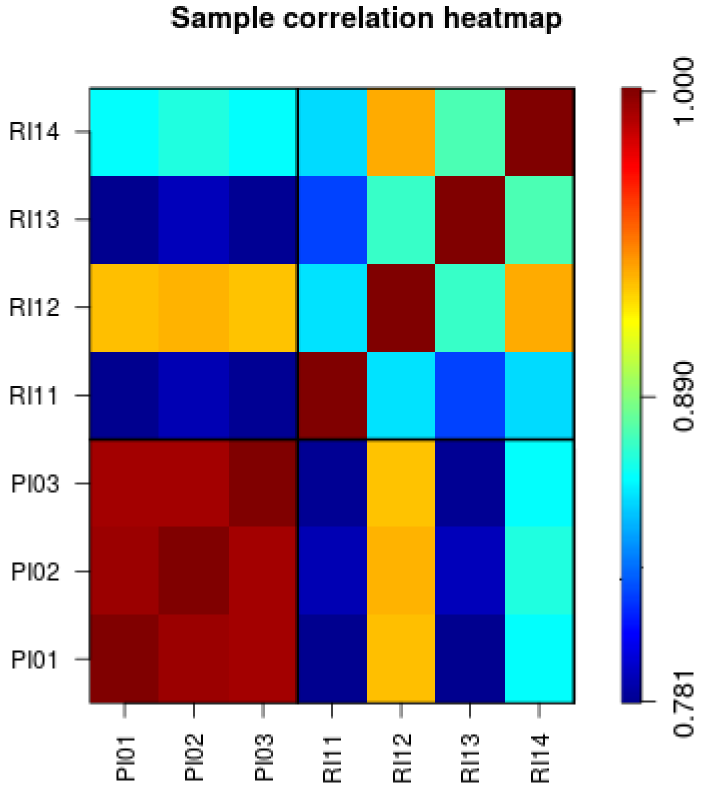


Figure 40. *Leishmania* transcriptome Euclidean distance heatmap. Samples that are more similar are indicated with dark red; less similar samples are indicated with dark blue. RI = RAG-infected derived parasites; PI = parasite inoculum derived parasites.

Figures 41 and 42 show the PCA plots for the *Leishmania* transcriptomes. Siginifcant differentiation between clusters of samples can be seen on plots of both 1st and 2nd, and 2nd and 3rd components, indicating the expression profiles of the two groups are dissimilar.

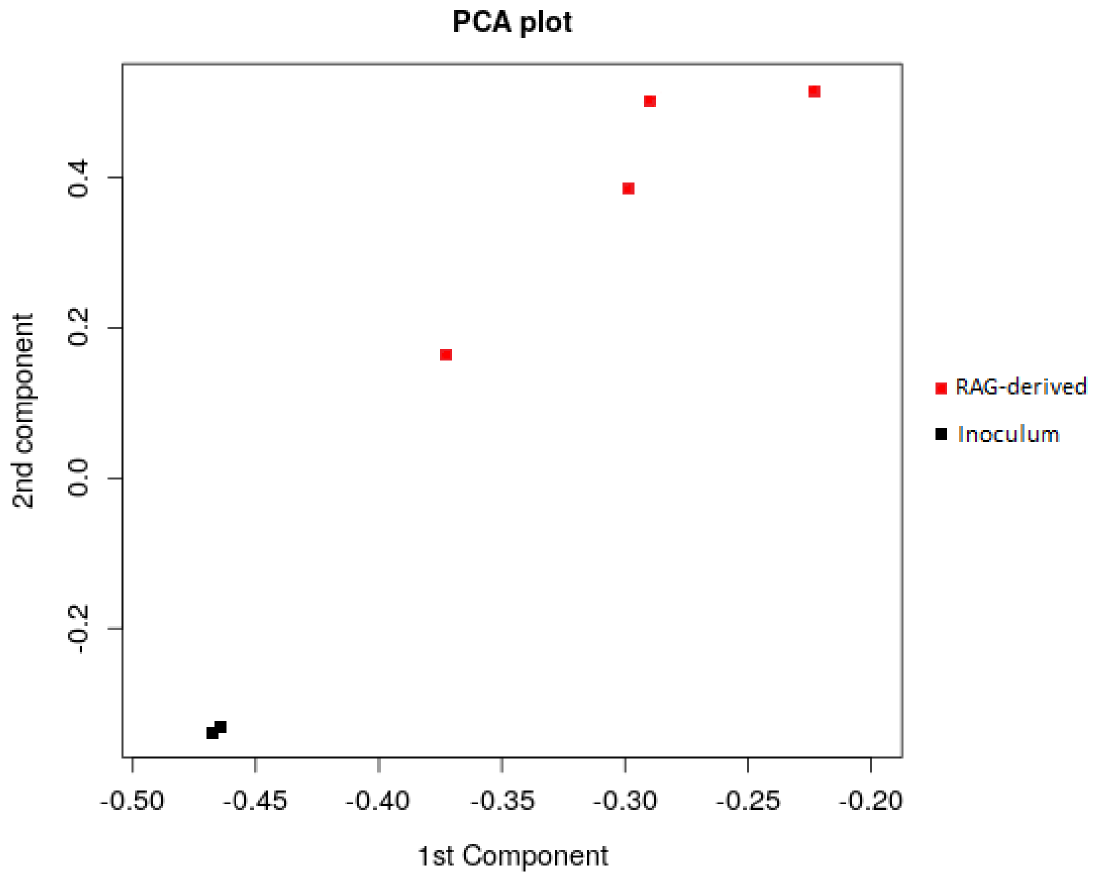


Figure 41. 1st & 2nd PCA components for *Leishmania* transcriptome data. Inoculum-derived samples are represented in black, and RAG-derived samples are represented in red.

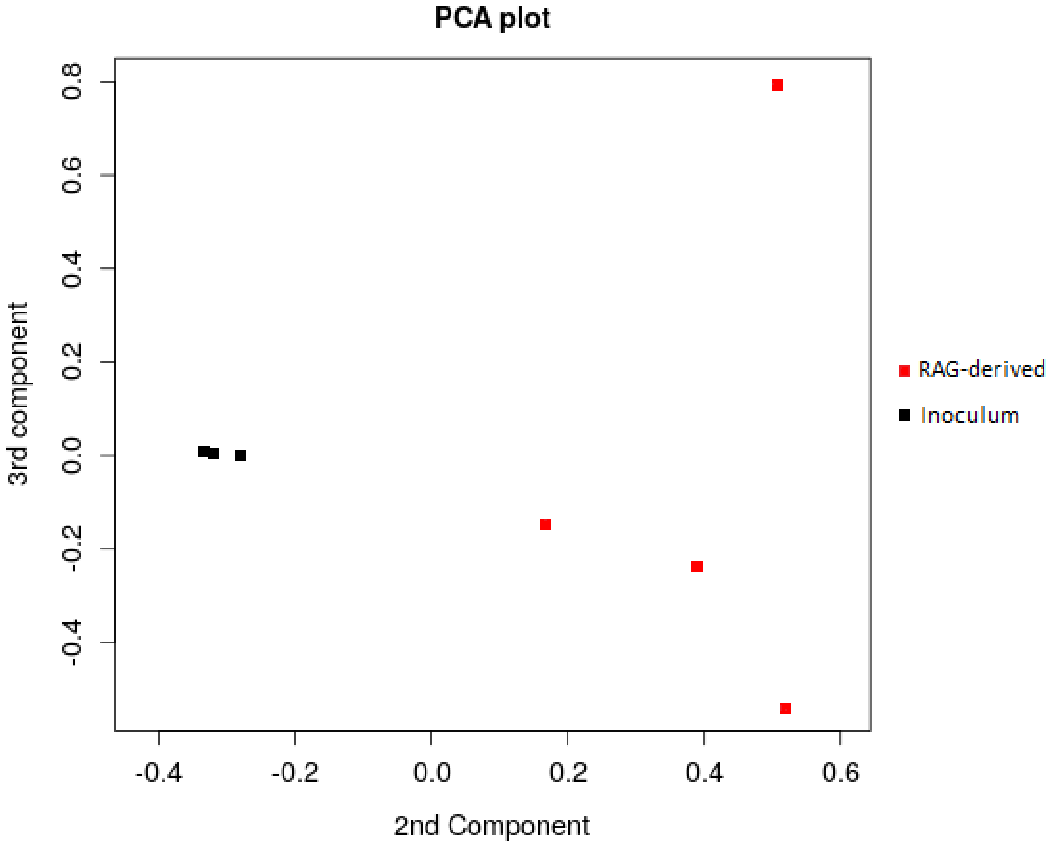


Figure 42. 2nd & 3rd PCA components for *Leishmania* transcriptome data. Inoculum-derived samples are represented in black, and RAG-derived samples are represented in red.

CHAPTER 4: DISCUSSION

4.1 THE MOUSE TRANSCRIPTOME

4.1.1 DIFFERENCES IN THE RESTING TRANSCRIPTOMES OF WT AND RAG MICE

A wide variety of immunity-related genes and cell types are found upregulated and enriched in WT B6 mice when compared with RAG2KO mice, even in an uninfected state. In terms of raw expression, the most highly differentially expressed genes in the WT were related to immunoglobulin heavy and light chains, and Fc receptors, indicating these genes are expressed at comparatively low levels in RAG mice. Matching this trend, the pathway and GSEA analyses found significant enrichment in terms related to immune cells, NK-cell mediated cytotoxicity, and the nitrogen metabolism. Overall, the results point to the presence of a resting adaptive immune system in one group of samples (WT), and the absence of one in the other (RAG). This was to be expected given that a lack of RAG recombinase prevents the development of mature B-cells, T-cells and immunoglobulins, but no other structural, physical or behavioural defects are present (Mombaerts et al. 1992).

Curiously, the most highly expressed genes in RAG mice relative to the WT group are resistin, leptin and perilipin 1, which are all involved in the regulation of hunger and adipose tissue (Vaisse et al. 1996; Martinez-Botas et al. 2000; Sul 2004). Despite the expression of these diet-related genes, no differences in weight or physiology are apparent between the two mice backgrounds. These genes are also expressed highly in the RAG samples when comparing WT and RAG mice under infection.

4.1.2 CHARACTERISTICS OF A HEALTHY IMMUNE RESPONSE

Comparing the transcriptomes of a WT mouse at rest and during infection should give an indication of how mouse gene expression reacts to the presence of the *Leishmania* parasite. Serum amyloid A3, a protein associated with acute inflammatory response, is the most highly upregulated gene in infected mice when compared with uninfected mice (Shimizu et al. 1992). Additionally, pathway analysis shows coagulation and complement cascades as being more highly expressed in infected mice. GO analysis shows enrichment for nitrogen and hydrogen peroxide metabolic processes and cellular detoxification, while GSEA matches a number of the genes found to be differentially expressed with immune memory and inflammatory response. The presence of pro-inflammatory genes, along with genes related to immune memory and nitrogen metabolic processes, is suggestive of an active immune system, aggressive in response to an infection (Franco et al. 2012). Healthy mice are known to respond to *L. donovani* infection in 2 phases, corresponding to innate and acquired immunity (Lipoldová and Demant 2006). Evidence of both phases can be seen in our own data, from upregulation in complement cascades and immune memory genes.

4.1.3 IMMUNOCOMPROMISED RESPONSE TO INFECTION

Given the lack of acquired immunity in RAG2 KO mice, and their consequent inability to generate pathogen-targeting immunoglobulins, T-cells, and B-cells, mice rely solely on their innate immune system to deal with infections. RAG mice are known to have high levels of NK cell activity (Belizário 2009) but other studies have found mice that are NK cell deficient are still able to clear the liver of *L. donovani* unimpeded (Lipoldová and Demant 2006). This indicates that NK cells are not crucial in the immune system's attempts to clear parasites.

Comparatively fewer genes were found to be differentially expressed between the RAG infected and uninfected samples - under 100 genes, when, for example, comparing uninfected WT and RAG mice found over 1,200. However, a small number of differentially expressed genes still constitutes a difference and a potential response to infection. In the infected samples, the most highly upregulated genes are related to immunoglobulins - heavy variable and light kappa chains. It's possible that the presence of *Leishmania* in the body is being detected, triggering a response to infection, and the initial genes relevant for initiating an acquired immune response are being upregulated; however, even if transcripts for immunoglobulins are being produced, this does not necessarily translate into the production of antibodies. B-cell progenitors and precursors are still present in the bodies of RAG mice, which may explain the apparent upregulation of B-cell-related immunity, but without RAG recombinase, the cells do not reach maturity.

Too few genes were differentially expressed between the samples for GO and pathway analysis. GSEA, however, found that a number of the differentially expressed genes matched with the Hallmark set for haeme metabolism. The role of iron metabolism in *Leishmania* infections is complicated, but indirect evidence suggests that the availability of iron from the host may affect the success of the parasite (Silva-Gomes et al. 2013). Depriving the parasite of iron, and the additive effect of iron involvement in NO_x and reactive oxygen species, is thought to have a leishmanicidal activity (Silva-Gomes et al. 2013). Thus, in modulating the expression of genes in the haeme metabolic pathway, infected RAG mice may be attempting to defend themselves despite the lack of acquired immunity.

4.1.4 COMPARING IMMUNE REACTIONS TO OTHER INTRACELLULAR PATHOGENS

The majority of transcriptomic profiling work on the effects on the host of other *Leishmania* species has been done on isolated macrophages rather than whole-tissue RNA samples, or multi-organ profiling. Given that the macrophages are both home to the parasites, and responsible for a portion of defence, highlighting macrophage response is insightful. However, the influence of genetic background and subsequent differential gene expression can decide the outcome of an infection in humans and mice. For example, in human *L. braziliensis* infections, patients that developed mucocutaneous leishmaniasis from cutaneous infection were found to have differential expression in genes that were involved in inflammatory response and cell migration (Maretti-Mira et al. 2012). Another recent study on *L. braziliensis*, based on transcriptomes generated from skin lesion biopsies, found that patients with and without detectable parasite transcripts had considerably different gene expression profiles. Samples positive for parasite transcripts showed upregulation of cell migration, cellular cytotoxicity, and inflammation; negative samples instead were found to upregulate genes related to skin defences and epidermal cell development (Christensen et al. 2016).

As with *Leishmania* infections, the presence of obligate intracellular protist *Toxoplasma gondii* has a notable effect on the host transcriptome. Unsurprisingly, a number of immunity related genes are upregulated during infection, such as immunoglobulins, chemokines, interferons, and MHC II, in similar inflammatory patterns to those of *Leishmania* infection (Tanaka et al. 2013). As with human *Leishmania* infections, the outcome of mouse infections with *T. gondii* can be determined by the host gene expression profile - symptomatic mice had higher expression of TGF-β and interferon regulatory factor 4 (Tanaka et al. 2013).

In the same manner, the success of murine malaria *Plasmodium berghei* infection is strongly influenced by host genetic background. BALB/c and B6 mice are known to have different resting and immunological gene expression profiles, like B6 and RAG mice. Pre-infection gene expression patterns can be analysed to convey information about resistance; mice with higher resting expression of immune response and defence related genes are more likely to resist infection than mice with comparatively lower levels of expression. For example, B6 mice are known to have relatively low levels of such genes, leaving them both susceptible to infection and sluggish to respond (Lovegrove et al. 2006).

Similarly, resistance to infection with intracellular bacterial pathogens are determined by genetic background and gene expression. In 4 strains of mice infected with *Mycobacterium tuberculosis*, hosts displayed varying patterns of resistance and susceptibility. It was found that macrophages from more susceptible mice stimulated recruitment of cell types that caused inflammatory immunopathology rather than microbial clearance (Keller et al. 2004). *Listeria monocytogenes*, another vacuole-residing bacterial pathogen, infects intestinal epithelial cells and hepatocytes though surface receptors (Pamer 2004). Unmodified mice of different genetic backgrounds are known to show differential susceptibility and resistance; A/J mice were found to be extremely susceptible to infection, while Black 6 mice were resistant to inoculation, showing slower bacterial dissemination and less severe lesions (Czuprynski et al. 2003).

Given that mouse reaction to the above mentioned intracellular pathogens - and intracellular stages of *P. berghei* infection - it appears that mice have a generalised response to infection (Schnappinger et al. 2006), consisting of a carefully controlled inflammatory response, immunoglobulins, chemokines and cytokines. This reaction is apparent in the B6 WT mice upon infection with *L. donovani*, but given the lack of ability to coordinate an acquired immune response, absent in the RAG mice.

4.2 THE LEISHMANIA TRANSCRIPTOME

4.2.1 PARASITES BEFORE AND AFTER INFECTION

Of 8,224 possible genes annotated on the *L. donovani* reference genome, 8,066 had at least 1 read detected in either the inoculum or RAG-derived samples, implying good overall coverage. However, almost 50% of genes were considered differentially expressed by edgeR (before filtering), which may be a consequence of the significantly lower level of reads in the RAG-derived samples. After filtering, which involved a minimum logCPM threshold, only 1.1% of genes were considered differentially expressed.

Almost 8% of the short list of genes found to be differentially expressed between *Leishmania* in RAG mice and those in inoculum were amastins and amastin-like proteins. Amastins are surface glyco-proteins known for their life-cycle stage dependent expression; as the name suggests, amastins are highly expressed by amastigotes, replacing the LPG coat of promastigotes (Saxena et al. 2007). Given that both the inoculum and RAG-derived parasites are amastigotes, and that different amastins were upregulated or downregulated in different samples, this may reflect attempts to react to different microenvironments, i.e. culture or host tissue.

Previous studies on the transcriptomic differences between axenic and intracellular amastigotes found that 13% of *Leishmania mexicana* transcripts had a significant difference in abundance (Fiebig et al. 2015). The same study found that the most highly expressed genes in axenic parasites - when compared to intracellular amastigotes - were related to proteolysis, DNA binding and nucleosomes (Fiebig et al. 2015). Our own results are in agreement with this data, as raw fold-change data showed significant upregulation of histones and histone-like proteins in RAG-derived parasite samples, in addition to gene ontology analysis revealing enrichment for DNA packing and nucleic acid metabolism-related genes.

Additionally, amastigotes are known to adapt to intracellular life by changing their metabolic preference from glucose and proline to amino acid and fatty acid beta oxidation, a change which may be reflected in the gene ontology analysis detection of enrichment of genes related to glycine, serine and disaccharide metabolisms (Fiebig et al. 2015).

4.2.2 OBTAINING A "TRUE" SNAPSHOT OF GENE EXPRESSION

The reliability of the conclusions drawn from the *Leishmania* data may be called into question when considering the ultra-low read counts in RAG-derived samples, and other technical issues. Before the application of the extra reliability criteria to the differentially expressed genes list produced by edgeR, almost 50% of the total genes detected across all *Leishmania* samples were considered differentially expressed. Applying a minimum threshold for log counts per million to ensure genes with only a handful of reads are excluded from the analysis cuts the number of genes that are considered differentially expressed to only 1%. Although edgeR attempts to compensate for low read counts, for genes where one library has a count in the hundreds, and the other library has a read count of less than 10, it can be hard to reliably and robustly scale and normalise appropriately. For example, analysis found that no genes were uniquely expressed in the RAG-derived samples, 8 genes were found to be uniquely expressed in the inoculum-derived parasites, which could be a symptom of low read counts.

Obfuscating problems with read count discrepancies is the relatively poor annotation of the *L. donovani* reference genome. Although the most up-to-date reference genome available was used to align reads to, of the 89 genes found to be differentially expressed, 28 were entirely hypothetical, with a further 41 having declared putative function and 11 having a noted similarity (e.g amastin-like) to another protein - leaving only 11% of genes with a concrete, known function. 40% of the most highly differentially expressed genes are entirely hypothetical, which has a knock-on impact on the analyses which build upon the differentially expressed gene list, such as enrichment analyses. By comparison, the analyses based on the mouse genome generated differentially expressed gene lists with an average of 7.7% protein-like genes and 12.1% hypothetical genes, leaving 80.2% of genes with a known, confirmed function.

Continuing the discussion of potential technical issues, the unexplained inconsistency with the inoculum technical replicates may have affected the results: BioAnalyser analysis of RNA and cDNA samples found no significant difference between the technical replicates, library sizes were similar in terms of total reads sequenced before and quality control, and in terms of uniquely mapping reads. PCA and Euclidean distance plots show that the transcriptomic profiles of the three replicates are almost identical. The mapping process was repeated in case of alignment error but the results were the same; it is still entirely unknown why there was such a drastic difference in the percentage of reads mapping to each genome for the 3rd technical replicate, and whether this has impacted the analysis.

Leishmania have additional complexities to add to their analysis in their genome organisation and gene expression systems. For example, the abundance of transcripts may or may not have direct functional consequences. *Leishmania* are highly unusual, when compared with the majority of studied organisms, in their regulation of expression, which is largely post-transcriptional and post-translational (Rastrojo et al. 2013). Theoretically, most genes are expected to be expressed at a similar, basal level (Kramer 2012). With the above in mind, changes in transcript abundance may not have a direct effect on protein levels. Rastrojo et al. 2013 found that transcript abundance and protein levels only roughly correlate, and as such, may not be a good measure for changes in gene expression (Cohen-Freue et al. 2007; Kramer 2012; Rastrojo et al. 2013). Similarly, *Leishmania* species are not only constitutively aneuploid, but additionally exhibit mosaic aneuploidy on a population level (Downing et al. 2011; Sterkers et al. 2011; Lachaud et al. 2014; Rogers et al. 2014). The abundance of gene transcripts will be directly affected by the copy number of the gene, and the ploidy of the chromosome - as genes are expressed at a similar rate, an extra copy will essentially double the level of expression (Kramer 2012; Dumetz et al. 2017; Iantorno et al. 2017). RNA sequencing is not used for the detection of ploidy or CNV - but DNA sequencing would allow for the detection of changes in ploidy/CNV through read depth, which in turn could be applied to RNA data to compensate for - or compare to - differences between groups of samples.

Finally, compromises made in the experimental design process may waiver some of the legitimacy of the findings. Due to problems finding an adequate solution for storing RNA samples early on in the experiment, the mixed mouse/*Leishmania* RNA samples and the inoculum RNA sample were generated from two entirely separate experiments, more than 6 months apart, as the original inoculum sample degraded in storage. Although the inoculums were generated using identical methods, it is entirely possible that the parasites further adapted to the RAG mice in which they were cultivated in during the time between experiments, and that adaptation is what the differential gene expression analysis is detecting, rather than a true "before and after", where a parasite is adapting to a new host. It could also be argued that the process of parasite isolation itself prevents the examination of "before and after" snapshots of the *Leishmania* transcriptome - the inoculum results instead show a parasite reacting to the shock of being extracted from the spleen, resuspended in culture and treated with chemicals. The process of extracting and purifying amastigotes from tissue takes at least 40 minutes, and must be performed promptly, as exposure to room temperatures is enough to trigger differentiation into the promastigote life cycle stage. Given that *Leishmania* can react to the temperature that quickly, it is not unreasonable to assume that the parasites are also capable of reacting to other environmental stresses with similar speed.

4.3 IMPACT OF METHODOLOGY

4.3.1 THE CHALLENGES OF ANALYSING DUAL RNA-SEQ DATA

Every technique and method has inherent flaws. RNA sequencing approaches have, for example, biases introduced by the biochemistry of library prep and next-generation sequencing platforms (Zhang et al. 2014). Measures of abundance are affected by nucleotide composition and gene length; sequencing depth and sample replicates can have significant effect on results; alternative splicing, isoforms, biological and technical variation further complicates things (Zhang et al. 2014). However, methods have been optimised to, wherever possible, eliminate or reduce such bias. For example, during library prep, to ensure unbiased coverage of the transcriptome, random primers were used for the conversion of RNA into cDNA. Additional steps to normalise metrics affected by gene length and library size are undertaken by edgeR (Robinson et al. 2010).

The relative abundance of host RNA when compared with parasite RNA is a particular challenge for dual RNA Seq approaches. The *M. musculus* genome, at 2.5Gb, is almost 70 times larger than the *L. donovani* genome at 32.4Mb (Downing et al. 2011). If genome size was equal to the amount of RNA produced, this difference in genome size would cause a far larger proportion of the RNA to belong to the mouse.

Without even considering the percentage of genes that are actually expressed, mammalian genomes are widely understood to have an extremely high proportion of non-transcribed DNA, in the form of pseudogenes, transposons, and other "junk" DNA – in humans, as much as 98% of genome is thought to be non-coding (Chi 2016). Our own data shows that while *Leishmania* express the majority of their genome (99.3% of annotated genes have a read count of >10), mice express relatively less, with 21.4% of genes having a read count of <10. Despite this difference, the tissue from which the mixed RNA samples were generated from contain substantially more mouse cells than *Leishmania* cells, such that the proportion of host RNA is much higher. Consequently, in mixed host-parasite samples, the amount of *Leishmania* RNA is relatively low, contributing to low read counts and poor transcriptome coverage.

Similar CRACKIT dual RNA-Seq experiments using mice and *L. donovani* found that the methods used were sufficient to generate enough *Leishmania* data for robust statistical analysis (Kaye et al. 2017). However, these previous experiments had used a different strain of mouse – BALB/c – which show a difference in their immune response to the B6 used in this experiment. BALB/c mice are known for their Th2-biased adaptive response, while B6 mice have a Th1 bias (Sellers et al. 2012). Th1 responses are known for their ability to clear intracellular parasites (such as *L. donovani*) and as such, B6 are able to cope with infection. BALB/c mice, with their Th1 response, are unable to clear the parasites (Sellers et al. 2012). The stronger immune response to intracellular parasites in B6 mice may be a contributing factor in the low abundance of parasite reads from the mixed host/parasite RNA samples.

4.3.2 COPING WITH ALIGNER TOLERANCE

Increasing the error tolerance of the STAR aligner will have very likely increased the number of errors in the overall alignment process, as the tolerance is a blanket process, and not targeted specifically to deal with splice leader sequences. The presence of *Leishmania*-aligned reads in the uninfected mouse samples are almost certainly evidence of overly tolerant aligner settings. Investigation of the incorrectly aligned reads, such as their dispersion across the genome, and the nature of the region, such as non-coding or coding, may have held clues as to the extent and effects of the alignment choices. Overall, the settings chosen mean that as long as a minimum of 40bp of the read matches the reference genome, the read will not be rejected. In retrospect, the issue of the splice leader sequences could have been solved by a simple CutAdapt treatment, by specifying the “adaptor” as the same as the splice leader sequence. With no SL sequence present in any of the reads, the alignment settings can be stricter, and more accurate. For example, the end-to-end setting for STAR uses a much harsher algorithm for calculating alignment score, and would produce a much more reliable alignment (Dobin et al. 2013). Another option, or an additional option, would be to computationally separate the reads during the alignment process, mapping each library to the genomes of both the host and the parasite – only in two separate runs, instead of together in a concatenated file. This has the advantage of not confusing the placement of genes that might be highly conserved between genomes, however unlikely that may be for taxonomically distant *M. musculus* and *Leishmania*.

4.3.4 THRESHOLDS OF EXPRESSION AND SIGNIFICANCE

In addition to inappropriate alignment settings, the decision to consider the alignment of a single read to a gene as evidence of the gene's expression was also unsuitable. The highest raw read count for any mouse gene was 578,310, and 169,473 for *Leishmania*, which further highlights the potential noise introduced by the low threshold. A more appropriate threshold could have been chosen by statistically examining either raw or normalised read counts.

Given the low read counts of the *Leishmania* genes and how the majority of "significant" ($p \geq 0.05$) genes are filtered when even minor logFC filtering is applied, it's likely that the single-read threshold was even more inappropriate for the *Leishmania* data. The conclusions drawn from the data are extremely unreliable, even though the read counts have been normalised and had statistical tests applied.

4.3.5 MODELLING DIFFERENTIAL GENE EXPRESSION

Various statistical distributions are used to model the probability of a random read drawn from the library mapping uniquely to the target of interest. Early RNA sequencing work found that data typically fitted a Poisson distribution, where the variance of the dataset was equal to the mean, rather than the Gaussian distribution used for microarray work. However, as RNA-Seq included more biological replicates, it was found that the Poisson distribution underestimated variance, an issue known as overdispersion (Zhang et al. 2014). The negative binomial distribution was found to better model dispersion, accounting for overdispersion, even when few biological replicates are available, and is now commonly used by differential gene expression software (Robinson et al. 2010; Zhang et al. 2014). As such, other distributions are available for use in detection of differential gene expression, but are not considered as appropriate for RNA Seq analysis.

Zhang et al. 2014 also tested edgeR against other differential expression software, such as Cufflinks/Cuffdiff2 and DESeq. EdgeR was found to be more effective at finding true positives, however, the other software was better if false positives are a concern of the dataset. As with choice of distribution/model, other DE detection software are available, but may not be as appropriate.

4.4 APPLICATION OF FINDINGS

4.4.1 CONTRIBUTION TO THE CRACKIT MODEL

Due to the constraints of time and the unexpectedly low read counts for *Leishmania* in the mixed RNA samples, the aims of this project changed focus from studying the transcriptome of the parasite to that of the host. The original intention was to populate a complex computer model of *L. donovani* infection with bioinformatic data about the response of the parasite to the immune pressure of the host. Unfortunately the parasite data are insufficient for the original purpose of the experiment, and combined with the decoupling of the inoculum and mouse parasite experiments, means the in vivo experiment will likely have to be repeated.

However, some use may still come from the mouse data and the little available parasite data. Providing additional data to a large project such as the CRACKIT project is useful in establishing context for other data types, such as pharmacokinetic/pharmacodynamic and imaging data (Timmis et al. 2016). More experimental mouse data to support a model's assumptions and fine tune parameters could also be useful (Albergante et al. 2013; Timmis et al. 2016). More specifically, empirical evidence for relevant immunological details - such as the regulation of cytokines and CD markers - could even be used to help design thresholds and parameters of the model (Albergante et al. 2013). Aside from the inner workings of the model, qualitative characteristics generated from the bioinformatics data may also be used to validate in silico models, for example, the extent of biological variance and stochasticity, and whether model heterogeneity reflects these accurately (Albergante et al. 2013; Timmis et al. 2016).

4.4.2 EXTRAPOLATING POTENTIAL

Although the data generated from this project might not be used for its original purpose, it can still be utilised for other research. First and foremost, given that the experiment is most likely to be repeated, the methodology and experimental design choices can be refined in order to achieve better results with regards to the collection of *Leishmania* transcriptome data. Mouse transcriptome data may also be useful for researchers studying *Leishmania* or other pathogens outside of the CRACKIT project, for example, the comparison between WT B6 and RAG2KO mice has potential to be used in the study of immunodeficient mice. Or, a study comparing mouse response to intracellular infections may make use of the transcriptomes of uninfected and infected B6 mice.

4.5 FUTURE WORK

4.5.1 IMPROVEMENTS TO EXPERIMENTAL DESIGN

Given that the experiment is likely to be repeated, several lessons can be learned from this project. One obvious improvement is to use RNA from the same inoculum used to infect the mice, rather than a second inoculum generated months after the original experiment, in order to better obtain a picture of *Leishmania* adaptation to the host. Additionally, steps can be taken to improve the yield of *Leishmania* RNA in the mixed samples; for example, using a mouse strain known to have a less strong immune response to maximise parasite growth. Previous studies on other trypanosomatids have attempted to solve problems similar to those faced by dual host-*Leishmania* studies. In terms of isolation of parasite cells from blood and tissue, Mulindwa et al. 2014 found Diethylaminoethyl cellulose columns to be especially effective at separating *Trypanosoma brucei* parasites from whole blood. Use of splice leader sequences to selectively prime cDNA synthesis and PCR amplification is also an effective way to ensure a better parasite to host signal ratio (Mulindwa et al. 2014).

Further work could also be undertaken in understanding transcript heterogeneity for both host and parasite. Though more study may be necessary in order to better interpret *Leishmania* data, it has been suggested that the presence of multiple SL sites and differential use of UTRs and polyadenylation may have significant effect on the parasite transcriptome (Rastrojo et al. 2013).

Additional analyses on the available data could also be insightful. Exclusion of reads mapped to multiple locations in the genome(s) reduces the ambiguity in the placement of the read. However, in the case of the *Leishmania* genome, the presence of tandem gene arrays means that reads would not uncommonly correctly map to multiple loci. Instead of entirely ignoring these reads during analysis, they can be analysed through use of specific tools to appropriately estimate the coverage of the involved genes, or by randomly deciding between the placements of each read in order to not lose data. For a more comprehensive understanding of the immunology underpinning changes in gene expression, Chaussabel analysis, which involves analysing transcriptome data for cell-type and response-specific biomarkers, would help characterise the differences in mouse immune responses (Chaussabel et al. 2008). This type of analysis could establish which branches of the immune system – such as innate or acquired, or Th1 and Th2 – are responsible for the differences in transcriptomes. Venn diagrams comparing significantly expressed genes between samples would also be a useful addition in highlighting the differences in immune responses. For a more biochemical approach, inclusion of ligand data for the various receptors found significantly differentially expressed between samples, such as the cytokine and chemokine receptors included in the gene panels, may indicate whether the changes in gene expression have a strong biological effect.

4.5.2 APPLICATION OF NOVEL TRANSCRIPTOMICS APPROACHES

Alternative methods to study the host and parasite transcriptomes are available. If the objective of the experiment is to provide data to inform a model, it could be argued that although no longer cutting edge, microarray analysis could be useful. Using a microarray would allow for the study of specific targets, such as those being coded into the model, though poor annotation of the *Leishmania* genome would make production of *Leishmania* probes more difficult.

A multi-omics approach, including genomics and proteomics, would provide better context for the transcriptome profiles and better support claims about changes in gene expression. More specifically, DNA sequencing of *Leishmania* inoculum and RAG-derived samples may contribute to determining their response to host immune pressure, as *Leishmania* are well-established in their reactive genomic flexibility (Mannaert et al. 2012; Sterkers et al. 2012; Rogers et al. 2014). Proteomics work would be extremely valuable, to validate whether changes in the parasite transcriptome correlate with protein levels, or if their alternative methods of gene regulation render transcriptome data meaningless.

Use of innovative techniques such as single-cell transcriptome sequencing would be invaluable in future work. Single-cell based techniques would allow for the reduction of ‘background noise’ with regards to other cells in the tissue, for example, isolating macrophages and other immune cells from the mouse spleen, or better purification of *Leishmania* amastigotes from mouse tissue.

BIBLIOGRAPHY

Abolins S, King EC, Lazarou L, Weldon L, Hughes L, Drescher P, Raynes JG, Hafalla JCR, Viney ME, Riley EM. 2017. The comparative immunology of wild and laboratory mice, *Mus musculus domesticus*. *Nat Commun* **8**: 14811. doi:10.1038/ncomms14811.

Abrudan J, Ramalho-Ortigão M, O'Neil S, Stayback G, Wadsworth M, Bernard M, Shoue D, Emrich S, Lawyer P, Kamhawi S, et al. 2013. The characterisation of the *Phlebotomus papatasi* transcriptome. *Insect Mol Bio* **22**: 211-232.

Ahmed S, Colmenares M, Soong L, Goldsmith-Pestana K, Munstermann L, Molina R, McMahon-Pratt D. 2003. Intradermal infection model for pathogenesis and vaccine studies of murine visceral leishmaniasis. *Infect Immun* **71**: 401-410.

Akhoundi M, Kuhls K, Cannet A, Votýpka J, Marty P, Delaunay P, Sereno D. 2016. A Historical Overview of the Classification, Evolution, and Dispersion of Leishmania Parasites and Sandflies. *PLoS Negl Trop Dis* **10**: e0004349. doi:10.1371/journal.pntd.0004349.

Akopyants N, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, Dobson DE, Beverley SM, Sacks DL. 2009. Demonstration of Genetic Exchange During Cyclical Development of *Leishmania* in the Sand Fly Vector. *Science* **324**: 265-268.

Albergante L, Timmis J, Beattie L, Kaye PM. 2013. A Petri Net Model of Granulomatous Inflammation: Implications for IL-10 Mediated Control of *Leishmania donovani* Infection. *PLoS Comput Biol* **9**: e1003334. doi:10.1371/journal.pcbi.1003334.

Alcolea PJ, Alonso A, Domínguez M, Parro V, Jiménez M, Molina R, Larraga V. 2016. Influence of the Microenvironment in the Transcriptome of Leishmania infantum Promastigotes: Sand Fly versus Culture. *PLoS Negl Trop Dis* **10**: e0004693. doi:10.1371/journal.pntd.0004693

Anders S, Pyl PT, Huber W. 2015. HTSeq - a Python framework to work with high-throughput seuencing data. *Bioinformatics* **31**: 166-169.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res* **38**: D457-D462.

Atayde VD, Aslan H, Townsend S, Hassani K, Kamhawi S, Olivier M. 2015. Exosome Secretion by the Parasitic Protozoan *Leishmania* within the Sand Fly Midgut. *Cell Rep* **13**: 957-967.

Babai B, Louzir H, Cazenave PA, Dellagi K. 1999. Depletion of peritoneal CD5+ B cells has no effect on the course of *Leishmania major* infection in susceptible and resistant mice. *Clin Exp Immunol* **117**: 123-129.

Belizário JE. 2009. Immunodeficient Mouse Models: An Overview. *Open Immunol J* **2**: 79-85.

Bradley DJ, Taylor BA, Blackwell J, Evans EP, Freeman J. 1979. Regulation of Leishmania populations within the host. III. Mapping of the locus controlling susceptibility to visceral leishmaniasis in the mouse. *Clin Exp Immunol* **37**: 7-14.

Bylund J, Karlsson A, Boulay F, Dahlgren C. 2002. Lipopolysaccharide-induced granule mobilization and priming of the neutrophil response to Helicobacter pylori peptide Hp(2-20), which activates formyl peptide receptor-like 1. *Infect Immun* **70**: 2908-2914.

Cantacessi C, Dantas-Torres F, Nolan MJ, Otranto D. 2015. The past, present and future of *Leishmania* genomics and transcriptomics. *Trends Parasitol* **31**: 100-108.

Cao W, Rosen DB, Ito T, Bover L, Bao M, Watanabe G, Yao Z, Zhang L, Lanier LL, Liu YJ. 2006. Plasmacytoid dendritic cell-specific receptor ILT7-Fc epsilonRI gamma inhibits Toll-like receptor-induced interferon production. *J Exp Med* **203**:1399-1405.

Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, Stichweh D, Blankenship D, Li L, Mungala I, et al. 2008. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**: 150-164.

Chi KR. 2016. The dark side of the human genome. *Nature* **538**: 275-277.

Christensen SM, Dillion LAL, Carvalho LP, Passos S, Novais FO, Hughitt VK, Beiting DP, Carvalho EM, Scott P, El-Sayed NM, et al. 2016. Meta-transcriptome Profiling of the Human-*Leishmania braziliensis* Cutaneous Lesion. *PLoS Negl Trop Dis* **10**: e0004992. doi:10.1371/journal.pntd.0004992

Clayton C, Shapira M. 2007. Post-transcriptional regulation of gene expression in trypanosomes and leishmanias. *Mol Biochem Parasitol* **156**: 93-101.

Cohen-Freue G, Holzer TR, Forney JD, McMaster WR. 2007. Global gene expression *Leishmania*. *Int J Parasitol* **37**: 1077-1086.

Colmenares M, Puig-Kröger A, Muňiz Pello O, Corbí AL, Rivas L. 2002. Dendritic Cell (DC)-specific Intercellular Adhesion Molecule 3 (ICAM-3)-grabbing Nonintegrin (DC-SIGN, CD209), a C-type Surface Lectin in Human DCs, Is a Receptor for *Leishmania* Amastigotes. *J Biol Chem* **227**: 36766-36769.

Croft SL, Sundar S, Fairlamb AH. 2006. Drug Resistance in Leishmaniasis. *Clin Microbiol Rev* **19**: 111-126.

Czuprynski CJ, Faith NG, Steinberg H. 2003. A/J Mice Are Susceptible and C57BL/6 Mice Are Resistant to *Listeria monocytogenes* Infection by Intragastric Inoculation. *Infect Immun* **71**: 682-689.

de Morais-Teixeira E, Souza Damasceno Q, Kolos Galuppo M, José Romanha A, Rabello A. 2011. The in vitro leishmanicidal activity of hexadecylphosphocholine (miltefosine) against four medically relevant *Leishmania* species of Brazil. *Mem Inst Oswaldo Cruz* **106**: 475-478.

Desjardins M, Descoteaux A. 1997. Inhibition of Phagolysosomal Biogenesis by the *Leishmania* Lipophosphoglycan. *J Exp Med* **12**: 2061-2068.

Dillon LAL, Suresh R, Okrah K, Corrada Bravo H, Mosser DM, El-Sayed NM. 2015. Simultaneous transcriptional profiling of *Leishmania major* and its murine macrophage host cell reveals insights into host-pathogen interactions. *BMC Genomics* **16**. doi:10.1186/s12864-015-2237-2.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram J, et al. 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* **21**: 2143-2156.

Dumetz F, Imamura H, Sanders M, Seblova V, Myskova J, Pescher P, Vanaerschot M, Meehan CJ, Cuypers B, De Muylder G, et al. 2017. Modulation of Aneuploidy in *Leishmania donovani* during Adaptation to Different *In Vitro* and *In Vivo* Environments and Its Impact on Gene Expression. *mBio* **8**: e00599-17. https://doi.org/10.1128/mBio.00599-17

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualisation of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**: doi:10.1186/1471-2105-10-48.

Engwerda CR, Ato M, Stäger S, Alexander CE, Stanley AC, Kaye PM. 2004. Distinct Roles for Lymphotoxin-α and Tumor Necrosis Factor in the Control of *Leishmania donovani* Infection. *Am J Pathol* **165**: 2123-2133.

Faleiro RJ, Kumar R, Hafner LM, Engwerda CR. 2014. Immune Regulation during Chronic Visceral Leishmaniasis. *PLoS Negl Trop Dis* **8**: e2914. doi:10.1371/journal.pntd.0002914.

Fernandes MC, Dillon LAL, Belew AT, Corrada Bravo H, Mosser DM, El-Sayed NM. 2016. Dual Transcriptome Profiling of *Leishmania*-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. *mBio* **7**: e00027-16. doi:10.1128/mBio.00027-16.

Fiebig M, Kelly S, Gluenz E. 2015. Comparative Life Cycle Transcriptomics Revises *Leishmania mexicana* Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates. *PLoS Pathog* **11**: e1005186. doi:10.1371/journal.ppat.1005186.

Forestier C, Machu C, Loussert C, Pescher P, Späth G. 2011. Imaging Host Cell-*Leishmania* Interaction Dynamics Implicates Parasite Motility, Lysosome Recruitment, and Host Cell Wounding in the Infection Process. *Cell Host Microbe* **9**: 319-330.

Franco LH, Beverley SM, Zamboni DS. 2012. Innate Immune Activation and Subversion of Mammalian Functions by *Leishmania* Lipophosphoglycan. *J Parasitol Res* **2012**: 165126. doi:10.1155/2012/165126.

Gazit R, Gruda R, Elboim M, Arnon TI, Katz G, Achdout H, Hanna H, Qimrom U, Landau G, Greenbaum E, et al. 2006. Lethal influenza infection in the absence of the natural killer cell receptor gene Ncr1. *Nat Immunol* **7**: 517-523.

Ghoshal K, Majumder S, Zhu Q, Hunzeker J, Datta J, Shah M, Sheridan JF, Jacob ST. 2001. Influenza virus infection induces metallothionein gene expression in the mouse liver and lung by overlapping but distinct molecular mechanisms. *Mol Cell Biol* **21**: 8301-8317.

Gillespie PM, Beaumier CM, Strych U, Hayward T, Hotez PJ, Bottazzi ME. 2016. Status of vaccine research and development of vaccines for leishmaniasis. *Vaccine* **34**: 2992-2995.

Gluenz E, Ginger ML, McKean PG. 2010. Flagellum assembly and function during the *Leishmania* life cycle. *Curr Opin Microbiol* **13**: 473-479.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333-351.

Gossage SM, Rogers ME, Bates PA. 2003. Two separate growth phases during the development of *Leishmania* in sand flies: implications for understanding the life cycle. *Int J Parasitol* **33**: 1027-1034.

Hardison JL, Wrightsman RA, Carpenter PM, Kuziel WA, LAne TE, Manning JE. 2006. The CC Chemokine Receptor 5 Is Important in Control of Parasite Replication and Acute Cardiac Inflammation following Infection with *Trypanosoma cruzi*. *Infect Immun* **74**: 135-143.

Iantorno SA, Durrant C, Khan A, Sanders MJ, Beverley SM, Warren WC, Berriman M, Sacks DL, Cotton JA, Grigg ME. 2017. Gene Expression in *Leishmania* Is Regulated Predominantly by Gene Dosage. *mBio* **8**: 8:e01393-17. https://doi.org/10.1128/mBio.01393-17.

Illumina Inc. 2011. Quality Scores for Next-Generation Sequencing. *Technical Note: Sequencing*. Available: http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

Illumina Inc. 2014. Understanding Illumina Quality Scores. *Technical Note: Informatics*. Available: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E, Mueller O, Schroeder A, Auffray C. 2005. Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Res* **33**: e56. https://doi.org/10.1093/nar/gni054

Inbar E, Akopyants NS, Charmoy M, Romano A, Lawyer P, Elnaiem DA, Kauffmann F, Barhoumi M, Grigg M, Owens K, et al. 2013. The Mating Competence of Geographically Diverse *Leishmania major* Strains in Their Natural and Unnatural Sand Fly Vectors. *PLoS Genetics* **9**: e1003672. doi:10.1371/journal.pgen.1003672.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, et al. 2005. The Genome of the Kinetoplastid Parasite, *Leishmania major*. *Science* **309**: 436-442.

Jaramillo M, Adelaida Gomez M, Larsson O, Tiemi Shio M, Topisirovic I, Contreras I, Luxenburg R, Rosenfeld A, Colina R, McMaster RW, et al. 2011. *Leishmania* Repression of Host Translation through mTOR Cleavage Is Required for Parasite Survival and Infection. *Cell Host Microbe* **9**: 331-341.

Janeway CA, Travers P, Walport M, Shlomchik MJ. 2001. Immunobiology. 5th ed. Garland Science, New York.

Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software]. Available at https://github.com/najoshi/sickle.

Kamhawi S, Ramalho-Ortigão M, Pham VM, Kumar S, Lawyer PG, Turco SJ, Barillas-Mury C, Sacks DL, Valenzuela J. 2004. A Role for Insect Galectins in Parasite Survival. *Cell* **119**: 329-341.

Kanazawa N, Okazaki T, Nishimura H, Tashiro K, Inaba K, Miyachi Y. 2002. DCIR acts as an inhibitory receptor depending on its immunoreceptor tyrosine-based inhibitory motif. *J Invest Dermatol* **118**: 261-266.

Kaye P, Forrester S, Mottram J. 2017. *Potential causes of low parasite read counts in mixed host-parasite RNA samples* [Project meeting]. 11 May 2017, 14:00.

Kaye P, Scott P. 2011. Leishmaniasis: complexity at the host-pathogen interface. *Nat Rev Microbiol* **9**: 604-615.

Kazemi B. 2011. Genomic Organisation of *Leishmania* Species. *Iran J Parasitol* **6**: 1-18.

Kedzierski L, Zhu Y, Handman E. 2006. *Leishmania* vaccines: progress and problems. *Parasitol* **133**: S87-S112.

Keller C, Lauber J, Blumenthal A, Buer J, Ehlers S. 2004. Resistance and susceptibility to tuberculosis analysed the transcriptome level: lessons from mouse macrophages. *Tuberculosis* **84**: 144-158.

Killick-Kendrick R. 1990. The life-cycle of *Leishmania* in the sandfly with special reference to the form infective to the vertebrate host. *Ann Parasitol Hum Comp* **65**: 37-42.

Kima PE, Soong L. 2013. Interferon gamma in Leishmaniasis. *Front Immunol* **4**: 156. doi: 10.3389/fimmu.2013.00156.

Kitamura A, Takahashi K, Okajima A, Kitamura N. 1994. Induction of the human gene for p44, a hepatitis-C-associated microtubular aggregate protein, by interferon-alpha/beta. *Eur J Biochem* **15**: 877-883.

Kleiman E, Salyakina D, De Heusch M, Hoek KL, Llanes JL, Castro I, Wright JA, Clark ES, Dykxhoorn DM, Capobianco E, et al. 2015. Distinct transcriptomic features are associated with transitional and mature B-cell populations in the mouse spleen. *Front Immunol* **6:** doi: 10.3389/fimmu.2015.00030.

Kramer S. 2012. Developmental regulation of gene expression in the absence of transcriptional control: The case of kinetoplastids. *Mol Biochem Parasitol* **181**: 61-72.

Kumar S, Mitnik C, Valente G, Floyd-Smith G. 2000. Expansion and molecular evolution of the interferon-induced 2’-5’ oligoadenylate synthetase gene family. *Mol Biol Evol* **17**: 738-750.

Kumar R, Engwerda C. 2014. Vaccines to prevent leishmaniasis. *Clin Transl Immunology* **3**: e13. doi:10.1038/cti.2014.4.

Lachaud L, Bourgeois N, Kuk N, Morelle C, Crobu L, Merlin G, Bastien P, Pagès M, Sterkers Y. 2014. Constitutive mosaic aneuploidy is a unique genetic feature widespread in the *Leishmania* genus. *Microbes Infect* **16**: 61-66.

Landfear SM. 2003. Trypanosomatid transcription factors: Waiting for Godot. *PNAS* **100**: 7-9.

Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-2089.

Liehl P, Zuzarte-Luis V, Mota MM. 2015. Unveiling the pathogen behind the vacuole. *Nat Rev Microbiol* **13**: 589-598.

Lin H, Grosschedl R. 1995. Failure of B-cell differentiation in mice lacking the transcription factor EBF. *Nature* **376**: 263-267.

Lipoldová M, Demant P. 2006. Genetic susceptibility to infectious disease: lessons from mouse models of leishmaniasis. *Nat Rev Genet* **7**: 294-305.

Loeuillet C, Bañuls AL, Hide M. 2016. Study of *Leishmania* pathogenesis in mice: experimental considerations. *Parasit Vectors* **9**: 144. doi:10.1186/s13071-016-1413-9

Loría-Cervera EN, Andrade-Narváez FJ. 2014. Animal models for the study of leishmaniasis immunology. *Rev Inst Med Trop Sao Paulo* **56**: 1-11.

Lovegrove FE, Peña-Castillo L, Mohammad N, Conrad Liles W, Hughes TR, Kain KC. 2006. Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. *BMC Genomics* **7**: doi:10.1186/1471-2164-7-295.

Luo W, Brouwer C. 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualisation. *Bioinformatics* **29**: 1830-1831.

Maggi L, Santarlasci V, Capone M, Peired A, Frosali F, Crome SQ, Querci V, Fambrini M, Liotta F, Levings MK, et al. 2010. CD161 is a marker of all human IL-17-producing T-cell subsets and is induced by RORC. *Eur J Immunol* **40**: 2174-2181.

Mannaert A, Downing T, Imamura H, Dujardin JC. 2012. Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. *Trends Parasitol* **28**: 370-376.

Maretti-Mira AC, Bittner J, Oliveira-Neto MP, Liu M, Kang D, Li H, Pirmez C, Craft N. 2012. Transcriptome Patterns from Primary Cutaneous *Leishmania braziliensis* Infections Associate with Eventual Development of Mucosal Disease in Humans. *PLoS Negl Trop Dis* **6**: e1816. doi:10.1371/journal.pntd.0001816.

Maroof A, Brown N, Smith B, Hodgkinson MR, Maxwell A, Losch FO, Fritz U, Walden P, Lacey CNJ, Smith DF, et al. 2012. Therapeutic Vaccination With Recombinant Adenovirus Reduces Splenic Parasite Burden in Experimental Visceral Leishmaniasis. *J Infect Dis* **205**: 853-863.

Martin M. 2011. CutAdapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**: 10-12. doi:http://dx.doi.org/10.14806/ej.17.1.200.

Martinez-Botas J, Anderson JB, Tessier D, Lapillonne A, Chang BH, Quast MJ, Gorenstein D, Chen KH, Chan L. 2000. Absence of perilipin results in leanness and reverses obesity in Lepr(db/db) mice. *Nat Genet* **26**: 474-479.

Maxwell-Silverman J, Clos J, Camargo de'Oliveira C, Shirvani O, Fang Y, Wang C, Foster LJ, Reiner NE. 2009. An exosome-based secretion pathway is responsible for protein export from *Leishmania* and communication with macrophages. *J Cell Sci* **123**: 842-852.

Maxwell-Silverman J, Clos J, Horakova E, Wang AY, Wiegigl M, Kelly I, Lynn MA, McMaster R, Foster LJ, Levings MK, et al. 2011. Leishmania Exosomes Modulate Innate and Adaptive Immune Responses through Effects on Monocytes and Dendritic Cells. *J Immunol* **185**: 5011-5022.

Maxwell-Silverman J, Reiner NE. 2011. Exosomes and other microvesicles in infection biology: organelles with unanticipated phenotypes. *Cell Microbiol* **13**: 1-9.

Maxwell-Silverman J, Reiner NE. 2012. *Leishmania* exosome deliver preemptive strikes to create an environment permissive for early infection. *Front Cell Infect Microbiol* **1**: 26.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**: 4288-4297.

McConville MJ, Mullin KA, Ilgoutz SC, Teasdale RD. 2002. Secretory Pathway of Trypanosomatid Parasites. *Microbiol Mol Biol R* **66**: 122-154.

Melo GD, Marcondes M, Vasconcelos RO, Machado GF. 2009. Leukocyte entry into the CNS of *Leishmania chagasi* naturally infected dogs. *Vet Parasitol* **162**: 248-256.

Mendes Roatt B, de Oliveira Aguiar-Soares RD, Coura-Vital W, Gama Ker H, das Dores Moreira N, Vitoriano-Souza J, Cordeiro Giunchetti R, Martins Carneiro C, Barbosa Reis A. 2014. Immunotherapy and immunochemotherapy in visceral leishmaniasis: promising treatments for this neglected disease. *Front Immunol* **5**: 272. doi:10.3389/fimmu.2014.00272.

Metas J, Hughes CCW. 2004. Of Mice and Not Men: Differences between Mouse and Human Immunology. *J Immunol* **172**: 2731-2738.

Mombaerts P, Iacomini J, Johnson RS, Herrup K, Tonegawa S, Papaioannou VE. 1992. RAG-1-Deficient Mice Have No Mature B and T Lymphocytes. *Cell* **68**: 869-877.

Moore EM, Lockwood DN. 2011. Leishmaniasis. *Clin Med* **11**: 492-497.

Moradin N, Descoteaux A. 2012. *Leishmania* promastigotes: building a safe niche within macrophages. *Front Cell Infect Microbiol* **2**: 121.

Mulindwa J, Fadda A, Merce C, Matovu E, Enyaru J, Clayton C. 2014. Methods to Determine the Transcriptomes of Trypanosomes in Mixtures with Mammalian Cells: The Effects of Parasite Purification and Selective cDNA Amplification. *PLoS Negl Trop Dis* **8**: e2806. doi:10.1371/journal.pntd.0002806.

Nandan D, Camargo de Oliveira C, Moeenrezakhanlou A, Lopez M, Silverman JM, Subek J, Reiner NE. 2012. Myeloid Cell IL-10 Production in Response to Leishmania Involves Inactivation of Glycogen Synthase Kinase-3β Downstream of Phosphatidylinositol-3 Kinase. *J Immunol* **188**: 367-378.

O'Brien AD, Rosenstreich DL, Taylor BA. Control of natural resistance to *Salmonella typhimurium* and *Leishmania donovani* in mice by closely linked but distinct genetic loci. *Nature* **287**: 440-442.

Okwor I, Uzonna J. 2016. Social and Economic Burden of Human Leishmaniasis. *Am J Trop Med Hyg* **94**: 489-493.

Pace D. 2014. Leishmaniasis. *J Infect* **69**: S10-S18.

Pamer EG. 2004. Immune responses to Listeria monocytogenes. *Nat Rev Immunol* **4**: 812-823.

Parsons M, Worthey EA, Ward PN, Mottram JC. 2005. Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. *BMC Genomics* **6**: 127. doi:10.1186/1471-2164-6-127.

Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al. 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat Genet* **39**: 839-847.

Perry MR, Wyllie S, Raab A, Feldmann J, Fairlamb AH. 2013. Chronic exposure to arsenic in drinking water can lead to resistance to antimonial drugs in a mouse model of visceral leishmaniasis. *PNAS* **110**: 19932-19937.

Pommerenke C, Wilk E, Srivastava B, Schulze A, Novoselova N, Geffers R, Schughart K. 2012. Global Transcriptome Analysis in Influenza-Infected Mouse Lungs Reveals the Kinetics of Innate and Adaptive Host Immune Responses. *PLoS ONE* **7**: e41169. doi:10.1371/journal.pone.0041169.

Rabhi I, Rabhi S, Ben-Othman R, Rasche A, Sysco Consortium, Daskalaki A, Trentin B, Piquemal D, Regnault B, Descoteaux A, et al. 2012. Transcriptomic Signature of *Leishmania* Infected Mice Macrophages: A Metabolic Point of View. *PLoS Negl Trop Dis* **6**: e1763. doi:10.1371/journal.pntd.0001763.

Rai K, Cuypers B, Raj Bhattarai N, Uranw S, Berg M, Ostyn B, Dujardin JC, Rijal S, Vanaerschot M. 2013. Relapse after Treatment with Miltefosine for Visceral Leishmaniasis Is Associated with Increased Infectivity of the Infecting *Leishmania donovani* Strain. *mBio* **4**: e00611-13. doi: 10.1128/mBio.00611-13.

Rastrojo A, Carrasco-Ramiro F, Martín D, Crespillo A, Reguera R, Aguado B, Requena JM. 2013. The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. *BMC Genomics* **14**: 223. doi:10.1186/1471-2164-14-223.

Ravel C, Cortes S, Pratlong F, Morio F, Dedet J, Campino L. 2006. First report of genetic hybrids between two very divergent *Leishmania* species: *Leishmania infantum* and *Leishmania major*. *Int J Parasitol* **36**: 1383-1388.

Reithinger R, Aadil K, Kolaczinski J, Mohsen M, Hami S. 2005. Social Impact of Leishmaniasis, Afghanistan. *Emerg Infect Diseases* **11**: 634-636.

Reithinger R, Dujardin JC, Louzir H, Pirmez C, Alexander B, Brooker S. 2007. Cutaneous leishmaniasis. *Lancet Infect Dis* **7**: 581-596.

Ritter U, Frischknecht F, van Zandbergen G. 2009. Are neutrophils important host cells for *Leishmania* parasites? *Trends Parasitol* **25**: 505-510.

Roberts LJ, Baldwin TM, Curtis JM, Handman E, Foote SJ. 1997. Resistance to *Leishmania major* is Linked to the H2 region on Chromosome 17 and to Chromosome 19. *J Exp Med* **185**: 1705-1710.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.

Rogers ME. 2012. The role of *Leishmania* proteophosphoglycans in sand fly transmission and infection of the mammalian host. *Front Microbiol* **3**: 223. doi: 10.3389/fmicb.2012.00223.

Rogers MB, Downing T, Smith BA, Imamura H, Sanders M, Svobodova M, Volf P, Berriman M, Cotton JA, Smith DF. 2014. Genomic Confirmation of Hybridisation and Recent Inbreeding in a Vector-Isolated *Leishmania* Population. *PLoS Genet* **10**: e1004092. doi:10.1371/journal.pgen.1004092.

Sacks D, Noben-Trauth N. 2002. The immunology of susceptibility and resistance to *Leishmania major* in mice. *Nat Rev Immunol* **2**: 845-858.

Saether PC, Westgaard IH, Flornes LM, Hoelsbrekken SE, Ryan JC, Fossum S, Dissen E. 2005. Molecular cloning of KLRI1 and KLRI2, a novel pair of lectin-like natural killer-cell receptors with opposing signalling motifs. *Immunogenetics* **56**: 833-839.

Sato N, Kuziel WA, Melby PC, Reddick RL, Kostecki V, Zhao W, Maeda N, Ahuja SK, Ahuja SS. 1999. Defects in the Generation of IFN-γ Are Overcome to Control Infection with *Leishmania donovani* in CC Chemokine Receptor (CCR) 5-, Macrophage Inflammatory Protein-1α-, or CCR2-Deficient Mice. *J Immunol* **163**: 5519-5525.

Saxena A, Lahav T, Holland N, Aggarwal G, Anupama A, Huang Y, Volpin H, Myler PJ, Zilberstein D. 2007. Analysis of the *Leishmania donovani* transcriptome reveals an ordered progression of transient and permanent changes in gene expression during differentiation. *Mol Biochem Parasitol* **152**: 53-65.

Schlein Y. 1993. *Leishmania* and Sandflies: Interactions in the Life Cycle and Transmission. *Parasitol Today* **9**: 255-258.

Schnappinger D, Schoolnik GK, Ehrt S. 2006. Expression profiling of host pathogen interactions: how *Mycobacterium tuberculosis* and the macrophage adapt to one another. *Microb Infect* **8**: 1132-1140.

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**: 839-851.

Sellers RS, Clifford CB, Treuting PM, Brayton C. 2012. Immunological variation between inbred laboratory mouse strains: points to consider in phenotyping genetically immunomodified mice. *Vet Pathol* **49**: 32-43.

Shimizu H, Mitomo K, Yamamoto K. 1992. Regulation of mouse serum amyloid A3 gene expression during acute phase reaction. *Folia Histochem Cytobiol* **30**: 141-142.

Shinkai Y, Rathbun G, Lam K, Oltz EM, Stewart V, Mendelsohn M, Charron J, Datta M, Young F, Stall AM, et al. 1992. RAG-2-Deficient Mice Lack Mature Lymphocytes Owing to Inability to Initiate V(D)J Rearrangement. *Cell* **68**: 855-867.

Siegel TN, Gunasekera K, Cross GAM, Ochsenreiter T. 2011. Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing. *Trends Parasitol* **27**: 434-441.

Silva-Gomes S, Vale-Costa S, Appelberg R, Gomes MS. 2013. Iron in intracellular infection: to provide or to deprive? *Front Cell Infect Microbiol* **3**: 96. https://doi.org/10.3389/fcimb.2013.00096

Späth G, Schlesinger P, Schreiber R, Beverley SM. 2009. A Novel Role for Stat1 in Phagosome Acidification and Natural Host Resistance to Intracellular Infection by *Leishmania major*. *PLoS Pathog* **5**: e1000381. doi:10.1371/journal.ppat.1000381.

Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. 2011. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. *Cell Microbiol* **13**: 274-283.

Sterkers Y, Lachaud L, Bourgeois N, Crobu L, Bastien P, Pagès M. 2012. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in *Leishmania*. *Mol Microbiol* **86**: 15-23.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**: 15545-15550.

Sul HS. 2004. Resistin/ADSF/FIZZ3 in obesity and diabetes. *Trends Endocrin Met* **15**: 247-249.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarises and Visualises Long Lists of Gene Ontology Terms. *PLoS One* **6**: e21800. doi:10.1371/journal.pone.0021800.

Tailleux L, Waddel SJ, Pelizzola M, Mortellaro A, Withers M, Tanne A, Ricciardi Castagnoli P, Gicquel B, Stoker NG, Butcher PD, et al. 2008. Probing Host Pathogen Cross-Talk by Transcriptional Profiling of Both *Mycobacterium tuberculosis* and Infected Human Dendritic Cells and Macrophages. *PLoS ONE* **3**: e1403. doi:10.1371/journal.pone.0001403.

Tanaka S, Nishimura M, Ihara F, Yamagishi J, Suzuki Y, Nishikawa Y. 2013. Transcriptome Analysis of Mouse Brain Infected with *Toxoplasma gondii*. *Infect Immun* **81**: 3609-3619.

Thompson CB. 1992. RAG knockouts deliver a one/two punch. *Curr Biol* **2**: 180-182.

Timmis J, Alden K, Andrews P, Clark E, Nellis A, Naylor B, Coles M, Kaye P. 2016. Building Confidence in Quantitative Systems Pharmacology Models: An Engineer's Guide to Exploring the Rationale in Model Design and Development. *CPT Pharmacometrics Syst Pharmacol* **5**: doi:10.1002/psp4.12157.

Trinconi CT, Reimão JQ, Yokohama-Yasunaka JKU, Miguel DC, Uliana SRB. 2014. Combination Therapy with Tamoxifen and Amphotericin B in Experimental Cutaneous Leishmaniasis. *Antimicrob Agents Chemother* **58**: 2608-2613.

Trinconi CT, Reimão JQ, Coelho AC, Uliana SRB. 2016. Efficacy of tamoxifen and miltefosine combined therapy for cutaneous leishmaniasis in the murine model of infection with *Leishmania amazonensis*. *J Antimicrob Chemother* **71**: 1314-1322.

Tschoeke DA, Nunes GL, Jardim R, Lima J, Dumaresq ASR, Gomes MR, de Mattos Pereira L, Loureiro DR, Stoco PH, Leonel de Matos Guedes H, et al. 2014. The Comparative Genomics and Phylogenomics of *Leishmania amazonensis* Parasite. *Evol Bioinform* **10**: 131-153.

Vaisse C, Halaas JL, Horvath CM, Darnell JE Jr, Stoffel M, Friedman JM. 1996. Leptin activation of Stat3 in the hypothalamus of wild-type and ob/ob mice but not db/db mice. *Nat Genet* **14**: 95-97.

Van Dommelen SL, Sumaria N, Schreiber RD, Scalzo AA, Smyth MJ, Degli-Esposti MA. 2006. Perforin and granzymes have distinct roles in defensive immunity and immunopathology. *Immunity* **25**: 835-848.

Van Griensven J, Balasegaram M, Meheus F, Alvar J, Lynen L, Boelaert M. 2010. Combination therapy for visceral leishmaniasis. *Lancet Infect Dis* **10**: 184-194.

Van Luenen HGAM, Farris C, Jan S, Genest P, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A, Ramasamy G, et al. 2012. Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in *Leishmania*. *Cell* **150**: 909-921.

Victoir V, Dujardin JC. 2002. How to succeed in parasitic life without sex? Asking *Leishmania*. *Trends Parasitol* **18**: 81-85.

Vikeved E, Backlund A, Alsmark C. 2016. The Dynamics of Lateral Gene Transfer in Genus *Leishmania* - A Route for Adaptation and Species Diversification. *PLoS Negl Trop Dis* **10**: e0004326. doi:10.1371/journal.pntd.0004326.

Vinet AF, Fuduka M, Turco SJ, Descoteaux A. 2009. The *Leishmania donovani* Lipophosphoglycan Excludes the Vesicular Proton-ATPase from Phagosomes by Impairing the Recruitment of Synaptotagmin V. *PLoS Pathog* **5**: e1000628. doi:10.1371/journal.ppat.1000628

Viney M, Lazarou L, Abolins S. 2015. The laboratory mouse and wild immunology. *Parasite Immunol* **37**: 267-273.

Wells CA, Salvage-Jones JA, Li X, Hitchens K, Butcher S, Murray RZ, Beckhouse AG, Lo YL, Manzanero S, Cobbold C, et al. 2008. The macrophage-inducible C-type lectin, mincle, is an essential component of the innate immune response to Candida albicans. *J Immunol* **180**: 7404-7413.

World Health Organisation. 2010. Control of the Leishmaniases. *WHO Tech Rep Ser* **949**: 1-186.

World Health Organisation. 2013. Frequently asked questions on visceral leishmaniasis (kala-azar). Available: http://www.who.int/neglected_diseases/resources/B5042/en/

World Health Organisation. 2016. Leishmaniasis in high-burden countries: an epidemiological update based on data reported in 2014. *Wkly Epidemiol Rec* **91**: 285-296.

Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, et al. 2014. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS ONE* **9**: e103207. doi:10.1371/journal.pone.0103207.

Zhang X, Lai M, Chang W, Yu I, Ding K, Mrazek J, Ng HL, Yang OO, Maslov DA, Zhou ZH. 2016. Structures and stabilization of kinetoplastid-specific split rRNAs revealed by comparing leishmanial and human ribosomes. *Nat Commun* **7:** 13223. DOI: 10.1038/ncomms13223

Zhou Y, Bizzaro JW, Marx KA. 2004. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. *BMC Genomics* **5**: 95. doi:10.1186/1471-2164-5-95

Zschaler J, Schlorke D. Arnhold J. 2014. Differences in innate immune response between man and mouse. *Crit Rev Immunol* **34**: 433-454.