Using OMIC approaches to understand the genetic mechanisms controlling virulence in *Trypanosoma brucei rhodesiense* isolates

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Sarah Jayne Forrester

July 2015

UNIVERSITY OF LIVERPOOL

Acknowledgements

I would like to firstly thank my supervisors, Neil and Harry, who both gave me the opportunity to do this project in the first place, and for their support and encouragement during the last 4 years. To Neil, for your sarcasm, and in turn, tolerating mine (and the endless nagging on the final stretch), to Harry, for instilling your enthusiasm in all things "tryp" in me, and the endless cups of tea. I'd also like to thank certain academic members (you'll know who you are, presumably), who inspired me to embark on a PhD in the first place from my undergraduate days in Liverpool. I'd also like to thank Ian Goodhead, who's bearded wisdom often put me on the right path, directed me to sources of caffeine, and whose work was the forerunner to my project. Special thanks also go out to the members of the BSU whose help was invaluable, in particular Pam, Stephen, Jo and Lynn.

To the past and current members of the CGR and Lab E/F members, thanks for helping me maintain a modicum of sanity throughout the last 4 years, in particular for your support during the last 12 months. I would also like to thank Laura, Jen, Tom and Ian for moments of bioinformatic wizardry, particularly during the earlier stages of my PhD. Thanks in particular to members of the office who were subject to my incessant tea drinking and for enabling me in my habit.

I reserve my sincerest thanks for my family, my mum, dad and Maureen, who without their emotional and calorific support, I would have struggled without. And last but not least, I would like to thank my nan, who even though you couldn't see me through to completion, cheered me on every step of the way, and whose relentless enthusiasm motivated me to do a PhD.

Abstract

Due to the resurgence of human African trypanosomiasis (HAT), the world health organization (WHO) and various non-governmental organisations NGO's have implemented strategies that have led to a significant drop in disease incidence, with only ~3,500 new HAT cases recorded in 2014. However the causative agent, *T. brucei* is still responsible for a heavy socio-economic burden, with *T. brucei* infections in cattle representing an estimated billion dollar loss annually. Of the two human infective *T. brucei* subspecies, *T.b. rhodesiense* is responsible for less than 3% of all HAT cases and is primarily considered a zoonosis. It causes acute disease comparative to the *T.b. gambiense* subspecies and two strains of differing phenotypes have been used to establish experimental infections, which reproduce the clinical manifestation observed in natural infections. This project utilized technological advances in order to understand the genetic mechanisms driving the phenotypic differences observed through the use of genomic, transcriptomic and metabolomic analysis.

Firstly this work discusses the feasibility of sequencing directly from field samples by using Whatman FTA™ card and sequence capture in combination, and benchmarking the data against available whole genome sequence data. The resulting data showed both successful enrichment and lack of allelic drop out effect. This methodology was subsequently applied to multiple other strains and used to look at deleterious variants potentially giving rise to these phenotypic differences. Gross differences in the abundance of bloodstream forms in these strains were also observed, which indicated the phenotype of these strains may result from the regulation of differentiation. Transcriptomic and metabolomic data was also used to identify differential regulation driving these differences in virulence, and showed that only a small subset of genes were differentially regulated. Amongst these several candidate genes, which had previously been associated with drug resistance, were identified. Genomic and transcriptomic data also indicated that iron regulation is one of the key mechanisms driving this phenotypic change, with a high density of deleterious SNPs located in iron transport, and the greater than 10 fold increase in the expression of the transferrin receptor found in the transcriptomic data. However further analysis ideally on a larger set of strains, or SNPs derived from the entire genome rather than a subset, would be necessary to ascertain whether this is true.

Table of Contents

List of tables

List of figures

CHAPTER 1

Introduction

1.1 The *Trypanosoma* genus

Trypanosomes are protozoan parasites responsible for a heavy health burden worldwide and the causative agent of several human and cattle diseases both in America and Sub Saharan Africa. The *Trypanosoma* genus has two distinct clades, the *Salivarian* and non Salivarian (*Stercorarian*) trypanosomes, which cause disease in Africa and America respectively and were first described by Hoare in 1972 (Stevens et al., 1999; Hoare, 1972; Stevens and Gibson, 1999b; 1999a). Although both are heteroxenous, with an insect vector, the method of transmission and life cycle of the trypanosome varies (Teixeira and Soulsby, 1987; Stevens et al., 1999; Stevens and Gibson, 1999b; 1999a).

1.1.1 Stercorarian trypanosomes

The non-salivarian trypanosomes include bird and reptile trypanosomes, which are the most well known of the Stercorarian trypanosomes (Truc et al., 2013; Teixeira and Soulsby, 1987; Ramsey et al., 2015). Included within the Stercorarian trypanosomes is *Trypanosoma cruzi*, which is well known for being the causative agent of Chagas disease in America, but also less well renowned trypanosomes such as *T. rangeli* (Ramsey et al., 2015; Truc et al., 2013). These trypanosomes complete their lifecycle by transmission through a Triatominae host. Similarly to the Salivarian trypanosomes, the insect vector takes a blood meal from an infected mammalian host, ingesting the parasite. In Stercorarian trypanosomes, the parasite then develops within the insect host, becoming infectious, and is then subsequently excreted when the insect defecates following its blood meal. The parasites then pass through the bite wound (Stevens and Gibson, 1999b; Ramsey et al., 2015).

1.1.2 Salivarian trypanosomes

The Salivarian trypanosomes include one of the most well studied trypanosomes, *T. brucei*, the causative agent of trypanosomiasis, and the less studied but economically very important *T. congolense* and *T. vivax* (Stevens and Gibson, 1999b). Compared to the non-Salivarians, the vector, tsetse flies (*Glossina morsitans*), have been far more extensively studied, including the release of the genome of the tsetse fly in 2014 (Brun et al., 2010; International Glossina Genome Initiative et al., 2014).

1.2 Human African trypanosomiasis (HAT)

Human African trypanosomiasis (HAT) is caused by the parasite *Trypanosoma brucei*. Its vector, the tsetse fly (*Glossina* spp) restricts the distribution of infections to within the tsetse belt, which is situated in Sub-Saharan Africa. Human African Trypanosomiasis in East Africa, as caused by the species *Trypanosoma brucei rhodesiense*, is a particularly acute form of "sleeping sickness", with patients often suffering severe symptoms sometimes only one week after infection (Giroud et al., 2009; Brun et al., 2010). There are other described subspecies of *T. brucei* in Africa, *Trypanosoma brucei gambiense*, which causes sleeping sickness in West Africa (usually resulting in chronic infections) and *Trypanosoma brucei brucei*, which infects cattle (Brun et al., 2010; Giroud et al., 2009). Collectively, *T.b. gambiense* and *T.b. rhodesiense* account for all human African trypanosomiasis (HAT). Human infections are predominantly caused by *T.b. gambiense*, with *T.b. rhodesiense* infections thought to only responsible for approximately 3% of HAT cases (Hamilton et al., 2004; Brun et al., 2010; Leonard et al., 2011; Balmer et al., 2011).

The relationship between these subspecies is not well understood nor is the underlying molecular cause of their host range or clinical phenotype (Berriman, 2005; Hamilton et al., 2004; Leonard et al., 2011; Brun et al., 2010; Balmer et al., 2011). However insights made from the sequencing of the *T. brucei. brucei* strain Tb927 in 2005 suggest that there is very little genetic variation that accounts for this vast difference in clinical manifestation, with over 99% genome identity between all 3 subspecies (Jackson et al., 2010) . This is similar to what has been seen in other parasites, for instance *Leishmania* was found to have only 200 genes with a differential distribution between sub species, with vastly differing phenotypes (Simarro et al., 2008; Peacock et al., 2007; Simarro, 2011).

T. brucei has to be considered a neglected tropical disease (NTD), due to lack of public interest despite its potential to infect 500,000 people a year (Simarro, 2011; Simarro et al., 2008; World Health Organization, 2014). However compared with the majority of infectious diseases, disease incidence in humans has dropped significantly, with WHO reporting ~3,500 cases in 2014, compared to 30,000 cases in 1990 (Simarro, 2011; World Health Organization, 2014). This has been done largely through improved surveillance and prevention strategies, and indicates that despite no current vaccine leads, and very old drug treatments which are becoming more redundant with surges in drug resistance, eradication of human African trypanosomiasis is conceivable (Bainbridge et al., 2010; Simarro, 2011; Mardis, 2008). Until recently there hasn't been much interest in understanding the interplay between parasite and host, however this is starting to change as the decreasing cost of sequencing and other high-throughput technologies is allowing for more robust analysis on these more complicated datasets (Luikart et al., 2003; Bainbridge et al., 2010; Forrester and Hall, 2014; Mardis, 2008; Morrison et al., 2010).

A greater understanding on the molecular mechanisms controlling disease progression and virulence is also being achieved by starting to combine data from these technologies, as seen in Chapter 4, which combines RNAseq and metabolomics analysis. The combination of these technologies, from genomic to transcriptomic and metabolomics analysis is far more powerful in providing biological insight. These advances in OMIC technologies and the decrease in price has also allowed for analysis to turn from individual strain to population genomics, which again is important for determining loci important in virulence and resistance (Forrester and Hall, 2014).

This decline in cost has also allowed for attention to been turned towards species that were previously largely unstudied, this includes *T. congolense*, *T. vivax*, which are responsible for the majority of trypanosome infections in cattle, and the less common human infective sub-species, *T.b. rhodesiense* (Kristjanson et al., 1999; Forrester and Hall, 2014; O'Gorman et al., 2009). Trypanosomes do continue to cause a heavy burden of disease, however the majority of this occurs within cattle, which has a heavy socio-economic effect.

1.3 *T. brucei* is a heteroxenous parasite requiring two hosts to complete its life cycle

Trypanosoma brucei requires an insect vector, in this case tsetse flies from the *Glossina* genus, and a mammalian host to complete its life cycle. Despite their unicellular nature, *T. brucei* undergoes a number of complex morphological changes in order to adapt to both tsetse fly and mammalian environments (Sternberg and Maclean, 2010; Matthews et al., 2004; Rico et al., 2013). The life cycle can be separated into two key sets of stages, the tsetse and human/mammalian stages, these are illustrated in Figure 1.1 beneath.

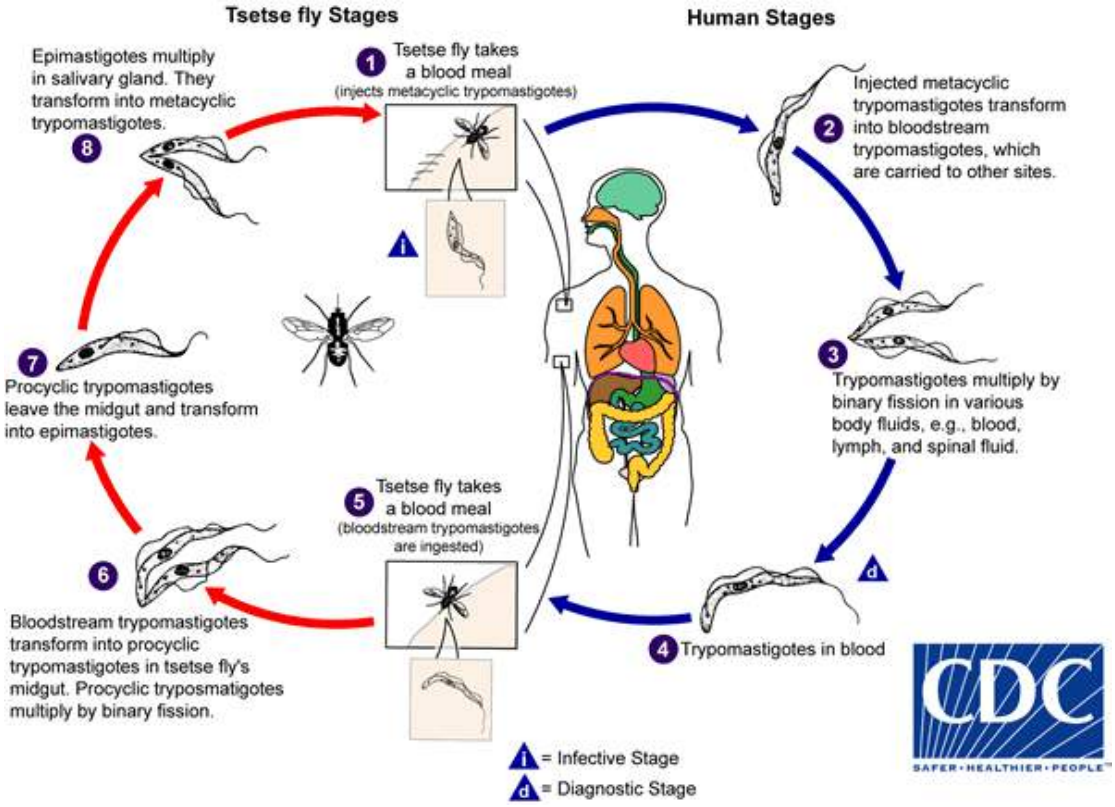

Figure 1.1: Taken with permission from the centre of disease control and prevention (www.cdc.gov). Stages of *T. brucei's* life cycle that occur within the tsetse fly vector are shown in red, stages that occur within the mammalian host are shown in blue. Stages 1 and 5 are important because these are the infective stages, stage one is when the parasites are infective to mammalian hosts and stage five is when the parasites are infective to tsetse hosts.

The mammalian stages of infection begin by the injection of metacyclic trypomastigotes into the mammalian host's bloodstream via bite. Occasionally swelling occurs around the bite region, which is referred to as a chancre and is associated with typically more virulent *T. brucei* strains (Matthews et al., 2004; Sternberg and Maclean, 2010; MacLean et al., 2010). These injected forms then differentiate within the bloodstream into the first bloodstream form stages, the long slender forms. These are highly

proliferative and often represent a high percentage of the bloodstream forms present at the beginning of an infection (Reuner et al., 1997; Matthews et al., 2004; MacGregor et al., 2011). However as the parasitaemia increases, the burden of the parasites on the host increases, and so to prevent the host being overwhelmed and death, these parasites differentiate into short stumpy forms (Seed and Wenck, 2003; Reuner et al., 1997; Duszenko et al., 2006; MacGregor et al., 2011). The parasites method of "quorum sensing" and the mechanism it uses to trigger subsequent differentiation is not wholly understood, but is believed to be controlled by an as yet unidentified stumpy inducing factor (SIF) (Seed and Wenck, 2003; Duszenko et al., 2006).

Differentiation from the long slender forms (LS) to the short stumpy forms (SS) requires a number of highly regulated morphological changes. Once differentiated into these SS stages, the parasite's cell cycle is arrested and many metabolic pathways and previously highly expressed genes, are downregulated (Rico et al., 2013; Seed and Wenck, 2003; Duszenko et al., 2006). These stages last 48 hours prior to apoptosis and can be split into two populations. The younger population of SS forms are tsetse fly infective and are adapted to survive within the tsetse fly host. The older population of SS forms are no longer tsetse fly infective and improve the chances of uptake of the younger SS population, in what has previously been coined, an altruistic manner (Kennedy, 2004; Rico et al., 2013).

The highly proliferative LS forms are capable of either staying in the bloodstream or migrating to extravascular regions and passing the blood-brain barrier (BBB) and entering the central nervous system (CNS). This migration triggers a number of symptoms within the host and signals the late stage of infection (Matthews et al., 2004; Kennedy, 2004). In both the bloodstream and within the CNS, the LS forms replicate asexually via binary fission, as noted in Figure 1.1, stage 3 (Fenn and Matthews, 2007; Matthews et al., 2004).

The SS forms are prepared for tsetse fly infection, and infect the tsetse fly during its blood meal. Initially the trypanosomes colonize the midgut of the tsetse fly, and here SS forms differentiate into procyclic trypomastigotes (Vassella et al., 2009; Fenn and Matthews, 2007). A key molecular change here is the switching of the variable surface glycoprotein (VSG) coat, which enables the parasite to evade the host's immune system within the mammalian host, to a procyclic coat (Van Den Abbeele et al., 1999; Vassella

et al., 2009). Once differentiated to trypomastigotes, the trypanosomes resume cell division and multiple by binary fission as shown in Figure 1.1.

Midgut procyclic trypanosomes then migrate to the salivary gland via the peritrophic matrix, the foregut, the proventriculus and the salivary ducts, as can be seen in Figure 1.2. Whilst in the proventriculus, the procyclic trypomastigotes undergo asymmetric division to generate a short and long epimastigote (Van Den Abbeele et al., 1999). The short epimastigote attaches to the epithelial cells of the salivary gland. Once attached, they then replicate and undergo another asymmetric division to generate metacyclic trypomastigotes which are adapted for mammalian survival (Gibson and Bailey, 2003; Van Den Abbeele et al., 1999).

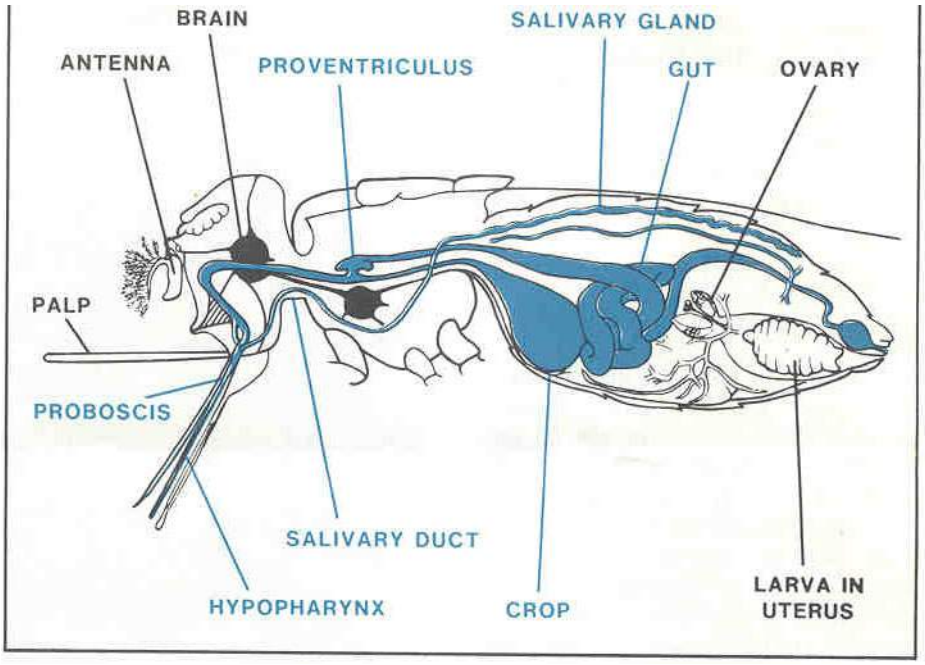

Figure 1.2: Taken with permission from www.CDC.gov. *T. brucei* initially colonizes the midgut of the tsetse fly, once here the short stumpy forms (SS) differentiate into procyclic trypomastigotes and exchange a VSG coat for a procyclic coat. These then migrate to the salivary glands via the proventriculus where they differentiate into epimastigates. Once in the salivary glands the replicate and become metacyclic trypomastigotes, ready to infect another mammalian host.

1.4 *T. brucei* is transmitted by tsetse flies but are poorly adapted to this host

HAT is transmitted by tsetse flies from the *Glossinia* genus. However less than 1% of wild flies have salivary gland infections, which suggests that trypanosomes are poorly adapted to their tsetse host (Gibson & Bailey, 2003). Although *T. brucei* takes at least two weeks to complete its life cycle and be transmissible, the fly will continue to produce human infective metacyclics for the remainder of its life (Gibson and Stevens, 1999; Gibson and Bailey, 2003; Tait et al., 2002). Tsetse flies require blood meals every few days and so continue to infect new hosts. The trypanosome life cycle stages in the tsetse fly are particularly important because this is where genetic exchange between parasites takes place (Gibson and Stevens, 1999; Sternberg and Maclean, 2010; Tait et al., 2002).

1.5 The *Trypanosoma brucei* genus can be divided into three subspecies

1.5.1 *Trypanosoma brucei brucei (T.b. brucei)* is the non-human infective subspecies used as a model of disease

T.b. brucei is the only non-human infective *T. brucei* sub species, and causes Nagana in cattle. *T. congolense* and *T .vivax* are also responsible for animal African trypanosomiasis (AAT) and collectively lead to an estimated loss of a billion dollars annually (O'Gorman et al., 2009; Kristjanson et al., 1999). Over approximately 60 million cattle are at risk from AAT in 37 endemic countries, which significantly reduces cattle productivity (Berriman, 2005; O'Gorman et al., 2009). Despite its inability to infect humans, the ease at which *T.b. brucei* is manipulated makes it a common model for human disease, particularly due to the high degree of similarity between sub species (~99%) (Berriman, 2005; Jackson et al., 2010). The most commonly used reference, Tb927, is a *T.b. brucei* strain which was used in the first sequencing of the trypanosome genome, and remains in use due to comparatively poor genome annotation of the released *T.b. gambiense* reference, DAL (Berriman, 2005; Jackson et al., 2010).

1.5.2 The roles of APOL1 and HPR in the lysis of trypanosomes

T.b. brucei is unable to infect humans because it is susceptible to the trypanolytic factors TLF-1 and TLF-2. These are complexes which contain the haptoglobin related protein (Hpr) and apolipoprotein L1 (ApoL1). These complexes differ in the protein components they contain, with TLF-1 comprised primarily of apolipoprotein A-1 (apoA-1) and Hpr, and TLF-2 primarily immunoglobulin M, (IgM), apoA-1 and Hpr. Unlike TLF-1, which is a high density lipoprotein (HDL), TLF-2 only contains less than 1% of lipid and is a high molecular weight protein binding complex (Pays et al., 2006; Hajduk et al., 1989; Vanhollebeke and Pays, 2006; Raper et al., 1999).

1.5.3 Apolipoprotein L genes are involved in apoptosis

The apolipoprotein L family (apoL), is comprised of six members, named apoL1-6, which arose as a result of tandem duplication (Pays et al., 2006; Vanhollebeke and Pays, 2006). Until recently, their function was unknown. ApoL1 has been more extensively studied, due to its known role in killing bloodstream trypanosome forms, and is a secreted protein. However the remaining members of the family are intracellular, and their proposed roles were in lipid transport and metabolism, due to ApoL1s known association with HDLs (Vanhollebeke and Pays, 2006; Pays et al., 2006). Due to their structural similarity to Bcl-2 proteins, this protein family's primary role is considered to be in regulating the mechanisms trigging apoptosis (Drain et al., 2001; Vanhollebeke and Pays, 2006; Seed et al., 2007).

1.5.4 The haptoglobin related protein and the haptoglobin-hemoglobin receptor compete for haem binding

Similarly to the expansion process of the apoL family, the haptoglobin related protein (Hpr) gene is a result of a gene triplication event. Hpr is primate specific, but apart from its protective role against *T.b. brucei,* its function is unknown (Langlois and Delanghe, 1996; Drain et al., 2001; Shimamura et al., 2001; Seed et al., 2007). Haptoglobin (hp) is a tetrameric plasma protein comprised of two alpha and beta chains. Hp has a high affinity for haemoglobin and promotes its clearance from the blood and subsequent transport to the lysosome by binding to it and forming a haptoglobin-haemoglobin (Hp-Hb) complex (Drain et al., 2001; Langlois and Delanghe, 1996; Shimamura et al., 2001).

HPR has a greater than 90% homology to Hp, which is an abundant serum protein. However its abundance in human serum is several hundredfold lower than haptoglobin (Hp) (Drain et al., 2001). Interestingly, Hp naturally inhibits the TLF-1 complex and the variation in trypanosome lytic activity is correlated with an individual's serum concentration of Hp (Harrington et al., 2010; Drain et al., 2001). HPR contains hydrophobic peptides that target the killing of specifically bloodstream forms, leaving procyclic forms untouched (Raper et al., 1999; Harrington et al., 2010; Drain et al., 2001). However Hp does not inhibit TLF-2 formation (Widener et al., 2007; Raper et al., 1999; Vanhollebeke et al., 2007; Drain et al., 2001).

Both APOL1 and HPR are currently believed to be essential for optimal parasite lysis (Capewell et al., 2011; Widener et al., 2007; Vanhollebeke et al., 2007). The mechanism has been characterized in the TLF-1 complex, but in TLF-2 is still not wholly understood. The TLF-1 complex is first taken-up by the parasite by the binding of HPR to the parasite's haptoglobin-hemoglobin receptor (HpHbR). Once internalized, the TLF-1 complex is then targeted to the lysosome. pH changes within the lysosome trigger the activation of the APOL1 element of the complex, which then undergoes conformational changes which result in pores forming within the lysosomes membrane. Lysis results from the subsequent osmotic changes (Vanhollebeke et al., 2008; Capewell et al., 2011; Vanhollebeke and Pays, 2010). Due to their key roles in the uptake of TLF-1, and their presence within the TLF-2 complex, it is likely that they have important roles in TLF-2's mode of action, however binding and internalization does not involve the HpHbR receptor directly (Lugli et al., 2004; Vanhollebeke et al., 2008; Vanhollebeke and Pays, 2010).

As previously mentioned, these trypanolytic complexes are found only in primates because the Hpr and ApoL1 genes are unique to the primate genome. Due to this, these trypanolytic factors are found only in a handful of non-human species including baboons and gorillas, however chimpanzees do not produce these complexes (Pays et al., 2006; Jamonneau et al., 2012; Lugli et al., 2004; Sternberg and Maclean, 2010). The human-infective species have acquired mechanisms, which enable them to resist these trypanolytic factors and not be lysed in human serum.

1.5.5 Human-infective subspecies of *T. brucei*

HAT is caused by *T.b. gambiense* and *T.b. rhodesiense*. As previously mentioned, the majority of human infections are caused by the *T.b. gambiense* subspecies. These subspecies are broadly considered to cause chronic, and acute infections respectively, however these infections result in a whole spectrum of symptoms (Capewell et al., 2011; Jamonneau et al., 2012; Sternberg and Maclean, 2010). Both *T.b. gambiense* and *T.b. rhodesiense* infections are resistant to lysis but use alternative mechanisms, which are discussed below and outlined in Figure 1.3.

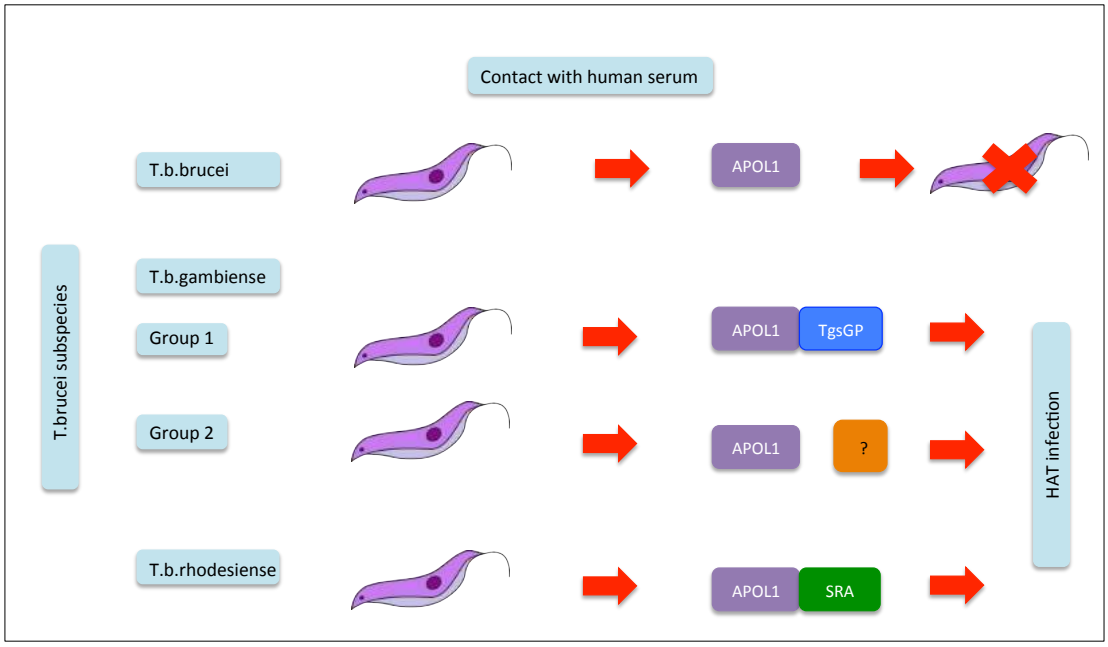

Figure 1.3: *T. brucei* is susceptible to lysis upon interaction with human serum due to the action of APOL1. *T.b. gambiense* and *T.b. rhodesiense* are resistant to lysis, however the mechanism of lysis resistance remains to be elucidated in group 2 T.b. gambiense. Group 1 *T.b. gambiense* resists lysis through TbgSP, and *T.b. rhodesiense* through SRA, the action of both is explained in the text beneath.

1.5.6 *T. brucei gambiense* is comprised of two groups, each with a different mechanism of resistance to lysis by TLF

T.b. gambiense causes African trypanosomiasis in western and central Africa, and is generally considered to cause slow onset chronic infections, however manifestations vary widely. Infected patients generally take months to years to present with symptoms, and the progression from the first to second stage of disease is often slow. Initially, *T. brucei* was considered to exist as just three individual subspecies, however *T.b. gambiense* can also be further divided into groups 1 and 2. This has implications for

their mechanisms of resistance to lysis, as described in 2011 (Capewell et al., 2011; 2013b).

Group 1 *T.b. gambiense* strains are clonal in nature, and have an invariant phenotype but are genetically distinct from group 2 *T.b. gambiense* strains and *T.b. brucei* (Mehlitz et al., 1982; Capewell et al., 2011; Balmer et al., 2011; Capewell et al., 2013b; Sternberg and Maclean, 2010). In contrast, group 2 strains are not genetically distinct from *T.b. brucei* and have variable phenotypes (Capewell et al., 2011; Mehlitz et al., 1982; Balmer et al., 2011; Sternberg and Maclean, 2010). It has been suggested previously that mating between group 1 *T.b. gambiense* and *T.b. brucei* could give rise to group 2 *T.b. gambiense* strains, but there is no current evidence to support this. However, there is evidence of mating within group 2 *T.b. gambiense* strains and between group 2 strains and *T.b. brucei*, which may explain the variability in phenotype (Capewell et al., 2013a; 2011). This also suggests that human infectivity has evolved via independent mechanisms in each group (Goodhead et al., 2013; Capewell et al., 2013a; Gibson et al., 2015). However variable phenotypes have previously been described in *T.b. rhodesiense* and other species, with hybrid genomes resulting from mating between species (Gibson, 1986; Goodhead et al., 2013; Paindavoine et al., 1989; Gibson et al., 2015; Capewell et al., 2013b).

Predominantly *T.b. gambiense* infections are caused by group 1 strains, group 2 strains have only been described in regions where group 1 strains coexist and only within Côte d'Ivoire, Cameroon and Burkina Faso (Capewell et al., 2013b; Gibson, 1986; Paindavoine et al., 1989). This led to questions over the origin of group 2 *T.b. gambiense* strains and the potential consequences of mating between and within these subspecies (Picozzi et al., 2005; Capewell et al., 2013b; 2013a).

Although the mechanism of evading lysis in group 2 *T.b. gambiense* strains has not been elucidated, the mechanism in group 1 has been well characterized, and uses an alternative method to group 1 for avoiding lysis (Kieft et al., 2010; Picozzi et al., 2005; Capewell et al., 2013a). In group 1 strains, there is reduced expression of the receptor TLF-1 binds to, HbHpR, which reduces uptake of the TLF-1 complex but doesn't entirely confer resistance (Capewell et al., 2011; Kieft et al., 2010). They still remain sensitive to APOL1, however they avoid lysis by preventing its uptake (Capewell et al., 2011). Gene regulation in trypanosomes is primarily controlled by the 3' UTR, and Capewell and colleagues identified several polymorphisms unique to the group 1 *T.b. gambiense*

strains within the 3'UTR of HbHpR which could explain its downregulation (DeJesus et al., 2013; Capewell et al., 2011). In 2013, a group 1 specific leucine to serine substitution at codon 210 of the HbHpR was shown to be responsible for abolishing TLF-1 binding (DeJesus et al., 2013).

Another key difference between groups 1 and 2 is that group 1 *T.b. gambiense* contains a *T.b. gambiense*-specific glycoprotein (TgsGP), which is absent from group 2 strains (Uzureau et al., 2013; Gibson et al., 2010; Capewell et al., 2011). TgsGP prevents the binding of APOL1, by stiffening membranes upon interaction with lipids using its hydrophobic beta-sheet (Uzureau et al., 2013).

Although this sub species can infect both animals and human, its predominant host is human, although animal reservoirs do exist (Berriman, 2005; Jackson et al., 2010; Mehlitz et al., 1982; Funk et al., 2013) . This is why the majority of HAT infections are caused by *T.b. gambiense*. It does have a reference genome, DAL972, however Tb927 is generally favoured over this, due to lack of annotation and manual finishing for gaps in the genome (Pays and Vanhollebeke, 2008; Berriman, 2005; Jackson et al., 2010).

1.5.7 *T.brucei. rhodesiense* is the causative agent of eastern and southern African trypanosomiasis

T.b. rhodesiense causes African trypanosomiasis in eastern and southern Africa. Similarly to *T.b. gambiense*, *T.b. rhodesiense* has evolved a strategy to combat the effects of ApoL1 binding. However, as illustrated in Figure 1.1, it uses the serum resistance associated protein (SRA), which is a truncated variable surface glycoprotein (VSG) like gene, however it only has less than 25% sequence homology (Xong et al., 1998; Pays and Vanhollebeke, 2008; Vanhamme et al., 1999). SRA is located within the polycistronic transcription units upstream of the VSGs in an active expression site, and is known as an expression site associated gene (ESAG) (Vanhamme et al., 2003; Xong et al., 1998; Shiflett et al., 2007; Vanhamme et al., 1999). Similarly to *T.b. gambiense*, *T.b. rhodesesiense* acts on ApoL1 to prevent lysis of the parasite. SRA does this by binding to the TLF complex once it has been trafficked to the endosome. The majority of the SRA protein is localized within the lysosome between the flagellar pocket and nucleus. Once the TLF complex is in close proximity to SRA, it binds to ApoL1 at its SRA interaction domain, which prevents its release from the complex and lysis of the parasite (Vanhamme et al., 2003; Shiflett et al., 2007).

Unlike *T.b. gambiense*, *T.b. rhodesiense* is generally considered to cause a more severe infection, with several characteristics, such as a tendency to cause a chancre, and a much quicker progression from the early bloodstream stages of the disease to the encephalitic stage. *T.b. rhodesiense* also infects both human and animal, however unlike *T.b. gambiense*, the majority of infections occur in the animal not human reservoir, with human infections considered more coincidental (Radwanska et al., 2002; Onyango et al., 1966; Picozzi et al., 2008; Hide et al., 1996; Noireau et al., 1989; Smith and Bailey, 2000; Anderson et al., 2011). Unlike *T.b. gambiense* and *T.b. brucei*, *T.b. rhodesiense* does not have a reference sequence, however due to the high degree of similarity between all three sub-species, the Tb927 reference is adequate for comparison.

1.6 *T. brucei's* subspecies are typically geographically isolated

As previously mentioned, the burden of disease in terms of HAT infections recorded annually has dramatically decreased within the last decade. However the incidence of non-human infections still results in a considerable socio-economic loss. Morphologically all three *T. brucei* sub-species are indistinguishable by microscopy, however *T.b gambiense* and *T.b. rhodesiense* infections primarily occur in different regions of Sub-Saharan Africa, and are only co-endemic in Uganda, which means that a strain's sub-species can often by identified by the location it was first isolated (World Health Organization, 2014; Radwanska et al., 2002; Picozzi et al., 2008). In contrast, AAT infections occur throughout Sub-Saharan Africa, and are co-endemic with both *T.b. gambiense* and *T.b. rhodesiense*.

In 2014, the world health organization (WHO) recorded that the most common sub-species, *T.b. gambiense*, is endemic in 24 countries within west and central Africa and was responsible for ~97% of reported HAT infections (World Health Organization, 2014). *T.b. rhodesiense* was found to be endemic in 13 countries in eastern and southern Africa and only represented ~3% of recorded cases. The incidence of HAT caused by both sub-species has fallen by greater than 70% between 1999-2014, with the reported cases of new *T.b. gambiense* infections falling from 27, 892 to 3,679, and 619 to 117 in *T.b. rhodesiense* (Simarro et al., 2010; World Health Organization, 2014; Simarro, 2011). This is a result of a large collaborative effort by the WHO and help from multiple non-government organizations (NGOs) in response to the resurgence of HAT

in the 1970s, when approximately 300,000-500,000 were infected (Simarro et al., 2010; Bucheton et al., 2011; Simarro, 2011).

The localization of *T.b. gambiense* and *T.b. rhodesiense* is shown in Figure 1.2, and shows *T.b. gambiense* to affect a larger region of Sub-Saharan Africa, and a higher density of infections, with countries recording over 1000 cases a year. Regions where infections were recorded in previous years, but where none were recorded within 2014 are shown in red. In particular, *T.b. gambiense* infections have effectively disappeared from savannah regions as a direct effect of active surveillance and treatment (Picozzi et al., 2005; Simarro et al., 2010; Bucheton et al., 2011). All these countries fall within the tsetse belt, which is the main limiting factor in the distribution of HAT infections. The line marked on Figure 1.4 represents the approximate boundaries between *T.b. gambiense* and *T.b. rhodesiense* infections.

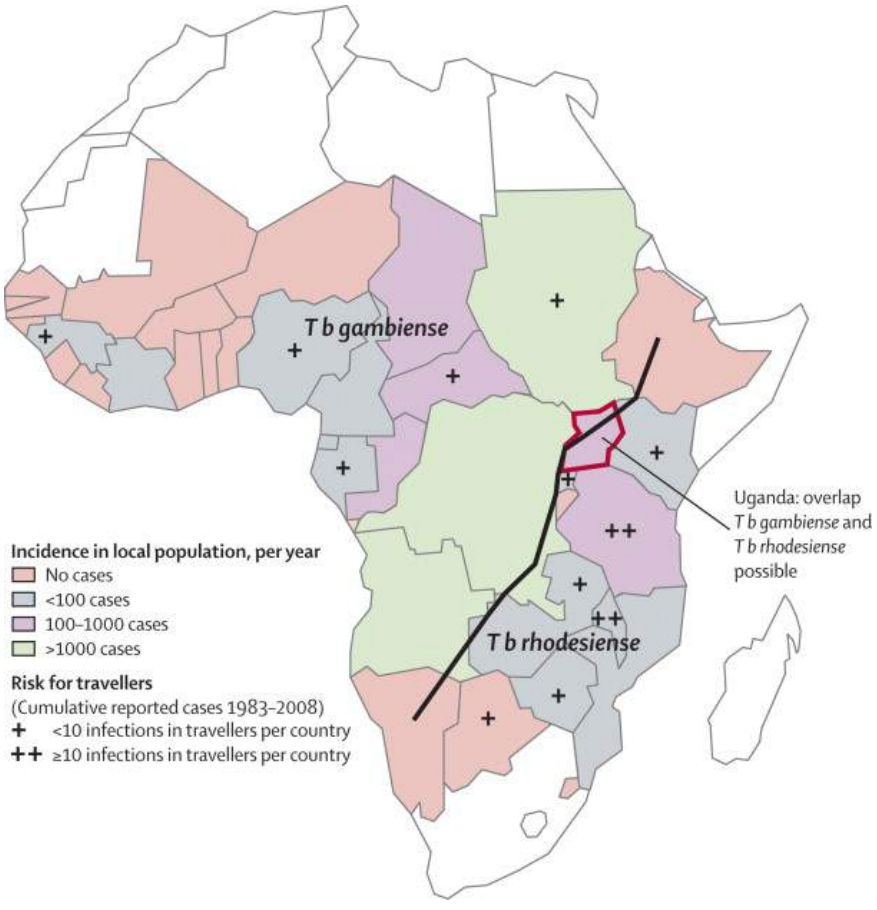

Figure 1.4: Taken with permission from Brun *et al.*, 2010. From " Human African trypanosomiasis". This shows T.b. gambiense infections occur within western and central Africa, and *T.b. rhodesiense* infections occur within eastern Africa. The line drawn represents the approximate divide between both subspecies, however Uganda has an overlap between *T.b. rhodesiense* and *T.b. gambiense* infections. In the savannah regions, no cases were recorded in 2014, where previously these regions were endemic.

1.6.1 Co-existance of both human infective sub-species in Uganda gives rise to the possibility of recombination between sub-species

As shown in Figure 1.4, Uganda is the only country where both *T.b. gambiense* and *T.b. rhodesiense* are known to co-exist (Bucheton et al., 2011; Picozzi et al., 2005; Jamonneau et al., 2012). Due to the phenotypes of group 2 *T.b. gambiense* strains and *T.b. rhodesiense* strains being more variable, this led to the idea of mating between and within sub-species leading to hybrid parasites with a more variable phenotype, as previously mentioned. The strains used within this body of work are *T. b. rhodesiense* strains originally isolated from Uganda. The phenotypes of these strains have been previously described and suggested introgression between *T. brucei* subspecies was associated with differences in virulence, as described in detail in section 1.11 (Goodhead et al., 2013).

1.6.2 Subpopulations in endemic countries contain mutations which confer trypanotolerant advantages

Trypanotolerance and susceptibility has been recorded both in human and animal populations (Murray et al., 1982; Bucheton et al., 2011; Naessens, 2006; Jamonneau et al., 2012; O'Gorman et al., 2009). In cattle, the trypanosusceptible Boran and trypanotolerant N'Dama cattle breeds have been extensively studied in order to understand the mechanisms of trypanotolerance, and to breed trypanotolerant cattle suitable for meat and milk production (Jamonneau et al., 2012; Murray et al., 1982; Naessens, 2006; O'Gorman et al., 2009). Until recently, little research effort has been placed into understanding the mechanisms of resistance in across human populations within these endemic regions, and this is the premise of the TrypanoGen project (http://trypanogen.net).

Host-parasite interactions are key to determining the outcome of an infection, and not only do the parasites vary in their virulence, but the host also varies in is ability to cope with the infection. HAT infections were considered invariably fatal without treatment until recently (Bucheton et al., 2011; Jamonneau et al., 2012). However small populations of infected persons in these endemic regions were found to spontaneously recover without medical intervention, or were asymptomatic despite being tested positive for parasite infection using one of the methods discussed in the treatments

section of this chapter (Drain et al., 2001; Bucheton et al., 2011; Jamonneau et al., 2012).

1.6.3 The effect of co-morbidities on trypanotolerance

There is an overlap in the regions where both malaria and trypanosomiasis are endemic. As previously discussed, low haptoglobin levels are correlated with an increase in the production of TLF complexes (Uzureau et al., 2013; Drain et al., 2001). In regions of malaria infection, haptoglobin levels are generally low as a result of haemolysis caused by *Plasmodium* when it releases merozoites into the blood. The subsequent rupturing of the erythrocytes causes the release of free haem into the blood, which is removed by haptoglobin (Uzureau et al., 2013). The increase in TLF complex formation due to low haptoglobin levels is protective against trypanosome infection. Co-morbidities can also make the reporting of HAT cases more difficult because particularly in the earlier stages of the disease, which can range in symptoms from asymptomatic to general signs of malaise, these non-specific symptoms can be easily mistaken for malaria or other diseases prevalent in that region.

1.6.4 Surveillance and the current methods for sub-species identification

As previously mentioned surveillance, i.e the detection and identification of the specific sub-species, was crucial in the efforts made by WHO and various NGOs to significantly reduce the health burden of HAT (Kibona et al., 2007; Simarro et al., 2010; Simarro, 2011). Although these sub species are typically geographically isolated, as previously mentioned, there are regions of overlap between *T.b. gambiense* and *T.b. rhodesiense* in Uganda, and migration between regions can lead to the introduction of sub-species in locations where they have not previously been identified, as has been suspected previously in Tanzania (Magnus et al., 1978; Kibona et al., 2007). One of the issues in determining the *T. brucei* subspecies present in an infection is the variability in the manifestation of the disease, particularly in group 2 *T.b. gambiense* and *T.b. rhodesiense* strains. Another is that all three subspecies are morphologically indistinguishable, and so alternative detection methods are required to elucidate whether a HAT infection is caused by *T.b. gambiense* or *T.b. rhodesiense*. The determination of the sub species present is also important because it determines the drug regimen used to treat the infection.

At present there are several methods of detection and they vary in their sensitivity. The card agglutination test (CATT) is a commonly used method of identifying *T.b. gambiense*, an example of which is shown in Figure 1.5. Since its development in 1978, the CATT test has been used widely in endemic areas because of its inexpensive and easy to use nature (Truc et al., 2002; Magnus et al., 1978; Chappuis et al., 2004). Despite a high sensitivity, which detects 90% of serologically positive patients, confirmation of the diagnosis subsequent to the test is required and relies on detection of trypanosomes either in the blood or via lumbar puncture (Truc et al., 2002; Picozzi et al., 2002; Chappuis et al., 2004). However not all serologically positive patients are identified as positive by microscopic methods and due to the serious side effects often exhibited by patients receiving treatment, a patient must be both serologically and parasitologically confirmed (Truc et al., 2002; Picozzi et al., 2002). The success of the CATT test has led to the development of other related diagnostic methods such as the micro-CATT and the latex agglutination test (LATEX), however the original CATT test remains the preferred method (Radwanska et al., 2002; Kayang et al., 1997; Njiru et al., 2004; Truc et al., 2002; Ng'ayo et al., 2005; Picozzi et al., 2008).

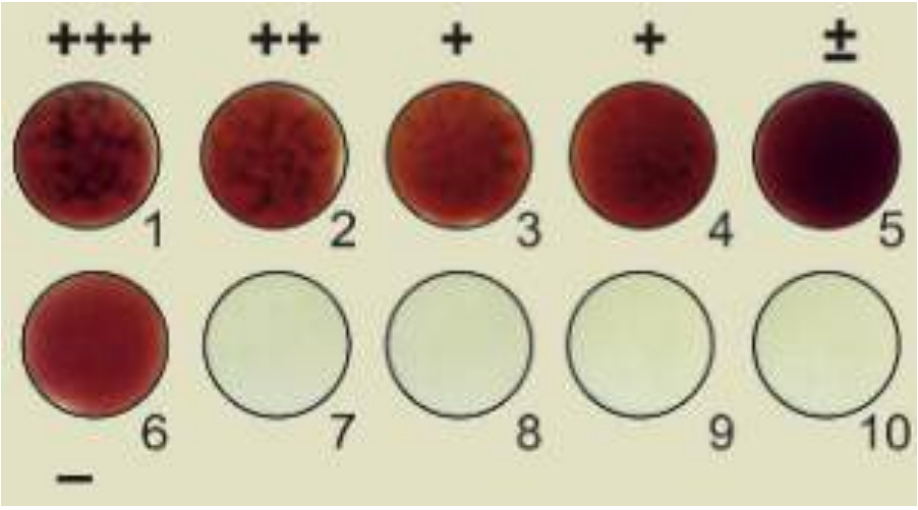

Figure 1.5: Adapted from Dhaliwal & Juyal, 2013. An example of a CATT test. Blood is serially diluted, each dilution shown on a subsequent number. When positive for parasites, the blood will agglutinate. In non diluted/less diluted samples, this agglutination will be more apparent as shown by the strong agglutination in 1, and the lower agglutination observed in 2-5.

Due to lack of specificity in the parasitological detection methods, effort was put into the development of more sensitive methods. These are primarily PCR based and markers such as SRA and TgsGP can be used for example for identifying *T.b. rhodesiense* and *T.b. gambiense* group 1 strains. SRA has been used for the detection of *T.b. rhodesiense* in cattle in multiple studies (Picozzi et al., 2002; Radwanska et al., 2002; Njiru et al., 2004; Ng'ayo et al., 2005; Picozzi et al., 2008). PCR methods have also been used to amplify trypanosome specific regions such as the expression site associated genes, which allows for detection prior to symptoms being displayed or microscopic detection (Zhang and Baltz, 1994; Picozzi et al., 2002; Kanmogne et al., 1996; Afework et al., 2006).

1.7 Previous methods of assigning taxonomy and analysing population structures have been superseded by sequencing

Beneath is a brief explanation of the methods that have been used previously for assigning taxonomy in T. brucei strains. Within this project, the strains used were assigned to zymodeme groups, which were generated using multilocus enzyme electrophoresis (MLEE), which is described in more detail in section 1.7.1.4. However these methods have been largely replaced by comparative genomic approaches.

1.7.1.1 Restriction fragment length polymorphisms (RFLPs)

RFLP analysis looks at differences in homologous DNA sequences at enzyme restriction sites. The DNA is digested by a combination of restriction enzymes, and the resulting fragments are separated by electrophoresis. Multiple previous studies have used these banding patterns to assign *T. brucei* isolates to groups and infer the population structure (Geysen et al., 2003; Zhang and Baltz, 1994; Kanmogne et al., 1996; Afework et al., 2006). RFLP-PCR has also been used to identify subgroups within trypanosome-infected cattle, and could distinguish between *T. congolense*, *T. brucei*, *T. vivax* and *T. theileri*. However the main limitation with this technique is that it is often incapable of distinguishing between highly similar species or the subgroups within a species (Tilley et al., 2003; Geysen et al., 2003).

1.7.1.2 Mobile genetic element PCR (MGE-PCR)

Mobile genetic element PCR (MGE-PCR) was used similarly to RFLP analysis on *T. brucei* strains to assign them to taxonomic groups. However this utilized RIME, a mobile genetic element. There is a higher degree of variability between strains using this method, and so individual strains and genotypes could be resolved (Tilley et al., 2003).

1.7.1.3 Microsatellites and minisatellites

Tandem repeat sequences (micro and mini satellites) were also used for strain identification. Microsatellites consist of 2-5 nucleotide length sequences that are repeated typically between 5-50 times. Minisatellites are larger and the repeat sequences are generally 6-100 nucleotides long and are repeated 5-50 times also (Sloof et al., 1983; Vergnaud, 2000; Koffi et al., 2007; Simo et al., 2010; 2011). Collectively, minisatellites and microsatellites are called variable number tandem repeats (VNTR). Similarly to the previous methods, multiple studies were done in *T. brucei* using these methods, however these techniques have been largely replaced by comparative genomic studies (Sloof et al., 1983; Koffi et al., 2007; Simo et al., 2010; 2011).

1.7.1.4 Isoenzyme and multilocus enzyme electrophoresis (MLEE)

Isoenzyme analysis as a method of strain characterization was first described in 1993 (Ben Abderrazak et al., 1993). Isoenzyme analysis uses the presence of the isoenzymes present in a strain, in conjunction with multilocus enzyme electrophoresis (MLEE) to generate banding patterns unique to a strain or set of strains. Typically, these isoenzymes exhibit Mendelian inheritance traits, with active isoenzymes representing particular loci which behave as alleles. These are often referred to as alloenzymes (Banuls et al., 1999; Ben Abderrazak et al., 1993; Soccol et al., 2002; Montilla et al., 2002). These isoenzyme patterns can be used to estimate strain relatedness and understand population structure. The strains used within this work were allocated to zymodeme groups using this method, and the isoenyzme banding patterns were sensitive enough to group strains within the *T.b. rhodesiense* subspecies. This method of characterization was highly popular, and was used to decipher population structure

in multiple studies (Banuls et al., 1999; Soccol et al., 2002; Montilla et al., 2002; Tibayrenc, 1998).

1.8 HAT infections result in a spectrum of symptoms but trypanosomiasis can be defined in two stages

Due to host-parasite interplay, a whole range of symptoms and patient outcomes can arise as a result of HAT infection (Nikolskaia et al., 2006; Giroud et al., 2009; Brun et al., 2010; Sternberg and Maclean, 2010). However, the diseases natural progression is from an early stage when the parasites are within the bloodstream, followed by a late stage which is characterized by the parasites passing through the blood brain barrier (BBB) (Jamonneau et al., 2012; Nikolskaia et al., 2006; Brun et al., 2010). Both stages exist in *T.b. gambiense* and *T.b. rhodesiense* infections, however depending on the severity of the disease a patient may be asymptomatic for a long period, or may present with a severe early stage infection. Depending on the time from infection and the severity of the disease, different drug regimens can be given and if treated early in infection, patients do not necessarily go on to develop the late stage of the disease. Until recently HAT infections without treatment were considered invariably fatal, however spontaneous recovery has been recorded, but this is considered very rare (Brun et al., 2010; Jamonneau et al., 2012; Kennedy, 2013).

In the early stage of infection, depending on the virulence of the strain, localized swellings/chancres can occur around the bite region. Other recognizable symptoms include winterbottom's sign, which is a result of lymphadenopathy, facial oedoma, muscle weakness and general malaise (Kennedy, 2013; Brun et al., 2010). The parasites go through several peaks of parasitaemia and these correlate with relapsing fever. The non-specificity of the symptoms often make initial diagnosis difficult and can lead to misdiagnosis with other endemic diseases causing similar symptoms. At this stage the parasites remain in the bloodstream and lymph and so this is often referred to as the haemo-lymphatic stage of infection (de Atouguia and Kennedy, 2000; Kennedy, 2013; 2004).

The late stage of infection is diagnosed by lumbar puncture and identification of trypanosomes within the cerebral-spinal fluid (CSF). This is the stage of the disease from which the disease derives its name, sleeping sickness. At this stage, parasites pass

through the BBB and can enter the tissues here and interfere with normal neuronal function (Kennedy, 2004; de Atouguia and Kennedy, 2000). As a result, the patient's cognitive ability may be impaired in multiple ways. Daytime sleeping is one of the common results of the late stage, in which between relapsed fevers, patients sleep into atypical sleeping patterns, which without treatment typically lead to coma. In part this is caused by the metabolism of tryptophan, which induces sleep in the patients (de Atouguia and Kennedy, 2000; Kennedy, 2004). The suprachiasmatic nuclei, which control circadian rhythms in humans, are dysregulated in *T. brucei* infections, and this also affects the neurological symptoms observed. This stage is also referred to as the encephalitic stage (Brun et al., 2010; de Atouguia and Kennedy, 2000). At this stage other major organs begin to fail, including commonly the liver, and other symptoms which may also be present include splenomegaly, amenorrhoea and severe anaemia (Steverding, 2010; Brun et al., 2010). Due to the severity of the late stage of the disease and the multiple organs it affects, treatment prior to the passing of the BBB is both preferential and more successful. Due to the chronic nature of *T.b. gambiense* infections, typically symptoms don't occur until months or years after infection. In contrast, in *T.b. rhodesiense* infections the disease manifests often occurs in weeks.

1.8 Trypanosomiasis treatment options are reliable but limited

Treatment for HAT is severely limited, dependent on drugs that were discovered empirically decades ago, and their mechanisms of action not well understood (Fairlamb, 1990; Steverding, 2010; Pépin et al., 1994; Keating et al., 2015). The need for novel treatments is also urgent due to the low efficacy, increasing resistance and severe side effects of these current treatments (Kennedy, 2004; Fairlamb, 1990; Pépin et al., 1994; Keating et al., 2015). In *T. brucei*, treatment of the late stage is difficult to treat in particular, because the drug needs to be able to pass through the blood-brain barrier (Kennedy, 2004).

Currently there are a small number of reliable treatments, however for the late stage especially, treatment is more complex and less successful and resistance is increasing whilst there is no current prospect of a vaccine (Masocha and Kristensson, 2012; Kennedy, 2004). Very few pathogens are capable of passing the blood brain barrier, and antibodies are too large to pass (Masocha and Kristensson, 2012; Barrett et al., 2007). This means that once *T. brucei* is within the CNS, it is protected from the

immunological challenges it faces as an extracellular parasite in the bloodstream. However due to the impenetrability of the BBB, once *T. brucei* reaches this stage, the damage to the host is typically severe, and finding drug treatments capable of also passing the BBB is difficult. There are currently four main drugs used in the treatment of trypanosomiasis, they are discussed briefly below. Due to the limited number of drug regimen options, treatment is selected by causative agent and stage, as shown in Table 1.1.

1.8.1 Pentamidine

Pentamidine is used in the treatment of early stage *T.b. gambiense* infections and is given either intramuscularly or intravenously. Compared to some of the other treatment options, pentamidine is generally well tolerated, however intramuscular injections can still lead to symptoms such as gastrointestinal problems and hypoglycaemia in up to 40% of patients (Doua et al., 1996; Benaim et al., 1993; Barrett et al., 2007). More serious side effects such as reductions in thrombocytes and leucocytes are also seen, but are far less common (Doua et al., 1996; Walter and Albiez, 1981; Barrett et al., 2007).

1.8.2 Suramin

Suramin is used in the treatment of first stage *T.b. rhodesiense* infections. However it is not used to treat stage one *T.b. gambiense* infections because *Onchocerca* species are also endemic in these regions. *Onchocerca* species elicit a high immune response when exposed to suramin, and so suramin is not used to reduce the risk of severe allergic reactions in patients (Anderson, 1976; Barrett et al., 2007; Walter and Albiez, 1981). The administration of suramin is more complicated and treatment lasts up to 30 days. Drug reactions to suramin are more frequent than pentamidine, but are usually much milder (Babokhov et al., 2013; Barrett et al., 2007).

1.8.3 Melarsoprol

Melarsoprol is used for the treatment of both second stage *T.b. gambiense* and *T.b. rhodesiense*. Although the most toxic of treatments due to it being an arsenic derivative, it is the only available treatment currently for late stage *T.b. rhodesiense* (Priotto et al., 2006; Babokhov et al., 2013; Chappuis, 2007). Eflornithine or eflornithine in

combination with nifurtimox are preferable treatments for late stage treatment of *T.b. gambiense*, however they are unaffordable in some of the poorer endemic countries (Babokhov et al., 2013; Priotto et al., 2006; Chappuis, 2007). For *T.b. gambiense*, the course of treatment is shorter, however for *T.b. rhodesiense* the treatment course is longer and more complex. Adverse drug reactions to melarsoprol are frequently serious, if not life-threatening (Gehrig and Efferth, 2008; Babokhov et al., 2013). Post treatment encephalopathic syndrome is life threatening, and occurs in ~5% of *T.b. gambiense* infected patients and ~8% of *T.b. rhodesiense* infected patients (Gehrig and Efferth, 2008). Subsequent treatment of the encephalopathic syndrome also often results in adverse skin reactions (Kennedy, 2013; Gehrig and Efferth, 2008). Treatment failures of up to 30% also suggest resistance to melarsoprol (Kennedy, 2013; Priotto et al., 2007).

1.8.4 Eflornithine and nifurtimox combination therapy

Eflornithine is the newest treatment for trypanosomiasis, and the only one to be developed within the last 50 years. It works by inhibiting ornithine decarboxylase, which is essential for differentiation and replication (Babokhov et al., 2013; Bacchi et al., 1983; Priotto et al., 2007). Eflornithine can be used in place of melarsoprol for late stage *T.b. gambiense* treatment, and has been shown to both be effective and have a significantly reduced mortality rate after treatment (Babokhov et al., 2013). However eflornithine cannot be used for *T.b. rhodesiense* treatment because the parasite is less susceptible (Priotto et al., 2009; Babokhov et al., 2013). For *T.b. gambiense* infections treatment lasts two weeks, but multiple infusions per day are required, limiting the use at more rural centres. Nifurtimox can also be used in conjunction with eflornithine to increase treatment success. Nifurtimox was originally used only as an orally administered treatment for Chagas disease (Priotto et al., 2007; 2009). For both eflornithine solely, and in combination treatment, the side effects are similar to those with pentamidine, gastrointestinal issues and altered blood counts amongst others (Priotto et al., 2007).

Table 1.1: This shows the current drug treatments available for HAT. The drug regimen selected is dependent on the pathogen species and the stage of disease. Melarsoprol can be used for *both T.b. gambiense* and *T.b. rhodesiense* in late stages, however eflornithine or eflortnithine and nifurtimox combination therapy is preferred due to the toxicity of melarsoprol.

	Stage	
Sub-species	Early	Late
T.b. gambiense	Pentamidine	Eflornithine on own or combined with nifurtimox
		Melarsoprol
T.b. rhodesiense	Suramin	Melarsoprol

1.9 *T. brucei* has an unusual genome structure comprising on large chromosomes, intermediate sized chromosomes and small circularized DNA fragments

T. brucei's genome consists of 11 paired megabase sized chromosomes, several intermediate sized chromosomes which are less than a megabase in size, and approximately a hundred minichromosomes which are between 50-100kilobase pairs (Daniels et al., 2010; Berriman, 2005). The larger chromosomes are diploid in nature, but the intermediate and minichromosomes appear to be aneuploid (Berriman, 2005; Daniels et al., 2010). Initial analysis of the sequenced large chromosomes in 2005 suggested *T. brucei* had approximately 9,068 protein coding genes and 904 pseudogenes spanning over 26 megabases in the large chromosomes (Berriman, 2005). These numbers have increased to 10,110 and 1,461 respectively in version 9 of Tb927's genome (www.genedb.org). A large proportion of the protein coding genes are located within these large chromosomes (Berriman, 2005; Daniels et al., 2010).

1.9.1 The large linear chromosomes of Tb927 were first sequenced in 2005

T. brucei was first sequenced in 2005 in draft form, and has been manually finished to remove gaps and improve overall coverage. The *T. brucei* strain sequenced was Tb927, a *T.brucei brucei* strain (Peacock et al., 2008; Berriman, 2005; Jackson et al., 2010). Since then, other *T. brucei* strains have been sequenced including DAL972 in 2010, however Tb927 is generally used as a reference strain because unlike some *T. brucei* strains such as *T. brucei* Lister 427, Tb927 can complete all stages of *T. brucei's* natural life cycle (Berriman, 2005; Peacock et al., 2008; Jackson et al., 2010). DAL972 is a *T.b. gambiense* reference, however this is far more fragmented than the Tb927 reference (Berriman, 2005; Jackson et al., 2010). Despite being different subspecies, *T.b. brucei*, *T.b. rhodesiense* and *T.b. gambiense* can all be mapped to the Tb927 reference because they share a greater than 99% genome identity. Due to the unusual structure of *T.*

brucei, the 11 paired megachromosomes were sequenced first and made publicly available (Forrester and Hall, 2014; Berriman, 2005). *T. brucei* is also easily amenable to bioinformatic analysis due to its diploid nature.

Sequencing the mega base chromosomes unveiled several features of *T. brucei's* genome, which were hard to decipher using more traditional genetic techniques alone (Forrester and Hall, 2014). Due to its extracellular nature, *T. brucei* spends the entirety of its life cycle evading the host's immune system and dedicates a high percentage of its genome to antigenic variation. Shotgun sequencing also gave significant insight into the organization of the genome and the potential transcriptional and regulatory differences, which arise as a result. The impact of sequencing on not only *T. brucei*, but on evolving the field of parasitology as a whole was reviewed by Forrester & Hall in 2014 (Forrester and Hall, 2014). Key insights and contributions to the parasite genomic field resulting from the publication of the Tb927 genome and the sequencing of other parasitic genomes are briefly discussed below.

1.9.2 A large proportion of *T. brucei's* genome is dedicated to antigenic variation

As previously mentioned, *T. brucei* dedicates a high percentage of its genome to antigenic variation, approximately 20%, in order to evade detection in the mammalian host (Aitcheson et al., 2005; Donelson, 2003). In *T. brucei,* antigenic variation is generated by variable surface glycoprotein (VSG) genes (Pays et al., 2001; Aitcheson et al., 2005). These are primarily located within the subtelomeric regions of the chromosomes and a catalogue of over a 1000 VSG genes can be activated (Aitcheson et al., 2005; Pays et al., 2001). The activation of these genes changes the glycoprotein covering the trypanosome surface. The host immune system can mount a response against these surface antigens, however only one VSG is active at one time, and so once a host response is primed, VSG switching occurs. This silences the expression of the previously expressed VSG gene and activates the expression of another (Berriman et al., 2002; Aitcheson et al., 2005). This activation happens within polycistronic transcriptional units within bloodstream expression sites (BES), which contain multiple expression site associated genes (ESAGs) and VSGs downstream (Pays et al., 2004; Berriman et al., 2002). VSGs can be activated by either the transposition of a silent VSG into an active expression site, rearrangements between telomeres or mediated by a change in transcriptional regulation (Berriman et al., 2002; Pays et al., 2004; Forrester and Hall, 2014). The structure of BESs had been described prior to the

release of Tb927's genome, however the variations in structure and the number of BESs could only be elucidated using the sequencing data (Berriman et al., 2002; Forrester and Hall, 2014). Sequencing also demonstrated that only 5% of these VSGs were functional, the rest are pseudogenes (Marcello and Barry, 2007). The mechanisms of VSG activation and switching and the structure the BESs have previously been reviewed (Hall et al., 2003; Barry and McCulloch, 2001; Donelson, 2003; Pays et al., 2004).

1.9.3 Shotgun sequencing illustrated a high degree of synteny cross species and conformity in genome organization

Shotgun sequencing also gave further insight into chromosome structure, organization and the positioning of housekeeping genes. Early chromosome papers for publications for *Plasmodium* and *T. brucei* (Tachibana et al., 2012; Hall et al., 2003; Bowman et al., 1999; Gardner et al., 2002) revealed that housekeeping genes were located in central regions of the genome and important antigen gene families were at the sub-telomeres with common complex DNA repeat units. In *Plasmodium*, the genomes of multiple *Plasmodium* species have been published and have been used to reveal much about how these species have adapted to their hosts and have again illustrated that the genomes have a highly syntenic core but vary at the telomeres (Tachibana et al., 2012).

Similar projects, such as the sequencing of the "tritryps" genomes again reaffirmed the idea of a very similar core gene content despite differences in their lifestyles and highly divergent subtelomeres that contained many of the surface antigens (El-Sayed et al., 2005b; Berriman, 2005; Parsons et al., 2005; El-Sayed et al., 2005a; Ivens et al., 2005). The sequencing of the other trypanosomatid genomes, *Leishmania major* and *Trypanosoma cruzi* also in 2005 also enabled cross species comparisons to be made and putative housekeeping genes to be identified (Parsons et al., 2005; El-Sayed et al., 2005b). Genes specific to *T. brucei* were also used to identify genes responsible for antigenic variation and enabling an extracellular lifestyle.

As expected, the lifestyles of these parasites are reflected in their genome, with intracellular parasites *L.major* and *T.cruzi* displaying a greater degree of similarity in comparison to extracellular parasite *T. brucei* (El-Sayed et al., 2005b; Parsons et al., 2005). These genomes allowed us to look at how differences in immune evasion strategies are reflected in the accessory parts of the genome. Due to its extracellular

lifestyle, *T. brucei* had the highest abundance of species-specific surface antigens proteins (El-Sayed et al., 2005b).

1.9.4 Genes are organized to enable polycistronic transcription

Access to the genome also revealed details about several key mechanisms that were previously not wholly understood, for instance transcriptional regulation in the kinetoplastids. The kinetoplastids have an unusual genome structure, with genes organized into long polycistronic arrays which undergo trans-splicing to excise genes from exons and attach a splice leader to each transcript (Parsons et al., 1984; El-Sayed et al., 2003). This is a conserved mechanism, which has been documented and reviewed in a variety of eukaryotes including nematodes, platyhelminthes and tunicates (Forrester and Hall, 2014; Pettitt et al., 2008; Vandenberghe et al., 2001). Although previously observed in *T. brucei*, it had not been explored on a genome wide scale prior to shotgun sequencing (Nilsson et al., 2010; Forrester and Hall, 2014).

In 2010, Nilsson et al used this knowledge in order to sequence *T. brucei* transcripts containing this splice sequence in order to discover splice site locations, understand alternative splicing events, and analyze the effect they had on gene expression (Nilsson et al., 2010). Sequencing also highlighted the scale of gene expansion in some gene families by tandem duplication, which compensates for a lack of transcriptional control (El-Sayed et al., 2005b).

Annotation of the Tb927 genome has also been pivotal in understanding metabolic processes present in *T. brucei* from the genes found to be present. Parasites typically reduce their genome size to only include genes encoding for essential functions, which often leads to parasite depending on the host for the provision of certain products (Sakharkar et al., 2004).

1.10 DNA sequencing methodologies

1.10.1 Sanger sequencing was the method of choice between the 1980s-mid 2000's

Two sequencing methods, the Maxam-Gilbert and the Chain-termination methods, were developed in 1977 to decipher short DNA sequences (Maxam and Gilbert, 1977; Sanger et al., 1977). However the chain-termination method developed by Frederick Sanger quickly became the preferred method and dominated sequencing until the development of next generation sequencing methods in 2005 (Hutchison, 2007). Advances in capillary sequencing enabled the automation of this process and made whole genome projects feasible. Sanger sequencing is still popular to date for small scale projects however with the introduction of next generation sequencing and subsequent advances to the technology, research has shifted towards much higher throughput sequencing (Lander et al., 2001; Hutchison, 2007). This has been coined Sanger sequencing and is still used today for smaller projects involving the sequencing of smaller products for example for PCR products, or for validating observations seen in next generation sequence (NGS) data.

However prior to NGS, Sanger sequencing was developed to increase throughput through improvements such as increase the number of capillaries, increased read length and by increasing the automation of the process. This led to many significant milestones using Sanger sequencing such as many of the first whole genome projects including the first human genome in 2001 (McCourt et al., 2013; Lander et al., 2001). Sanger sequencing also has a much lower error rate than seen by NGS technologies, hence its use in NGS validation (Forrester and Hall, 2014; McCourt et al., 2013).

1.10.2 New high throughout technologies and the start of the genomic era

Next generation sequencers are capable of much higher throughput because they are 'massively parallel', generating milllions of shorter reads instead of the long individual reads produced by Sanger sequencing (Forrester and Hall, 2014). These technologies have heavily invested in the development of chemistries that can produce longer and greater numbers of reads. Typically, Illumina reads are now at least 100 base pairs in length, however initially reads lengths were as short as 21 bases, whereas Sanger

sequences can generate upto 1kb length reads (Margulies et al., 2005; Forrester and Hall, 2014).

The first commercially available next generation sequencer was developed by 454 Life sciences and used pyrosequencing (Mardis, 2008; Margulies et al., 2005). This method used emulsion PCR, in which DNA was amplified inside water droplets containing a single DNA template attached to a single bead, in an oil solution. Luciferase was then used to decipher each nucleotide added. This technology was cheaper per base compared to Sanger sequencing, and gave much longer read lengths than other methods such as SOLiD and Illumina of up to 700 bases. However this technology was quickly surpassed by Illumina and SOLiD technology, due to cheaper runs and higher data yields (Bentley et al., 2008; Mardis, 2008).

In Illumina sequencing, DNA and primers are attached to a slide and amplified with a polymerase to form clusters (Bentley et al., 2008). The sequence is then determined by using reverse terminator bases, washing non-incorporated nucleotides away and imaging the fluorescently labeled nucleotides (Bentley et al., 2008). In SOLiD sequencing, the DNA is amplified by emulsion PCR prior to sequencing. Fluorescently labeled probes then compete to ligate to the sequencing primer (Rusk, 2011; Valouev et al., 2008). Both Illumina and SOLiD sequencing produce data with short read lengths, however Illumina can potentially generate a greater number of reads per run, and with new chemistry Illumina reads can now be upto 2 x 150 and 2 x 300 on Hiseq and Miseq models respectively (www.Illumina.com/systems.html).

Unlike 454, SOLiD or Illumina sequencing, ion torrent's sequencer uses hydrogen ion detection instead of fluorescence, to determine the incorporation of a specific base into the sequence. Here a microwell containing the template is flooded with a single type of nucleotide, which if incorporated, causes the release of a hydrogen ion (Mamanova et al., 2010; Rusk, 2011). Table 1.2 gives a summary of the aforementioned sequencing methods.

As was seen previously with Sanger sequencing, Illumina sequencing currently dominates the sequencing field with 454, SOLiD and ion torrent technologies largely becoming redundant. Sequencing methodologies are advancing towards the development of more tailored sequencing applications, and are moving away from the traditional methods of sequencing an individual genome. Examples of newer

sequencing strategies include the sequencing of multiple strains to infer population structure, enrichment sequencing to resolve target organisms from a mixture of genomes and methylation studies (Fonseca et al., 2012; Mamanova et al., 2010; Ellegren, 2014).

Table 1.2: This compares traditional Sanger sequencing with four next generation sequencing technologies on the basis of read length, accuracy, number of reads and time per run. Illumina is the forerunner in next generation sequencing currently, with SOLiD, 454 and ion torrent becoming redundant, in part this is due to individual sequence error issues such as palindromic sequences and homopolymer errors.

Method	Read length	Accuracy (%)	Reads per run	Time per run	Advantages	Disadvantages
Chain termination (Sanger sequencing)	Up to 1kb	99.9	N/A	Up to 3 hours	Long reads. Most accurate	Most expensive per base cost, more time consuming, least throughput
Next generation sequencers						
Ion torrent	Up to 400bp	98	Upto 80 million	2 hours	Less expensive. Fast runs	Homopolymer errors
Pyrosequencing (454)	700bp	99.9	1 million	24 hours	Long reads. Fast	Expensive per base cost, homopolymer errors
Sequencing by synthesis (Illumina)	50-300bp	99.9	Up to 6 billion	1-11 days, dependent on run and instrument type	Highest sequence yield	Expensive equipment and consumables. Needs high DNA concentrations
Sequencing by ligation (SOLiD)	35-50bp	99.9	1.2-1.4 billion	1-2 weeks	Low cost per base	Difficulty sequencing palindromic sequences

1.10.3 The generation of increasing volumes of data poses analytical and computational issues

Fonseca et al., 2012 highlighted the wealth of freely available bioinformatic tools, with more than 60 mapping tools currently available, the majority post 2008 (Fonseca et al., 2012). This reflects how sequencing has grown as a technology and diversified, and so tools have either had to be adapted to be or created to deal with these changes. Key software changes mark important milestones in sequencing, such as the move from

SOLiD's colourspace data to primarily Illumina sequencing. There are a whole variety of other OMIC applications too such as methylation sequencing, transcriptomics and more recently Pacbio data, all of which need bioinformatic software capable of dealing with these differently formatted data. Short read assembly in particular is computationally expensive due to the ever increasing number of reads produced per sequencing run, and the difficulty in concatenating only short sequences together. However there have been other fundamental changes to sequencing, such as ever increasing read lengths and the move towards mostly paired end sequencing, which most aligners now have to deal with (Smith and Bailey, 2000; Fonseca et al., 2012).

1.11 Previous work done on B17 and Z310 strains

Several isolates of *T. brucei rhodesiense* have been described that have distinct and reproducible phenotypes when used to infect mice. Two of these, which form the basis for this body of work, are strains from the zymodeme groups B17 and Z310. These names are derived from the regions where these zymodeme groups were first described, Busoga and Zambesi. Their phenotypes have been previously described in 2000 and are interesting in part because both are Ugandan strains which show striking phenotypic differences despite belonging to the same subspecies (Smith and Bailey, 2000). Patients infected with a B17 strain generally presented with particularly acute forms of *T. rhodesiense* sleeping sickness, a severe early stage infection and chancre (Smith and Bailey, 2000). In contrast to this, Z310 isolates caused a particularly chronic form of *T. rhodesiense* sleeping sickness and patients where often unaware they were infected (Beament, 2002; Smith and Bailey, 2000). Immunological studies on the disease manifestation in mice infected with B17 and Z310 strains have also been discussed previously (Goodhead et al., 2013; Beament, 2002). One B17 and one Z310 strain were shotgun sequenced, and subsequent analysis indicated introgression between subspecies (Jamonneau et al., 2012; Goodhead et al., 2013). The shotgun sequence data for these strains has been used for comparison within this work.

1.12 The primary aim of this thesis

The primary aim of this project was to utilize recent advances in different OMIC approaches to phenotype multiple *T.b. rhodesiense* strains, in order to correlate genetic differences with differences in clinical manifestation. This was done in an attempt to

determine the genetic factor(s) controlling virulence (as defined in this case by the clinical presentation of symptoms), to understand the difference in virulence between two Ugandan *T.b. rhodesiense* strains, belonging to two zymodeme groups, Z310 and B17. This was done through the generation of metabolomic, transcriptomic and genomic data and bioinformatic analysis.

In order to achieve the aim of the project, this work is divided into three main parts, a brief discussion of each is provided below.

1.12.1 Generating enrichment sequencing data and benchmarking against previous shotgun sequenced data

This work is discussed in Chapter two and involves both the refinement of the method for direct sequencing from blood and benchmarking this data against previous whole genome sequence data. This discusses the use of Whatman FTA™ cards as a sample collection method for clinical samples, and the preparation of these samples for downstream sequencing preparation. FTA™ cards have been used previously as a method for cataloging and preserving field samples but have only been used in limited downstream applications. The analysis within this chapter focuses on the feasibility of FTA™ card derived samples for enrichment sequencing, and how the resulting sequence data compares to whole genome sequence data using a number of parameters for sequence quality. This looks at both the actual enrichment observed, the evenness of coverage, potential allelic drop-out effects, and compares two target enrichment designs to see whether design can significantly improve the sequencing outcome.

1.12.2 Using data generated by enrichment sequencing to look at inter and intra zymodeme variation

Chapter three uses multiple strains from zymodeme groups B17 and Z310 to look at both conserved variants both zymodeme groups, which may contribute to a conserved "core" *T.b. rhodesiense* genome, and at variants unique to each zymodeme group. This chapter looks at the functional implications of these variants and what potential impacts deleterious variants may have, and whether these can explain the phenotypic variation seen in these strains. Additional human clinical samples, which were also subject to sample enrichment and analyzed for the feasibility of using very low

parasitaemia infections are also discussed here. Differences in cell cycle regulation, its effect on differentiation and the potential implications on virulence are also discussed within this chapter.

1.12.3 Combining transcriptomic and metabolomic data to understand host-parasite interplay during infection

This chapter uses both transcriptomic and metabolomic data to look regulatory differences between both strains. Differentially expressed genes (DEG) within the transcriptomic data can be used to understand what pathways are being up or down-regulated. Similarly, metabolomic data is generated by measuring the abundance of the end products of metabolic processes, and so the differential regulation of certain pathways can be inferred. Metabolomic data from samples taken during an infection also contain host metabolites. This presents challenges in extracting the parasite from host metabolites, but it also allows for host immune responses to be investigated.

CHAPTER 2

Developing a method for direct sequencing of host-parasite samples from blood

2.1 Introduction

DNA sequencing technology has advanced parasitology and enabled free access to the genomes of many parasites, both multicellular and protozoan. Primarily, the sequencing of parasites has been done from cultured material, mostly due to one of the first obstacles in sequencing parasites, obtaining enough parasite DNA to construct a library. Ease of access to this amount of starting material varies wildly on the availability of samples, and the nature of the sample. In RNAseq, particularly in high parasitiaemia infections, because the level of mRNA expression is very high in the parasite, the imbalance in host to parasite transcripts is not as exaggerated. In addition, the host reads in RNAseq are often useful in determining host responses. However in DNA sequencing, the ratio between host and parasite DNA is often severely imbalanced, making DNA sequencing directly from clinical samples which contain a mixture of host and parasite DNA, more complicated.

2.1.1 Aims of the chapter

Whatman FTA™ classic cards have been used for many years in order to catalogue samples, particularly in regions where a cold chain cannot be maintained between sample collection and downstream processing. Within this chapter, I will be determining whether blood samples collected on Whatman classic cards are viable samples for preparing sequencing libraries from. This chapter will make use of available whole genome sequence (WGS) data and compare this to sequence data generated using sequence capture technology, and assess whether this technology is robust and can be used in place of whole genome sequencing for targeted analysis. This will be assessed through the representation of heterozygotes within the sequence data, evenness of coverage and the level of off target sequence generated. The effect of target design on the quality of sequence data generated from sequence capture will also be discussed. The robustness of the enrichment technology will also be assessed through the degree of variation in probe performance across designs and samples.

2.1.2 Sequencing in parasites and the current limitations

Natural *T. brucei* infections typically have much lower parasitaemias than those observed in experimental mouse infections. This is seen especially in more chronic infections such as *T.b. gambiense* infections or during the early stage of the disease (Jamonneau et al., 2012; Chappuis et al., 2005). Here, the parasite can represent less than 1% of the total DNA. This is because the parasite is outnumbered by the white blood cells (WBCs), which contain host DNA, and this effect is exacerbated as the number of WBCs increases in response to infection; and is further compounded by the comparative genome sizes of the host and parasite (Waterston et al., 2002; Lander et al., 2001; Berriman, 2005). In the mammalian host, the genome can be over 1000 times larger that of the parasite (DePristo et al., 2011; Waterston et al., 2002; Lander et al., 2001). DNA sequencing directly from a host parasite mixed sample in *T. brucei* would require very deep sequencing in order to generate enough useable sequence data, and the majority (>99%) of the sequence data would be host. To look at either copy variants, or single nucleotide polymorphisms (SNPs), and to use these in comparative studies, would require a high level of coverage, and would be prohibitively expensive (Daily et al., 2005; DePristo et al., 2011).

Although the extent to which culturing alters the parasite's biology is debated, studies in *P. falciparum* have shown there is a distinct difference in the transcriptomes of parasites *in vivo* and *vitro* (Koumandou et al., 2008; Daily et al., 2005). Daily et al discovered that in *in vivo* samples, there was overexpression of entire gene families, which was not observed in *in vitro* samples. These gene families encoded surface antigens, which as in *T. brucei*, are essential for the propagation and maintenance of an infection. These genes also represent good potential vaccine targets and give insight into mechanisms such as host evasion. Subsequent studies by Koumandou et al in 2008 in *Trypanosoma*, again noticed these differences between *in vitro* and *in vivo*, however these were noted broadly across multiple pathways with no obvious functional group impacted (Koumandou et al., 2008). However the fitness/ virulence of the strain may be impacted, and this has been seen in parasites where they are no longer capable of completing the entire life cycle because they have adapted to culture, for example the *T. brucei* strain Lister 427, which is incapable of differentiating into fly transmissible stages (Koumandou et al., 2008; Cross and Manning, 1973; Peacock et al., 2008). Some strains are also less likely to propagate in culture, limiting sequencing to only strains

more adapted to culture, making the data less representative of the population of strains.

It is important to note that the analysis done by Koumandou and colleagues used large quantities of RNA, with 1 x 10^9 cells used, quantities that are often unobtainable using samples from experimentally infected animals or natural infections (Koumandou et al., 2008). Studies, particularly those including cattle infections, often use large volumes of blood, methods circumventing the use of such large quantities of material are largely undeveloped. More research has been dedicated to the development of micro-volume scale methods, particularly in sequencing where several low input library preparation techniques have been developed. These include the Illumina nextera kit, which allows libraries to be prepared from 1ng of starting DNA and the NEBnext ultra kit, which requires only 5ng (Caruccio, 2011). However in this instance alternative methods are required due to the DNA imbalance as mentioned above.

These limitations are not restricted to *Trypanosoma*, but highlight difficulties faced across the whole spectra of parasites. However there are other considerations, which are more specific to *T. brucei*. Amongst these, the human infective *T. brucei* subspecies. *T.b. rhodesiense* and *T.b. gambiense* are a CAT3 level organisms, and unlike CAT2 organisms, this places further restrictions on culture conditions, experimental infections used to raise parasites, and the transportation of samples from infective sources, i.e. blood. However the transportation of DNA is far less difficult. Whole blood samples, either from an experimental or patient sample, can be immobilized to prevent sample degradation using Whatman FTA™ cards (Mirchamsy et al., 1968; Smith and Burgoyne, 2004; Moscoso et al., 2004; Muthukrishnan et al., 2008). Dried blood spots have been used to overcome these difficulties in a variety of parasitic, viral and bacterial samples (Morrison et al., 2007; Mirchamsy et al., 1968; Moscoso et al., 2004; Muthukrishnan et al., 2008). Preventing the degradation of samples is a key consideration for trypanosomes because the source of natural infections is limited to within sub-Saharan Africa, where there may be greater difficulty in maintaining the cold chain, or preserving the sample between collection and processing time (Lonsdale-Eccles and Grab, 1987; Morrison et al., 2007). This is especially important with sequencing because DNA and RNA need to be of a high quality to produce high quality sequence data.

Several collection methods have been used routinely to collect and purify trypanosomes for downstream processing. Previously, DEAE columns were used widely as a method for purifying the parasites, however the popularity of this method has waned in part due to the long documented potential biochemical changes this method of preparation causes (Lonsdale-Eccles and Grab, 1987). Applying samples to Whatman cards stabilizes them at an ambient temperature, and prevents degradation between collection and processing time (Makowski et al., 2003; Kline et al., 2002; Smith and Burgoyne, 2004).

2.1.3 Whatman FTA™ Cards as a method of sample collection and storage

There have been several studies using Whatman cards for the preservation of samples from a variety of different bodily fluids, and many attempts to perfect the extraction of the sample from the card. Cards of blotting paper used solely for blood samples are also referred to as Guthrie cards, and are used routinely for neonatal blood screening (Kline et al., 2002; Makowski et al., 2003; Smith and Burgoyne, 2004; GE Healthcare, 2010). There are two main categories of Whatman FTA™ cards, the classic and the elute, with the classic binding the DNA to the card, so that a punched card piece is used directly in downstream applications, or the elute card, which can be washed to elute the sample off the card (Safar et al., 2010; Kline et al., 2002; Inoue et al., 2007; Smith and Burgoyne, 2004; Kraus et al., 2011; GE Healthcare, 2010a).

Several studies have been done using Whatman FTA™ classic cards as a source of DNA, from a whole variety of biological material and organisms, from extracting viral RNA from cloacal swabs in influenza infected birds, to detecting porcine reproductive and respiratory syndrome virus from pig blood, to human DNA for GWAS studies (Fowler et al., 2012; Safar et al., 2010; Inoue et al., 2007; Kraus et al., 2011). More recently alternative downstream applications have become a previously unexplored interest, with Fowler and colleagues looking at the potential of using DNA from FTA™ classic cards to sequence the DNA captured on them (Fowler et al., 2012). This not only reflects a shift from changes in methodology, as sequencing becomes more affordable and sequence data more accessible, but also a change to experimental design, with the improvement in sequencing technology leading to more tailored sequencing options available.

In parasitology, a by product of this is the sequencing of more parasites endemic to regions such a Sub-Saharan Africa. In regions like this collection methods such as Whatman FTA™ cards are essential and so technology capable of being used in combination with FTA™ cards are becoming more popular for processing clinical isolates. As previously mentioned, FTA™ cards are useful for cataloguing *T. brucei* because transport for CAT3 organisms is highly regulated. Kraus et al. used FTA™ cards for sampling the avian influenza virus (AVI) primarily because of the difficulties of transporting a live virus, demonstrating the clear safety and transport advantages (GE Healthcare, 2010b; Kraus et al., 2011).

Classic cards are the original Whatman FTA™ cards, and these have been used to catalogue samples for many years, originally only with the intention of using the DNA bound to the card in limited capacities such as PCR. These cards render the DNA stable by lysing cells on contact, denaturing proteins and protecting the DNA from potential damaging products such as free radicals for up to 17.5 years (Picozzi et al., 2002; GE Healthcare, 2010b). They also render infectious agents applied to the cards as safe, reducing disease risk from sample transportation (GE Healthcare, 2010b; Picozzi et al., 2002). The cards also protect DNA from bacterial and fungal growth on the card, as was shown in a study comparing FTA™ classic cards to Nucleosave™ cards, in which cards were exposed to UV in order to simulate aging and DNA damage over time, with no visible change in quality following exposure (GE Healthcare, 2010b).

Due to the limited nature of the classic cards, Whatman subsequently released FTA™ elute cards which allow DNA to be easily eluted from the card and allows the main contaminant, protein, to remain bound after elution. This is achieved by using a chaotrophic salt in the cellulose matrix, which denatures the proteins and dissociates them from nucleic acids, allowing them to be eluted off the card (Sawyer and Puckridge, 1973; Damodaran and Kinsella, 1983).

Field samples have been traditionally and continue to be, collected and sampled on classic cards, resulting in a vast catalogue of samples. However downstream analyses of the samples of these cards has been predominantly restricted to PCR, with card punches used directly in the PCR reaction. Although this is fine for PCR, this limits the number of downstream applications these samples are viable for, such as sequencing. In this chapter, I aim to bring together both the classic Whatman FTA™ card and sequencing technology, to show how samples collected and catalogued in this manner

can be used for whole genome analysis. This not only makes newly collected samples usable for this analysis, but allows samples previously collected open to analysis previously restricted to just PCR.

2.1.4 Whole genome amplification

In 1992 Telenius & Zhang and colleagues published papers on the development of a technique, which non-specifically amplified DNA using a degenerate primer (Zhang et al., 1992; Telenius et al., 1992). This has since been coined whole genome amplification (WGA), and there are three main types, multiple displacement amplification (MDA), degenerate oligonucleotide PCR (DOP –PCR) and primer extension preamplification (PEP). The DOP-PCR and PEP methods have largely been replaced by MDA and are based on traditional PCR.

PEP works by using a preamplification step to attach primer-binding sites to DNA fragments which can then be used during subsequent WGA. Random primers and Taq DNA polymerase are then used at a low annealing temperature during the WGA reaction. Whereas DOP-PCR has no preamplification step and ligates adaptors to DNA fragments to create primer-binding sites. Similarly to PEP, Taq DNA polymerase is used in the WGA reaction, however unlike PEP, DOP-PCR's primers are semi degenerate oligonucleotides and the annealing temperature increases during the reaction (Lee et al., 2008; Zhang et al., 1992; Arneson et al., 2008). Both methods are limited due to their use of Taq DNA polymerase which limits the maximum size of fragments to only 3kb, and on average produces 400-500bp fragments (Dean et al., 2002; Lee et al., 2008; Arneson et al., 2008). This also introduces errors into the sequence and bias into the coverage (Dean et al., 2002). This polymerase is also sensitive to secondary structure and these structures can result in dissociation of the enzyme from the DNA template or slippage of the polymerase, excessive amplification of certain regions of the genome and allelic drop out (Lee et al., 2008; Dean et al., 2002).

However in 2002, Dean et al published MDA as a new method of WGA, and this made these two previous techniques largely redundant and has revolutionized single cell biology, particularly the development of sequencing methods for single cells. Like PEP, MDA uses random hexamers, however instead of Taq DNA polymerase it uses the Phi29 polymerase (Lee *et al*., 2008). Long fragments instead of short fragments are used as a template, and the template needs to be denatured prior to amplification. Many of the

limitations of the previous methods are solved by MDA because the Phi29 polymerase can generate fragments of upto 100kb without dissociation or sequence bias. This enzyme also has 3' -5' exonuclease proofreading activity, which reduces the error rate to less than a 1000 times lower than Taq DNA polymerase (Handyside et al., 2004; Lee et al., 2008; Spits et al., 2006). The Phi29 polymerase also resolves secondary structures without enzyme dissociation, which greatly reduces allele drop out.

Since the development of MDA as a method of WGA, it has been commercialized by several companies, producing Picoplex by rubicon genomics, Genomiphi by GE healthcare and Repli-G by Qiagen. Multiple studies have used all three of these kits, all of which suffer from the same drawbacks in occasional allelic drop out of heterozygotes, however this is mostly an issue for single cell sequencing, with these effects reduced by using greater than 5 cells (Voet et al., 2013; Handyside et al., 2004; Spits et al., 2006). However a Genomiphi and Picoplex comparative study found that single cell Picoplex generates significantly more nucleotide copy errors (Marine et al., 2011; Voet et al., 2013). Whole genome amplification has been known to show bias in bacteria because it preferentially replicates circular genomes over linear ones, however this is not an issue in *T. brucei* due to the structure of its genome (Treff et al., 2011; Marine et al., 2011; Hellani et al., 2008).

Samples that have been whole genome amplified have been used successfully for many analyses downstream of sequencing, including copy number variation analysis (CNV) and single nucleotide polymorphisms (SNPs) (Paez et al., 2004; Treff et al., 2011; Hellani et al., 2008). Several studies have shown that WGA prior to sequencing does not adversely affect the quality of the data and generates results concordant with the sequencing of unamplified samples (Paez et al., 2004).

2.1.5 Targeted sequencing

Despite the cost of sequencing declining, sequencing an entire genome, or sequencing lots of strains in order to do population studies, is still in some cases prohibitively expensive. New sequencing technology also means that more data is being generated, and this can make analysis both computationally demanding and harder to analyse (Chilamakuri et al., 2014; Sboner et al., 2011). Often sequencing is not hypothesis free, and only certain regions are of interest. In these cases sequencing can be targeted to just certain regions of the genome, significantly reducing the cost and data to analyse.

Targeted sequencing is also referred to as target enrichment as the process of sequencing does not remove contaminating DNA, it just selectively enriches for the target. In solution based systems the process is also known as sequence capture, because the target sequence is captured by effectively being pulled out of the mixture of non target and target DNA in the sample by becoming physically bound to streptavidin beads (Chilamakuri et al., 2014).

One way of doing targeted sequencing is to just focus on the exome of an organism. Multiple studies have been done on library preparation comparisons, with Chilamakuri et al. (2014) focusing on the comparison between four commercially available solution based exome capture systems (Chilamakuri et al., 2014). These are the Agilent's sureselect human all exon, Nimblegen's seqcap ez exome library, Illumina's truseq exome enrichment and Illumina's nextera exome enrichment. All of these exome capture systems are available for custom design and species except for the Illumina truseq exome libraries (Bodi et al., 2013; Chilamakuri et al., 2014). In solution capture has the advantage of being more amenable to upscaling and does not require the same specialist equipment compared to arrays.

The two leading technologies for custom design solution based capture are Nimblegen and Agilent. In 2013, Bodi et al., did a comparison between these technologies and their ability to capture SNPs (Bodi et al., 2013). A key difference in the two systems is the Agilent uses RNA probes to complement the target region, whereas Nimblegen uses DNA probes (Chilamakuri et al., 2014; Margulies et al., 2005; Bodi et al., 2013). The Illumina truseq human exome and nextera exome technologys also use baits made from DNA (Bainbridge et al., 2010; Chilamakuri et al., 2014). Nimblegen's probes are the shortest (between 60-90 nucleotides) compared to Illumina's truseq and nextera designs, which are 95bp long and Agilent's probes which are 120 nucleotides by default, or longer for instance in AT rich genomes such as *P. falciparum* (Melnikov et al., 2011; Bainbridge et al., 2010; Chilamakuri et al., 2014). Melnikov et al., 2011 used Agilent's sequence capture but used 140 nucleotide baits to compensate for this (Melnikov et al., 2011). Similarly Gnirke et al demonstrated in 2009 the reproducibility and robustness of Agilent's custom sequence capture, this time with 170 nucleotide baits, which showed a high percentage of on target DNA and even coverage (Gnirke et al., 2009).

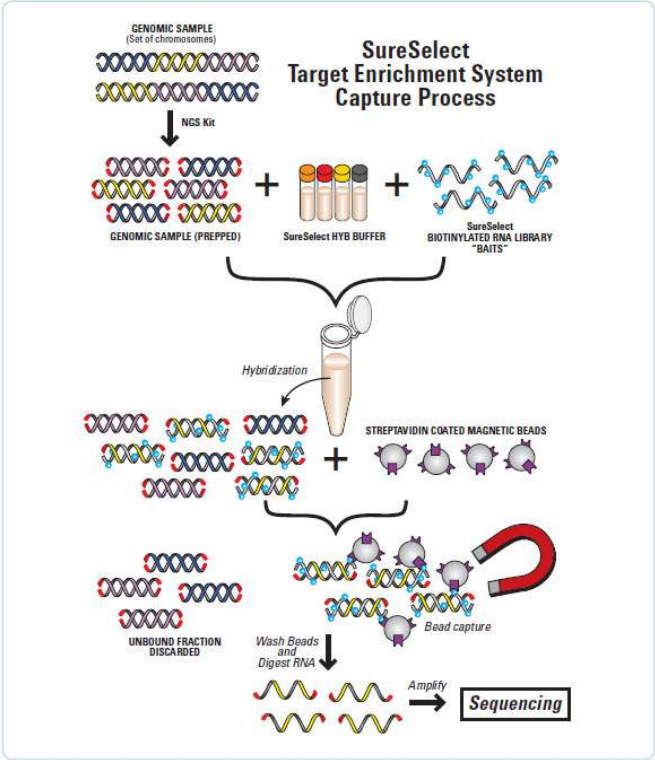

Figure 2.1: Diagram illustrating the sureselect enrichment process, taken from www.genomics.Agilent.com. The process of library preparation prior to the hybridisation of the biotinylated baits to the target DNA is similar to that of a Truseq library and was followed as per manufacturer's instructions. As shown, genomic DNA (gDNA) is sheared to small fragments, which are prepared using a variety of steps including end repair and addition of adaptors, like a typical library. The prepared DNA is then hybridised to a library of designed biotinylated sequences which bind to the target DNA in the sample, and can subsequently be removed using streptavidin beads. The bound DNA then amplified and prepard for sequencing. The biotinylated library baits used in this chapter were designed using the Tbrucei927 reference genome, which can be accessed at www.tritrypdb.org.

Agilent's probes are also uniform in length whereas Nimblegen's probes vary between 60-90bp, to help enrichment in more difficult to sequence regions of the genome (Bainbridge et al., 2010). Agilent boosts performance in these regions by increasing the number of copies of a probe instead. Both also offer tiling options so that the probes overlap in the design (Morgulis et al., 2006; Bainbridge et al., 2010). Agilent's sureselect system was used to generate the data within this chapter and the library preparation process is outlined in Figure 2.1.

2.1.6 Windowmasker can be used to mask repetitive regions and improve sequence capture design

In sequencing, the reads generated can often be biased in low complexity/ repetitive regions and/ or regions that have either very low or high GC content. In order to deal with this, because sequence capture only targets a proportion of the genome, the region can be designed to avoid highly repetitive regions. In sequence capture, the bias caused by repetitive sequence is often even more pronounced and so software such as Windowmasker can be used to find these regions in the target, and mask them so that they aren't included in the final design (Tarailo-Graovac and Chen, 2009; Morgulis et al., 2006). Unlike its predecessor, Repeatmasker, Windowmasker requires only the genome sequence you are using to find low complexity and repetitive regions (Tarailo-Graovac and Chen, 2009). In comparison, Repeatmasker required the alignment of subsections of the genome of interest to a curated repeat library. This was not only computationally intensive, but meant that only select genomes could be used in conjunction with the program (Morgulis et al., 2006; Tarailo-Graovac and Chen, 2009). Windowmasker uses an algorithm that works in two parts, first by calculating the number and size of Nmers in the genome, then iterating through the genome and masking regions that it determines to be repeats or of low complexity (Morgulis et al., 2006).

2.1.7 Bioinformatic analysis

2.1.7.1 Short read sequence alignment tools

The most commonly used alignment tools for DNA sequencing are short read sequence aligners, the two preferred and most commonly used ones being BWA (the burrows wheeler aligner) and Bowtie. These tools are primarily used because they handle data in a very memory efficient manner, which significantly decreases the time taken to complete an alignment. They do this by using programming functions, known as hashes. A hash is a memory efficient method of storing data, with every data value associated with a key. These keys can then be used to search through the data more efficiently instead of looking through every individual data value. Unlike some aligners, such as their predecessor MAQ, BWA and bowtie do not hash the read sequences and

then scan through the reference sequence (Li and Durbin, 2009). They also do not hash the genome, a method used by many aligners including SOAPv1, NovoAlign and BFAST, which can be very memory intensive depending on the size of the genome (Fonseca et al., 2012; Li and Durbin, 2009; Langmead et al., 2009) (http://www.novocraft.com; http://genome.ucla.edu/bfast). Instead they search backwards, aligning for the 3' end of the read using the Burrows-Wheeler Transform (BWT) algorithm, which makes the algorithm much more efficient (Fonseca et al., 2012; Li and Durbin, 2009; Langmead et al., 2009). BWT has been used in the development of other aligners such as SOAPV2 (http://soap.genomics.org.cn/).

In BWA, partial alignments/sequences with imperfect matches are scored. This score reflects any penalties given for mismatches or gaps and is used to prioritise alignments in order to find the most confident intervals (Smith and Bailey, 2000; Fonseca et al., 2012; Li and Durbin, 2009). Corresponding reverse reads are also processed in the same way, and the quality scores associated with both forward and reverse reads are used to find the optimal alignments.

Bowtie's algorithm is faster than BWA, however it sacrifices the quality scoring and the confidence of the alignments because it makes fewer confident matches compared to BWA and MAQ and it also misses the best imperfect matches (Li and Durbin, 2009; Smith and Bailey, 2000; Langmead et al., 2009). Both Bowtie and BWA use the FM index, which is a deviation of the burrows-wheeler transformation (BWT) (Ferragina and Manzini, 2001; Li and Durbin, 2009; Langmead et al., 2009). This allows the algorithm to be accurate and memory efficient and was first described in 2001 by Ferragina and Manzini (Ferragina and Manzini, 2001). The BWT transforms the DNA sequence into a sorted matrix, which can then be stored more compactly, and be reversed to provide the full DNA sequence again (Li and Durbin, 2009; Ferragina and Manzini, 2001; Langmead et al., 2009). Unlike blast, these two programs look at windows along the sequence and align these rather than looking at the sequence as a whole, and this allows it to accept mismatches and score the alignments based on these, but also reduces the memory required, because only part of the sequence is stored at any one time (Li and Durbin, 2009; Langmead et al., 2009).

BWA supports gapped alignments, whereas the original version of Bowtie didn't and MAQ is incapable of doing so for single end sequencing reads (Li and Durbin, 2009). BWA also gives mapping quality, as does MAQ, whereas Bowtie doesn't, which also

accounts for the additional time it takes to map in BWA (McKernan et al., 2009; Li and Durbin, 2009). However the main advantage to both Bowtie and BWA is that their standard output is a SAM file, rather than the MAP file created by MAQ.

SAM files are now used universally, and are amenable to conversion into multiple other file formats, and use by a high number of bioinformatic tools. In contrast, MAQ was initially used as a pipeline for quality control of reads, alignment and variant calling, which meant that it was used for start to end analysis and its output was not made for compatibility with other software. Since MAQ, many new bioinformatics tools have become available for each stage of the analysis, such as BWA for alignment, and GATK for variant calling, which have rendered MAQs map files obsolete, and are much better at assessing qualities such as basecall qualities. Improvements in sequencing chemistry also mean that new bioinformatic software is essential for handling changes to sequence data, such as the greater volume of reads generated, and increased reads lengths, which MAQ is less capable of analyzing because it is not still being actively developed. Both BWA and Bowtie also support both base space and colour space reads. Illumina reads are all in base space, however SOLiD data is in colour space and is not universally accepted by aligners (Li and Durbin, 2009; McKernan et al., 2009; Li and Durbin, 2010).

Within this chapter, the analysis has been done in BWA, and both SOLiD and Illumina data has been aligned using the BWA aln algorithm. This is a short read aligner algorithm, however BWA-MEM and BWA-SW are also available, and are preferred particularly for longer reads (Rimmer et al., 2014; Li and Durbin, 2009; 2010).

2.1.7.2 SamTools

Samtools is a collection of tools, which utilize files in SAM, BAM and CRAM formats (Li and Durbin, 2009; Rimmer et al., 2014). SAM stands for the Standard Alignment/Map format, has become the standard file format for next generation sequence data and is generated by many short read aligners including BWA (Li and Durbin, 2009; McKenna et al., 2010). One of its tools facilitates the conversion from SAM to BAM format, which is a compressed version of the SAM file. Both formats will be used in the analysis in this chapter. These tools are used to allow for the manipulation of the compressed BAM file in order to sort, index and prepare the data for further downstream processing such as SNP calling (McKenna et al., 2010; Li and Durbin, 2009; DePristo et al., 2011).

2.1.7.3 Genome Analysis Toolkit (GATK)

The Genome Analysis Toolkit is, similarly to Samtools, a collection of tools to help analyse NGS data (McKenna et al., 2010; DePristo et al., 2011). However the primary focus of the tools in GATK are for variant detection, validation and genotyping. GATK's main function is to call single nucleotide polymorphisms from the BAM file produced in Samtools (Liu et al., 2013; McKenna et al., 2010). Samtools also has a variant caller, samtools mpileup, however GATK is more stringent when calling SNPs and so because the primary focus behind this chapter is not variant discovery, but the validity of SNPs found in more than one type of sequencing technology, GATK was preferred over samtools for calling SNPs in this instance (Melnikov et al., 2011; Liu et al., 2013).

2.2 Methods

2.2.1 Designing a target region for sequence capture

Agilent's sequence capture technology was used to sequence DNA from experimental mouse infections, which contained a mixture of both host and parasite DNA. In order to test the effectiveness of this technology a target region was designed containing 985 target genes. These target genes were selected so that all of the 11 megachromosomes were equally represented, and genes were selected from regular intervals along each. The 11 megachromosomes are approximately 26Mb in combined length, and initially 1000 genes were identified and selected from the whole genome to be part of the first design. This was done on the basis that one gene would be taken from at least every 20,000bp interval along each chromosome, using a sliding window approach.

From these 1000, 985 were used in the first design. Genes were selected on the basis that they were from non-repetitive regions, and so the telomeric regions are underrepresented due to their highly repetitive nature. Annotations from the Tb927 v4.1 version of the genome were used in this selection, and to reduce the introduction of repetitive regions into the design, genes that had been identified as leucine rich repeats (LRRPs), VSGs or ESAGs were removed. However, as you can see from subsequent analysis, several of the genes in the first design were enriched successfully,

and so were used in the subsequent redesign, which were later identified as LRRPs/VSGS or ESAGs when the Tb927 genome underwent reannotation.

The entire sequences of these genes were used, and gene co-ordinates and sequences were obtained from the Tritrypdb website (www.tritrypdb.org) from the reference strain Tb927 version 4.2. An overview of the process undergone to design the target region is shown in Figure 2.2. The gene IDs of the genes used in this design, and in the second design mentioned beneath, are given in the appendices.

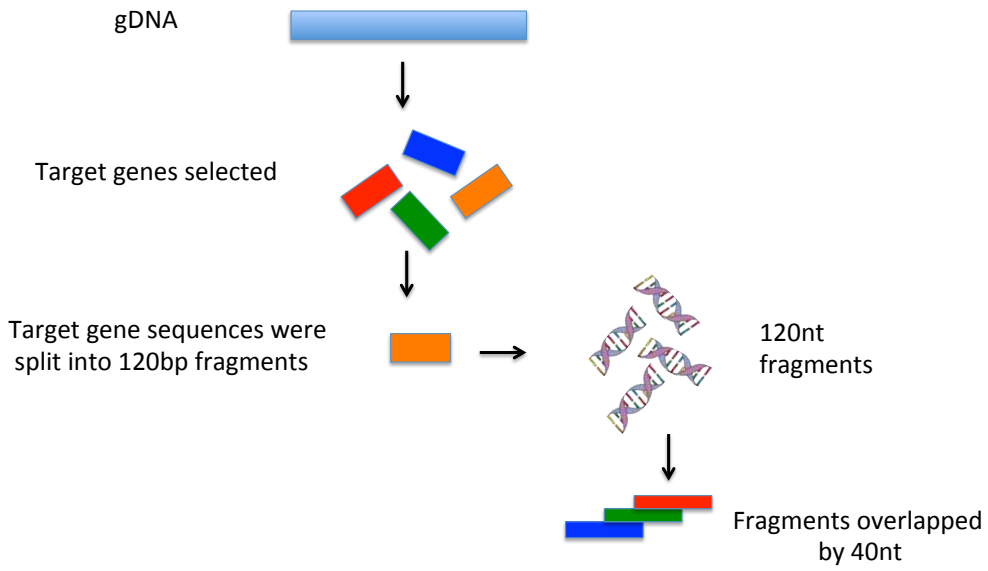

Figure 2.2: The 120nucleotide biotinylated baits were designed by selecting genes at regular intervals from the Tb927 v4.2 genome (accessed from www.Tritrypdb.org). For each of these selected genes, the genomic sequence was divided into 120nt fragments, and these fragments overlapped by 40nt.

Whole gene sequences from the selected genes were divided into 120 nucleotide long pieces in order to generate the bait sequences as shown in Figure 2.2. Each of these 120 nucleotide long baits were designed so that they overlapped each over by 40 nucleotides each, in order to reduce the chances of regions of low coverage/ allelic drop out. The baits were all blasted against themselves and any identical were removed. In total there were 49,161 120nucleotide long baits, over a target region of ~2.9Mb, which is approximately 12% of the genome (Melnikov et al., 2011). These were custom-made biotinylated baits, which were synthesized by Agilent technologies (www.genomics.agilent.com). More genes are targeted on the longer chromosomes; which reflect that the genes were selected at regular intervals along each chromosome. In this design only one copy of each oligo was used.

Table 2.1 shows the breakdown of the number of baits per chromosome and the number of genes targeted along each chromosome. The last column shows the percentage of the each chromosome covered, which is relatively uniform across the chromosomes (between 8-15%). The percentage covered per chromosome is derived by the percentage the length the baits cover of the whole chromosome length. The chromosome and gene lengths were generated from Tritrypdb annotation data from the Tb927 reference version 4.1. The newest version of this genome v8.1 was not used to generate gene lengths because the gene lengths have subsequently been reannotated, however at the time of design, the baits were only designed to cover the gene intervals within Tb927 reference version 4.1. The actual percentage covered was calculated using the gene lengths of each target gene per chromosome. This design also included the HpbHpr Haptoglobin haemoglobin receptor gene.

Table 2.1: This table shows the number of genes in the first design per chromosome. Chromosome and gene length information derived from tritrypdb from www.tritrypdb.org. The percentage of the chromosome targeted is based on the 120 nucleotide length of each bait, the number of baits per chromosome, and the proportion of the chromosome based on the length of each chromosome. These baits were designed to overlap by 40 nucleotides and so are, tiled three-fold. Coverage here is defined as the breadth of coverage across the target, which is the percentage of nucleotides in the chromosome covered by reads.

Chromosome number	Number of genes targeted (total number of genes on chromosome)	Number of baits per chromosome	Actual percentage of chromosome covered (%)
1	50 (532)	2896	12
2	56 (352)	3167	15
3	79 (623)	4074	12
4	54 (587)	2121	8
5	60 (579)	2719	10
6	52 (622)	2662	9
7	70 (774)	3835	9
8	92 (906)	4919	10
9	118 (1602)	6986	9
10	170 (1762)	9797	12
11	184 (1878)	5969	10

2.2.2 Redesign of target region

The target region was subsequently redesigned and nested within the original target. This second design used genes that were enriched well within the first design. This second design also included oligopeptidase B, which is a known virulence factor in *T. brucei*. Although this target was nested within the original target region, and so non-repetitive sequence should have been removed, Windowmasker was used as an additional check to remove any repetitive sequence from the final design. This masks low complexity and highly repetitive regions. As before, 120 nucleotide length baits were used and tiled 3 times to provide a 40 nucleotide overlap. Unlike the original design, these baits were designed against the Tb927 genome version 8.1.

This design was considerably smaller than the previous design, with 731 genes included and a total target covering 2Mb and a total of 43,298 probes, with a 40 nucleotide overlap as before. These targets were chosen from the original design based on their level of coverage within the sequence data. The level of coverage was based on an average across the whole gene rather than by each individual 120 nucleotide sequence. Genes that had an average depth of coverage greater than five and less than 500 across 6 strains (the other samples discussed in Chapter3), were chosen. These cutoffs were chosen because poorly performing baits either unsuccessfully enriched and had a coverage depth of 1, or excessively overenriched, and had a very high coverage of 1000 or more. Suredesign XT™ software was used to generate the baits in this design and regions of sequence with a GC content higher than 65% or at the start/end of a region were considered more difficult to capture. These regions were "boosted" and additional copies of these oligos were used.

Similarly to Table 2.1, Table 2.2 shows the number of genes targeted per chromosome and the percentage of the chromosome covered in design two. As before, the actual percentage covered is roughly uniform, but lower in chromosome 11 because fewer of the genes targeted in the initial design on chromosome 11 performed well compared to the other chromosomes and these were excluded from the second design. As was seen with the first design, chromosome 2 had a slightly higher percentage of genes included within the design (15% when the average was approximately 10% in design one, and 14% when the average was approximately 8% in the second design). Table 2.2 also shows the change in the percentage of the chromosome covered by the target region. The second design is smaller, and so a decrease in the percentage covered per

chromosome is expected, and this is between 1-2% per chromosome except for chromosome 11. As explained above, this is due to more genes from chromosome 11 being removed from the target region because they performed poorly.

Figure 2.3A shows the chromosomal position of the genes included in the first design, and shows that the genes targeted are at regular intervals along each chromosome, with no bias towards particular regions of the chromosome, or a particular chromosome. Figure 2.3B shows the positions of the genes that were included within the first design but excluded from the second design due to poor performance. This figure shows that those genes that performed poorly were not solely from one strand, or to one specific region, with those underperforming spread relatively evenly throughout the mega chromosomes except for as previously mentioned chromosome 11, where genes predominantly at the 3' end of the chromosome consistently performed poorly. Figure 2.3C shows the chromosomal position of the genes included in design two. Despite the genes shown in Figure 2.3B being excluded, there is still relatively uniform gene coverage per chromosome, except for chromosome 11, which is underrepresented.

Table 3.2: This table shows the genes included in the second design, per chromosome. As in table 2.1, chromosome and gene length information derived from www,tritrypdb.org. The last column shows the percentage change in coverage per chromosome between the two designs. This design is smaller, so we would expect a decrease in the coverage. This decrease is relatively uniform except for chromosome 11, in which a greater percentage of genes performed poorly and were excluded from the second design.

Chromosome number	Number of genes targeted (total number of genes on chromosome	Number of baits per chromosome	Base pairs covered in region (bp)	Percentage of chromosome covered (%)	Percentage decrease in coverage in the second design (%)
1	43 (532)	2232	107425	10	2
2	48 (352)	3475	165278	14	2
3	65 (623)	3506	166688	10	2
4	39 (587)	1817	95765	6	2
5	40 (579)	2692	132017	7	2
6	45 (622)	2426	120614	7	1
7	56 (774)	3100	156000	7	2
8	73 (906)	4035	209597	8	2
9	102 (1602)	5872	264796	7	1
10	138 (1762)	9154	428910	10	2
11	82 (1878)	4989	241530	5	6

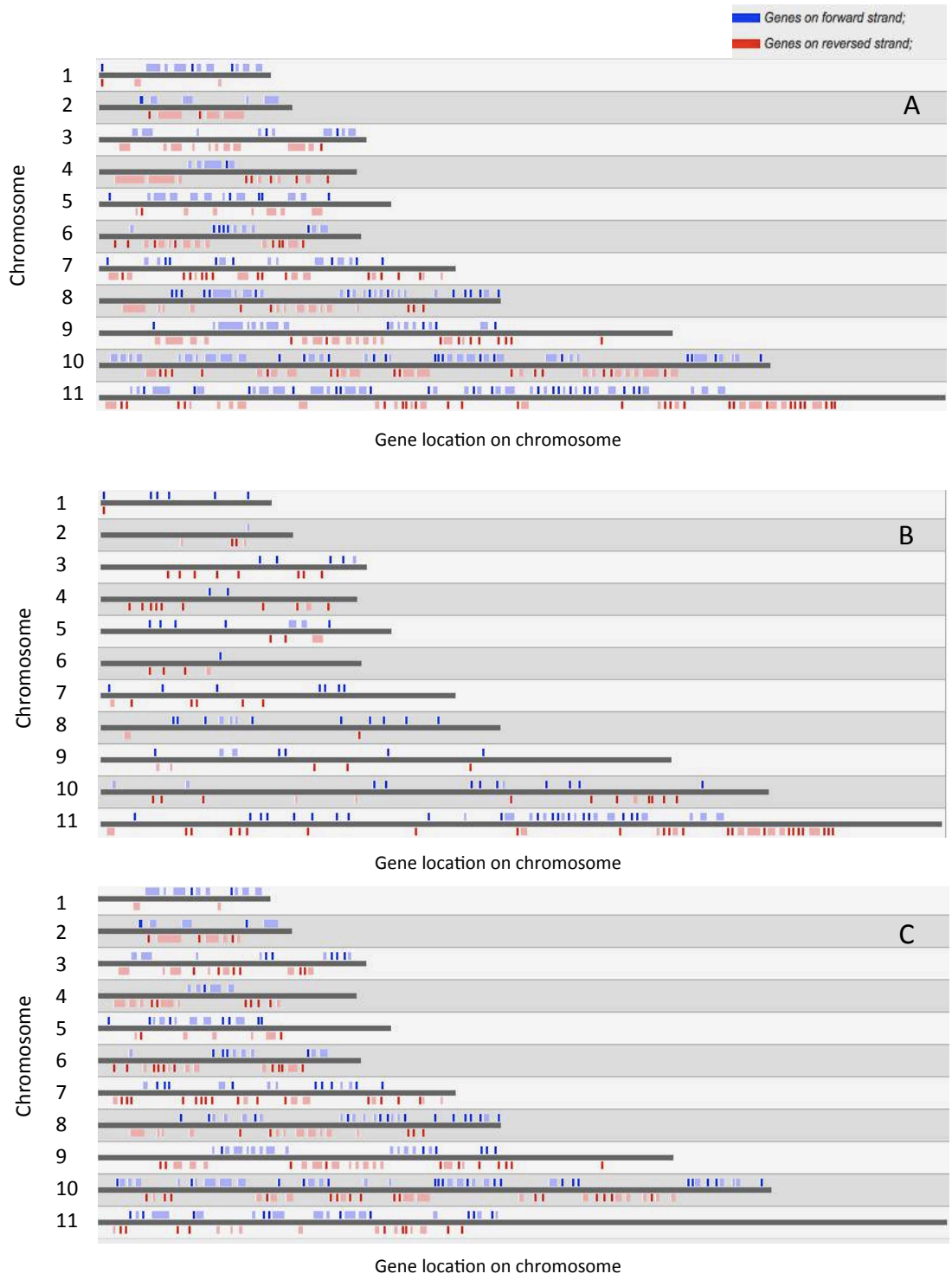

Figure 2.3: Shown are the positions of genes in both target regions. Genes positioned on the forward strand are coloured blue, red on the reverse. Dark individual blue or red bars represent one gene on either the forward or reverse strand, respectively. Genes that are adjacent to each other are shown collapsed into a lighter bar. The position is shown starting from the first position of each chromosome. A shows the position of all the genes in design one, with relatively uniform intervals covered across each chromosome. B shows only the locations of the genes, which were included in the first design, but excluded from the second due to poor performance. Those performing poorly were found on both strands, and at regular intervals, suggesting equal enrichment across chromosomes, except for chromosome 11, which was poorly enriched on one half of the chromosome. C shows the genes included in design two, with regular intervals between target genes except in chromosome 11. Diagrams generated using Tritrypdb resources available at www.tritrpdb.org

2.2.3 Strains selected

The two strains used in the initial design were blood samples taken from experimentally infected mice, and had previously been whole genome shotgun sequencing using SOLiD sequencing. These strains are referred to as B and E here and belong to zymodeme groups Z310 and B17 respectively. Additional metadata is included in Table 2.3 below. These parasites are *T.b. rhodesiense* parasites isolated originally in 1997 and the phenotypes have been previously described (Smith and Bailey, 2000). These strains originate from Uganda, one of the few locations where *T.b. rhodesiense* and *T.b. gambiense* infections co-exist (Picozzi et al., 2005). This presents an unusual circumstance in that these usually geographically isolated have the potential to recombine. These strains were selected due to their difference in phenotype despite them both being Ugandan *T.b. rhodesiense* strains. There are additional strains from the same zymodeme group, and these are discussed in chapter 3.

Table 2.3: The two strains listed below were used in target enrichment sequencing and are discussed in this Chapter. Additional strains were used in the targeted sequencing however these are mentioned in Chapter 3. These two strains were used to benchmark enrichment sequencing in this chapter because there is whole genome sequence data available for these. The strains were assigned to a zymodeme group based on the iso-enzymes they have and determined via multi locus electrophoresis (MLEE). These strains were of initial interest due to their difference in clinical manifestation despite both being Ugandan *T.b. rhodesiense* strains.

Zymodeme group	Strain	Phenotype in human infections	Additional metadata	Sample preparation method
Z310	B	Chronic	Patients had either asymptomatic first stage disease, or presented with symptoms at late stage. Lack of chancre.	Lysed blood applied to Whatman FTA ™ classic card
B17	E	Acute	Patients presented with severe early stage. Chancres present.	

2.2.4 Experimental infections

Per parasite strain, 2 female 8-10 week old A/J mice were inoculated intraperitoneally (IP) with $10^4 ml^{-1}$ parasites in 0.2ml of infected murine blood from frozen stabilates. This mouse strain is more susceptible to infection and was used to raise parasites before passaging into a more resistant mouse strain, C57BL/6. The parasitaemia in the mice was monitored by tail snip and thin film daily. The mice were humanely culled by CO_2 once the first peak of parasitaemia was reached; this was determined by a parasite density of between 5-15 parasites per field over an average of 5 fields of view at a 40x magnification on a light microscope, and using a haemocytometer to calculate parasitaemia. Terminal bleeds were taken by cardiac puncture using 2mM EDTA as an anticoagulant. 1 $x10^4$ parasites were then passaged into C57BL/6 female 8-10 week old mice by IP injection and monitored by tail snip. These were also humanely culled by CO_2 at the first peak of parasitaemia. Terminal bleeds were also taken from these mice using cardiac puncture and 2mM of EDTA.

2.2.5 Processing of samples prior to application on card

Following collection by cardiac puncture, the whole blood was spun at 6,000g for 2 minutes, the plasma removed, and the resulting pellet resuspended in an equal volume of PBS. To this 1ml of Ammonium-Chloride-Potassium (ACK) lysis buffer (Life technologies, cat: A10492-01) was added to lyse the red blood cells (RBC). This was then centrifuged at 6000g for 1 minute; the supernatant removed, and resuspended in 100µl of PBS and flash frozen.

2.2.6 Storage of samples

For the initial two strains in the pilot study, blood prepared as described above from the C57BL/6 terminal bleeds was diluted to 10^4 with phosphate buffered saline (PBS) pH 7.0 presuming an undiluted parasitaemia of approximately 10^6 parasites ml^{-1}. This was to test the sensitivity of both the subsequent MDA and library preparation to reflect that natural infections, particularly in humans, have a much lower parasitaemia. 20µl of lysed blood was applied to the centre of a FTA™ classic card (GE life sciences,

cat no: WB120208). ACK lysis buffer lyses the red blood cells and prevents excess proteins contaminating the sample. Once applied the blood was allowed to dry at room temperature for 2 hours. Dry cards were subsequently stored at room temperature.

2.2.7 DNA extraction

There are several tried and approved methods of DNA extraction, however the majority of methods for classic cards involve the direct use of card punches (Hoare, 1972; Cox et al., 2005; Stangegaard et al., 2013). Several papers have tried to determine the best method of eluting DNA off classic cards (Makowski et al., 2003; Ahmed et al., 2011). In this Chapter, single punches were used per extraction, and punches pooled subsequent to extraction to increase the elution of DNA from the card.

2.2.8 Disc wash

Prior to extraction, any discs from the FTA™ cards must first be washed adhering to Whatman's protocol (GE Healthcare, 2010a). This removes any erythrocytes on the cards, removing many potential contaminants from the subsequent DNA extraction. Despite little evidence for cross contamination when cutting discs from different cards, due to the multiple amplification and other processes downstream, in between punches the hole punch was rinsed with 70% ethanol, dried, and punches were taken from blank pieces of FTA™ card to reduce potential carryover (GE Healthcare, 2010a). Several methods were used to extract DNA from the classic FTA™ cards, of these, the most successful, and the one used to prepare libraries for sequencing was a Whatman approved method, as described below.

2.2.9 High pH and room temperature method

DNA was eluted from FTA™ paper using a high pH and incubating at room temperature. One 2mm punched disc was used to which 35µl of 0.1M NaOH, 0.3mM EDTA, pH 13.0 solution was added. This was left to incubate for 5 minutes at room temperature, and following the addition of 65µl of 0.1M Tris-HCL, the solution was flash vortexed 5 times and left to incubate for a further 10 minutes, also at room temperature. This was subsequently flash vortexed 10 times after which, the punch was squeezed and

removed. This eluate was then used directly in subsequent PCR reactions and whole genome amplifications. The recovery of parasite DNA from the card was maximized using single discs per extraction, and subsequently pooling.

The samples used for library preparation were prepared from two 2mm diameter punches of card; the eluates of these were then pooled. Yields from DNA extractions from FTA™ cards are often low, partly due to the small volume of sample applied to them. Pooled eluates were further concentrated prior to WGA by lyophilisation at 30°C and resuspension in 20µl of sterile H_20.

These crude extractions were then quality checked using their 260/280, 230/260 ratios on the Nanodrop™, and concentration by HS DNA Qubit™ (Fisher Scientific, UK; Invitrogen, UK). This DNA was PCR'd using the following programme, initial denaturation at 98°C, and 35 cycles of 98 °C, for 30 seconds, an annealing temperature of 53 °C for 30 seconds, extension at 72 °C for 30 seconds and a final extension at 72 °C for 10 minutes. 1µl of undiluted DNA was used in a 10µl reaction mix containing 5µl of 2x Bioline's Biomix™ red and 2µl of 1.5pmol forward primer and 2µl of 1.5pmol of reverse primer (Bioline,UK). Primers 5' GATGAATCTCCCGGCAGTAA 3' and 5' CTGCCT TTGCATCACCACTA 3' were used to detect trypanosome DNA and 5' GGAACATCGACATGGGGTAA 3' and 5' GTAGCCTGTGCATCCTC 3' to detect mouse DNA. These primers were designed and used previously (unpublished work), against housekeeping genes, in *T. brucei*, this was against gene Tb927.9.1540, and in *M. musculus* gene Nrxn1. PCR products were then ran on a 1.5% agarose gel. Amplification in these samples is usually detected in the host (mouse), but not in the parasite prior to WGA.

2.2.10 Whole genome amplification (WGA)

Even in high parasitaemia infections, the proportion of the total DNA that is trypanosome DNA is so small (0.01% in a $10^6 ml^{-1}$ infection), and the concentration of the DNA extract relatively low, that the eluate from the extraction requires whole genome amplification before it can be detected. This was done using the Genomiphi whole genome amplification kit and followed according to manufacturer's instructions (GE Healthcare, 2010a; Seth-Smith et al., 2013). 2µl of DNA were used in a 40µl reaction, and these were set up in triplicate. The manufacturer recommends 10ng of

DNA is used, however extractions from the FTA™ cards were often lower than this. Due to the low concentration of the original samples, samples were incubated for 16 hours at 30 degrees because a 3 hour incubation did not provide sufficient amplification. Successful amplification was observed by directly loading and running product on a 1.5% agarose gel. This not only showed positive amplification, but tight bands and lack of smear indicated the DNA was not degraded.

Following successful amplification the DNA was diluted to 20ng/µl and 1µl of product was then used in a 10µl PCR reaction mix using the same conditions as the unamplified DNA in order to determine amplification of trypanosome DNA. Only samples that were positive for trypanosome were used to construct libraries. 20µl of DNA from each of the amplified replicates (60µl in total) were pooled together and then cleaned using an AMPure clean up as described below. HS DNA Qubit™ values were used to determine the concentration and Nanodrop™ values for the sample quality (Fisher Scientific, UK; Invitrogen, UK).

2.2.11 AMPure clean up

AMPure beads were used to remove residual enzymes from the WGA reaction and improve the quality of the sample (Agencourt, Cat no: A63880). The ratio of beads to DNA, based on volume, not concentration, determines the size fragments removed from the sample. This method can be used as an alternative to size selection through gel excision and can reduce the percentage of sample loss. The bead volume was 1.8 times the volume of the DNA, which only removes fragments of 200bp or less. To ensure optimum binding of the DNA to the beads, DNA was added to the beads instead of beads to the DNA.

60µl of the pooled WGA'd DNA was added to 108µl of ampure beads and the clean up was performed as per manufacturer's instruction. 500µl of 70% ethanol was used per wash and samples were resuspended and eluted in 20µl of sterile H_20. The concentration and quality of these samples was then determined using the HS DNA Qubit™ and the quality using the Nanodrop™ (Fisher Scientific, UK; Invitrogen, UK).

2.2.12 Library preparation

4µg of total DNA was used for Sureselect enrichment, as determined by Qubit™ values (Invitrogen, UK). This exceeds the recommended 3µg input, however both 4µg and 3µg inputs were tested and 4µg did not overload the end repair reaction, and allowed the samples to undergo less cycles of PCR, reducing the likelihood of PCR duplicates. This was validated by shearing 3µg and 4µg aliquots of template, and observing whether the yield of AMPure purified end repaired DNA was decreased. Increasing the starting template concentration did not reduce the yield of successfully end repaired DNA. In order to generate enough total DNA for this input, these samples were WGA'd from 10ng aliquots of the cleaned WGA'd DNA and pooled and cleaned as previously mentioned. This did not adversely effect the quality of the DNA, which was determined by 1.5% agarose gel, PCR, 260/280 and 230/260 values. The libraries were prepared according to the manufacturer's instructions using the Sureselect for Illumina paired end libraries protocol (Illumina, inc). The chosen PCR cycling and hybridization conditions used are given beneath.

Pre capture PCR was done using 6 cycles in the following programme, 98°C for 2 minutes for initial denaturation and then 6 cycles of 98 °C for 30 seconds, 65°C for 30 seconds then 72°C for 1 minute then a final extension of 72°C for 10 minutes. For the first design, , 500ng of strain B and E were hybridized for 24 hours and post capture PCR and addition of adaptors was done after the DNA was eluted from the streptavidin beads. In the second design, 750ng of DNA were hybridized for 24 hours and the post capture PCR and addition of adaptors was done whilst the DNA was still attached to the beads. The libraries were single indexed. The sequencing for the first design was performed on one Miseq run, and sequencing for the second design was performed on a rapid run of the Hiseq, the prepared libraries were given to the Centre of Genomic Research (CGR) for sequencing. All data was paired end and for miseq runs was 2 x 150bp and 2 x 100bp reads for hiseq data (Illumina, inc).

2.2.13 Bioinformatic analysis

Data previously generated using SOLiD sequencing was available for both strain B and E, however the enrichment data was generated using Illumina technology. Due to this, there are slight differences in the treatment of the data, and these are outlined below and shown in Figure 2.4. Figure 2.4 gives an overview of the pipeline used to generate the BWA mapping data.

2.2.13.1 Mapping of the Illumina reads

BWA was used to align the enrichment sequence data to the *T.brucei brucei* reference Tb927 version 8.1, accessible from (www.tritrypdb.org). Prior to mapping, the short adaptor sequences used in the library preparation stages to allow samples to be identified post sequencing, were trimmed using Cutadapt version 1.2.1 using option –O 3, which trims the 3' ends which match the index sequence (Martin, 2011). Sickle was subsequently used to trim reads with a quality score of less than 20.

Due to the small size of *T.brucei's* genome, the reads were first indexed using BWA's algorithm, is. BWA's aln algorithm was then used to align the Illumina reads to the reference using default settings and generate sequence alignment (SA) co-ordinates. The default parameters were optimal for this dataset analysis. The effect of decreasing the number of mismatches allowed per read (decreasing from the default of 4% of the length of the read, which in the 150bp reads is 6 mismatches by default), further trimming the reads, and mapping the paired end data as fragment rather than paired end, were all investigated, and did not significantly improve the quality or percentage of mapping.

The reads were paired end which allows BWA to use both read files to pair reads and detect structural variations. BWA sampe was used to generate a SAM file from the SA co-ordinates produced by BWA aln. All mappings also added read group information at the BWA sampe stage using the parameter –r, this is needed for downstream analysis by GATK.

2.2.13.2 Mapping of SOLiD reads

SOLiD whole genome sequence data was already available for strains E and B. However colourspace data needs to be converted from a csfasta (colourspace fasta) and qual (quality scored) files into a fastq format, which contains the sequencing reads and quality data combined. This was done in BWA using the solid2fastq.pl script. These reads were also indexed using BWA is algorithm and aligned using BWA aln. These reads were considerably shorter than the Illumina reads (50bp) and single ended, so the BWA samse option was required subsequent to BWA aln.

2.2.14 SamTools

The SAM files generated by BWA samse/sampe were then converted into BAM files. The raw number of unmapped and mapped reads were calculated, both by extracting reads with a bitwise score of 4, which are unmapped, and using samtools flagstat function, which gives additional mapping information.

For single ended mappings, the BAM file was first extracted and then sorted. Unique reads were then extracted from this BAM file using reads with a X:T:A:U tag. The number of uniquely mapping reads was also counted. This removes reads that map to more than one place on the reference, however this does not remove PCR duplicates. PCR duplicates are often caused by the additional PCR cycles needed to produce a library. PCR can often bias a library by causing excessive amplification in certain regions of the genome. This is often exacerbated in low input samples where a greater number of cycles is required in order to make a viable library. Samtools rmdup was used to remove these PCR duplicates using the –s parameter, which is for single ended reads.

For paired end mappings, the BAM file was extracted and sorted, samtools flagstat was then used to generate information on the number of paired and unpaired reads, and singlet information. Another BAM file was created from a filtered version of the SAM file. In this file only reads with a SAM bitwise flag of either 99, 147, 83 or 163 were extracted, these meant that the reads were paired, mapped and correctly orientated. This filtered BAM file then had the unique reads extracted using the X:T:A:U flag, and PCR duplicates were subsequently removed using samtools rmdup using the –S parameter, which is for paired end reads.

The alignment files were then filtered for only positions within the target region to ascertain the off target effect using Samtools view –L and providing an interval list. For design one, these intervals were determined by using blastn with the Tb927 version 4.1 genome gene sequences, against the Tb927 version 8.1 reference and obtaining the new co-ordinates, because version 4.1 was used to design the probes in the first design. The second design used a newer reference, and these genes had newer annotations, so gene co-ordinates from the Tb927v8.1 reference could be used directly for this.

2.2.15 Sequence coverage/ Depth

The sequence coverage was determined using the script coverageStatsSplitByChr_v2.pl which was provided by Kevin Ashelford. This gave the coverage depth across each of the genes in the custom reference and was used in the second array design to determine which genes were enriched the most.

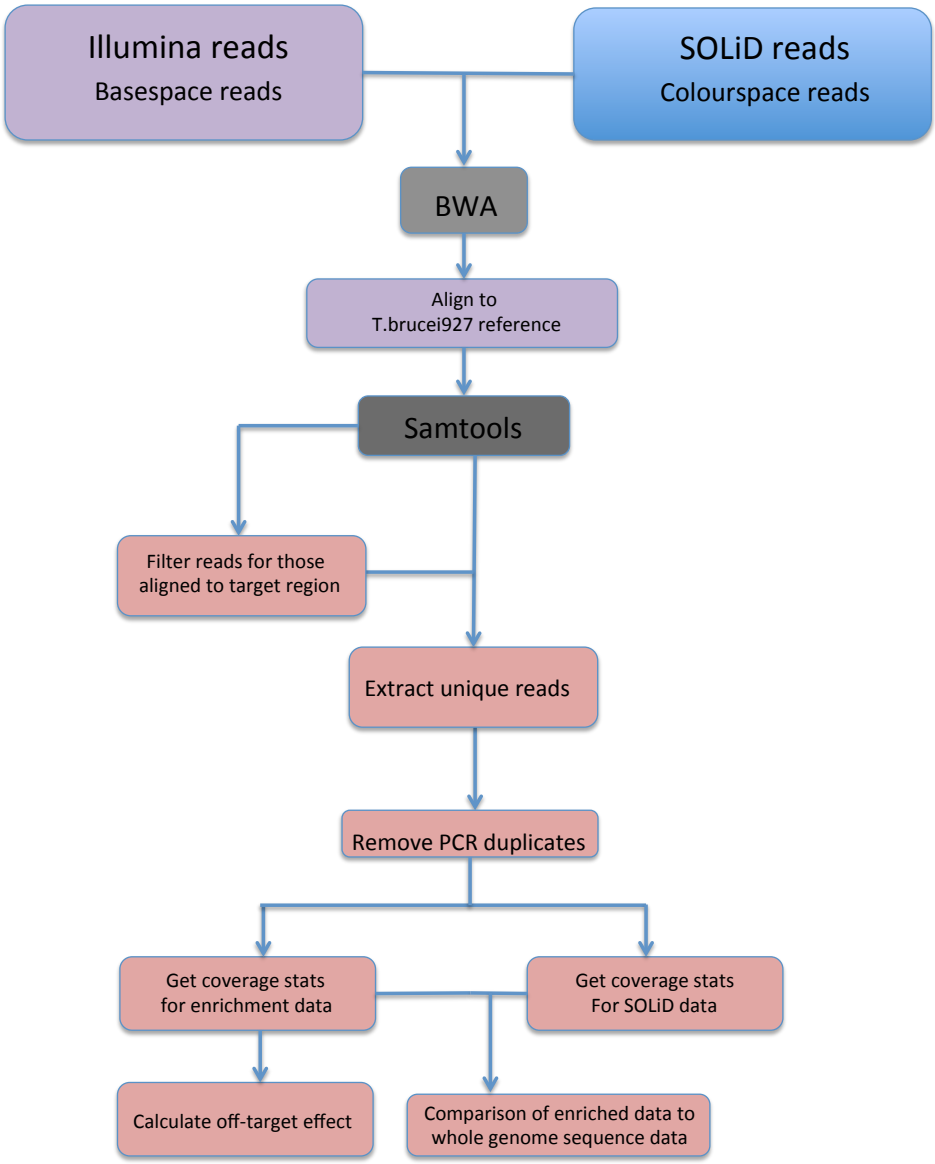

Figure 2.4: Overview of mapping strategy in BWA. Both Illumina and SOLiD reads were mapped using BWA aln to Tb927 reference v8.1 Mapped reads were then filtered so that PCR duplicates and non-uniquely mapping reads were removed. The BAM file created in Samtools was filtered for the intervals in the target region. Both the reads from the target intervals only, and the whole genome were compared to observe the off target effect.

2.2.16 SNP calling

2.2.16.1 GATK

SNPs were called using the Genome Analysis Toolkit (GATK) version 3 using the process outlined in Figure 2.5 below. The SNPs were called on the BAM files generated by BWA, that had been filtered (only uniquely mapping reads and no duplicates). The BAM file was locally realigned using the GATK walkers RealignerTargetCreator and IndelRealigner. RealignerTargetCreator identifies regions that need to be masked prior to realigning by identifying INDELs. This is because misalignments near INDELs are often mistaken for SNPs. Raw SNPs were then called from this realigned BAM using the UnifiedGenotyper walker. These SNPs were then filtered using the VariantFiltration walker, and SNPs that were hard to validate, were in a SNP cluster or had a quality score (MQ0), were identified and removed in subsequent analysis. SNPs were also filtered by coverage, this varied per dataset, and the parameters for this are specified in the analysis. SNPs were considered hard to validate if they had a low depth and a quality score of lower or equal to 40. SNP clusters were defined as 3 or greater SNPs within a 10bp window, and considered sequencing errors. Within this step, the zygosity of the SNPs was also determined. SNPs were called within and outside the target using the appropriate BAM file.

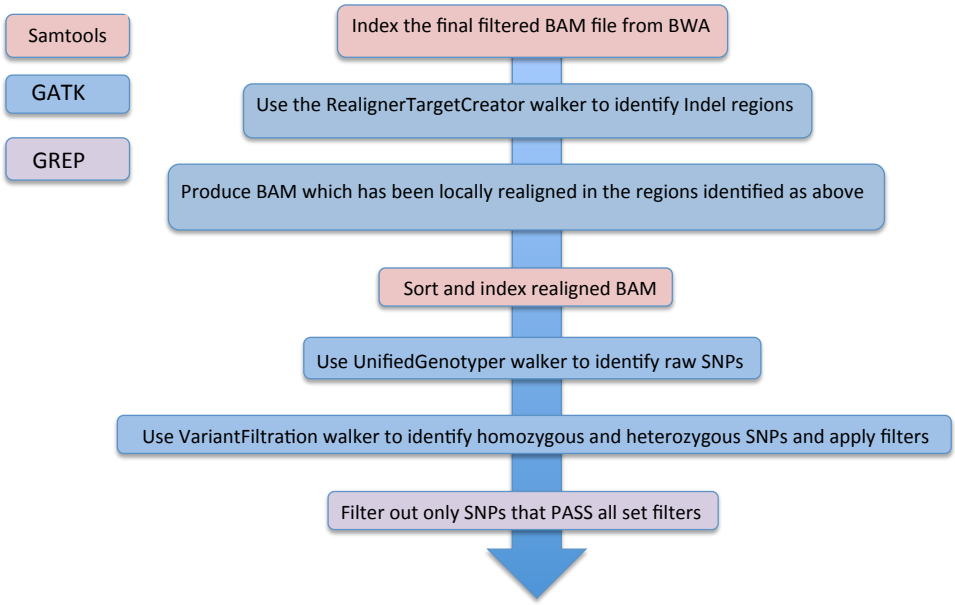

Figure 2.5: Overview of commands used to generate final SNPs. Steps shown in pink, blue and lilac show use of Samtools, GATK and GREP to complete stages respectively.

The uniqueness of the SNPs between strains, and WGS/enrichment comparison was done by looking at the SNPs position and genotype using Vcf-Compare and Vcf-stats.

2.2.16.2 Remapping the data using the SNPs called

In order to determine whether the SNPs generated in the WGS data were the same, the Mutate_reference.pl script provided by Laura Gardiner, was used to incorporate the SNPs found in the WGS data into the reference. This was done for each strain, and then once the SNPs were incorporated, the Illumina reads from the enrichment libraries were mapped to the altered reference, and SNPs called. SNPs were generated and filtered as previously mentioned, and so these SNPs showed SNP differences between the SOLiD and Illumina data for the same strain.

2.3 Results and discussion

2.3.1 Mapping stats over the entire Tb927 reference in the enrichment and WGS data

Table 2.4 shows the mapping of each strain (B and E) to the entire Tb927 version 8.1 reference in the WGS data and the enrichment libraries for both designs. The mapping percentage shows the total percentage of the trypanosome DNA in the sample. However this data also includes reads mapping to the Tb927 reference, which are not within the target region. This data can be used to determine the amount of off target data, and can be used to see whether the percentage of off target data is approximately uniform across strains.

Design one was based on the enrichment of 985 targets, design two, 731 targets, and so we would not expect to see the data to map to much greater than ~12% and ~8% of the whole genome in designs one and two respectively because this would indicate a high proportion of off target reads. The starting percentage of host DNA in the sample prior to enrichment is also very high (>99%) and so we would also expect a significant proportion of the reads to map to the host rather than the parasite. This can be observed in Table 2.4. Unfortunately the library for strain E in the second design

performed poorly compared to the other strains sequenced, and so had a much lower overall mapping and coverage. However a high percentage of the non-filtered reads do map to the *T. brucei* genome (>70% for the enrichment libraries excluding strain E in the second design). A high proportion of the uniquely mapping reads were removed following subsequent removal of PCR duplicates, however this is to be expected as a result of the high number of cycles needed during library preparation, and MDA prior to library preparation.

Due to the data being generated on the Miseq for the first design runs, and the Hiseq for the second design, there are far fewer reads for the first design; however for both, due to the small design region, the number of high quality reads is usable. If we compare this to the whole genome sequence data, which was sequenced from pure trypanosome cultures, we can see that enrichment sequence data doesn't adversely affect the quality from a mapping perspective. Despite far less amplification prior to sequencing, a significant number of reads are still filtered out due to low quality, however this is expected with older chemistries such as SOLiD (Ratan et al., 2013).

Figure 2.6 shows the average coverage per chromosome, and the percentage of the chromosome mapped to for both enrichment designs and the WGS data. Figure 2.6A-C show the mean coverage per chromosome, with the depth of coverage for B17/Strain E shown in blue, and Z310/Strain B shown in red. A shows the mean coverage in the WGS data, B the mean coverage in the first capture design and C, the mean coverage in the second capture design. These show that the mean coverage is actually higher in the first design than the WGS data, and is only slightly decreased in the second design compared to the WGS data.

Figure 2.6D-F show the percentage mapping per chromosome, with D-F representing the percentage mapping in the WGS, first design and second design respectively. The mapping percentage is highest within the WGS data as anticipated, and the smaller size of the design in two compared to one is also reflected in the percentage of chromosome covered. Z310 had a greater number of off target reads in design two compared to B17 and this is reflected in Figure 2.6.F

The capture array was designed to enrich each chromosome approximately equally, and so the on target percentages for each chromosome should be approximately 12% for the first design and 8% for the second design (except for chromosome 11 which

should have a lower percentage in design two). However because this data shows the on and off target data, we would expect the total mapping percentage per chromosome to be higher than the expected, to account for off target effects. The actual percentages that should be covered per chromosome for each design are shown in Tables 2.1 and 2.2.

The coverage is relatively equal between chromosomes and the mean coverage is also relatively consistent between strains, with the off and on target reads covering on average ~20% of each chromosome. The target region was largest for chromosome 1,2 and 3, and Figure 2.6 shows the highest percentage of the chromosome covered in these, which suggests uniform off target effects throughout the genome. The lowest mapping percentage was seen in chromosome 11, and this reflects the design. Chromosome 11 is the largest chromosome and has a greater percentage of repetitive sequence, and so the target region covers a much smaller percentage of the whole chromosome length. It also performed poorly compared to the other chromosomes, and so the second design had a disproportionally smaller target region for design two on chromosome 11.

Table 2.4: Shows the read counts and percentage of total reads mapped to the Tb927 v8.1 reference in the enrichment data for both designs, and WGS data for strains E and B. A high percentage of unfiltered reads mapped to the Tb927 reference (greater than 70% for all enrichment libraries excluding library E). The difference in the uniquely mapped reads and mapped reads is accounted for by the number of not only PCR duplicates but also incorrectly paired but mapped reads.

Data type	Target enrichment data								Whole genome sequencing			
	1st design				2nd design							
Zymodeme group	Z310		B17		Z310		B17		Z310		B17	
Isolate	B		E		B		E		B		E	
Total reads	7,471,188	Total reads (%)	6,663,392	Total reads (%)	38,438,548	Total reads (%)	29,348,266	Total reads (%)	117,196,475	Total reads (%)	121,086,789	Total reads (%)
Mapped reads	5,804,155	78	4,765,272	72	34,446,987	90	4,713,550	16	32,183,855	27	32,259,503	27
Unmapped reads	1,650,773	22	1,898,120	28	3,991,561	10	24,634,716	84	85,012,620	73	88,827,286	73
Uniquely mapped reads	4,168,802	56	3,284,132	49	26,522,280	69	3,484,764	12	20,707,438	18	20,787,940	17

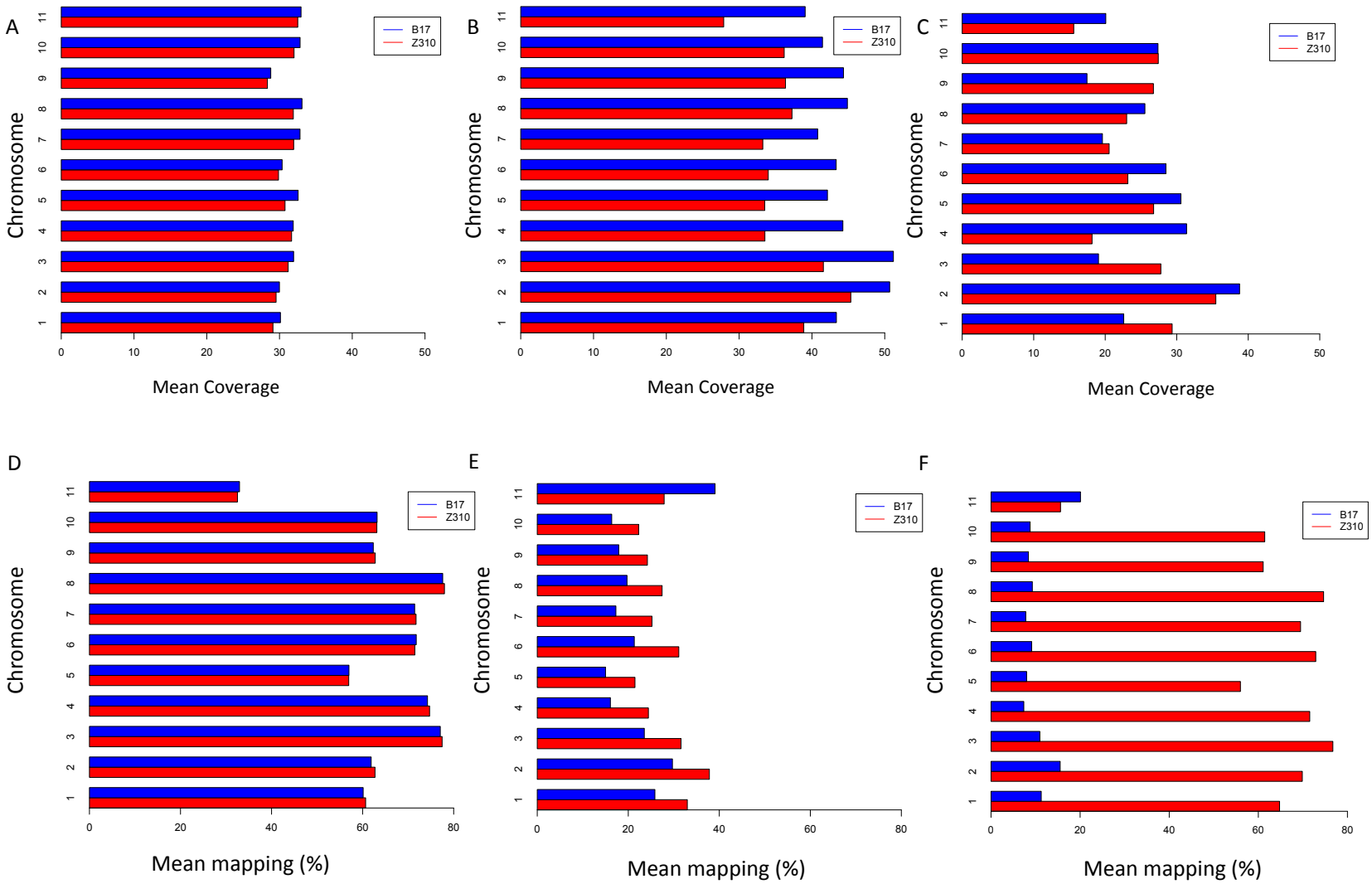

Figure 2.6: A-C show the mean coverage per chromosome of the WGS data, enrichment data from design one, and enrichment data from design two, respectively. Strain E is the representative B17 strain and is shown in blue, strain B is the representative Z310 strain and is shown in red. D-F show the percentage of the chromosome mapped in WGS and enriched data in the same order as A-C. The reads shown here are mapped to the entire Tb927 v8.1 reference, hence this data also shows off target data, which may non-uniformly enrich across non-target regions

2.3.2 Mapping stats for enriched samples over target region in the first and second design

The reads from the enrichment data that mapped within the target region for both designs was used so that SNPs could be called from data within the same region, because different libraries could be enriching off target regions non-uniformly. SNPs called against data mapped to off and on target regions showed a large difference in the number of heterozygous SNPs between strains E and B, an effect that could easily of been produced by one strain enriching different off target regions.

Table 2.5: Shows the read counts and mapping percentages for strains B and E within the target region for both designs. This table also uses the data from Table 2.4 to calculate the percentage of on target data.

	Design target region one		Design target region two	
Strain	B	E	B	E
Total reads	7,471,188	6,663,392	38,438,548	29,348,266
Number of reads mapped on and off target	5,804,155	4,765,272	34,446,987	4,713,550
Percentage of on target reads (%)	80	78	83	83
On target stats				
Number of reads mapped on target	4,661,397	3,694,601	28,734,337	3,925,427
Percentage of total reads mapped to target (%)	62	50	75	13
Number of correctly paired mapped reads	4,330,694	3,381,152	27,537,191	3,762,260
Percentage paired and mapped (%)	58	52	72	13
Uniquely mapping reads	3,779,170	2,936,547	24,941,554	3,366,468

As seen in Table 2.5, a high percentage of the reads mapped on target in design one (between 78-80%) and this on target mapping increased marginally with the second design (83%). This was consistent even in library E second design, where the library appears to not have hybridized as well. Even for this library, because of the small target region (2.0Mb) in the second design, and the read length of 100bp, there is sufficient coverage despite a small percentage of the reads uniquely mapping. The redesigned target region was based on targets that performed well, and the greater percentage of on target reads illustrates that design is an important factor in the performance of an

enrichment experiment. A high percentage of the reads were removed after identifying duplicates, this is often seen in samples with low initial amounts of DNA. Using WGA can exacerbate this.

Table 2.6 shows how this mapping data correlates to the coverage across the targets in each design. Here the coverage is based across whole gene sequences rather than per 120 nucleotide bait. The percentages of genes mapped to in column 1 are from the non-filtered mapping, the subsequent columns are generated from the data filtered for uniqueness and PCR duplicates. In the first design there were 985 genes in total targeted across the 11 megachromosomes, and greater than 92% of the gene targets were uniquely mapped to in both of the first design libraries. This was increased to greater than 99% in the second design, where there were 731 gene targets. 70-75% of the gene targets uniquely mapped to in design one had a coverage of greater than 50, whereas for the same strain in design two, because almost all targets were uniquely mapped to, over 99% of the total target genes had a coverage of greater than 50x.

In the unfiltered data, there are several extreme outliers in coverage, in design one certain targets had coverage greater than 5000 x (not shown), and this illustrates how important data filtering is, because these are not seen in the data filtered for PCR duplicates and non uniquely mapped reads. Although the number of genes uniquely mapped to in Strain E design two is comparable with Strain B in design two, the number of genes uniquely mapped to with a high level of coverage is much lower, with only 68% of the target genes having a coverage greater than 5. However, as discussed in Chapter 3, this is a reflection of poor hybridization in the sample, rather than poor design, because successful enrichment as shown in Strain B was seen in the other libraries used in the design and discussed in Chapter 3.

The mapping of the WGS data to these targets is shown to illustrate that the WGS data is suitable for mapping across all of these targets, are uniquely mapped to, and can be compared with the enrichment data in the same region.

Table 2.6: Shows the coverage across the target region in both enrichment designs. In design one, there were 985 genes and 731 in design two. Neither strain sequenced had coverage over all 985 genes, however the vast majority of the target region was covered (between 93-96%). In design two a greater percentage of the target region was mapped to (99.9%). However, the genes with a greater than 5 fold coverage was similar between both strains in design one, but poor performance in strain E shows a greater number of genes (32%) having a coverage of less than 5 in design two. The whole genome sequence data mapped to all of the genes in both targets. WGS data is mapped to 986 targets because this includes the 985 in target one, and the additional gene included in the second design but not included in the first

Strain		Genes mapped to	Genes mapped to (%)	Genes uniquely and no duplicates mapped to	Genes uniquely mapped to (%)	Genes with >5x coverage (Unique no duplicates)	Genes with >50x coverage (Unique no duplicates)	Genes with >1000x coverage (Unique no duplicates)
Design one	B	941	96	931	95	861	701	0
	E	916	93	901	92	857	656	0
Design two	B	730	99.9	730	99.9	730	727	0
	E	730	99.9	729	99.8	500	162	0
Whole genome sequence data 986 genes in both designs	B	986	100	986	100	N/A	N/A	N/A
	E	986	100	986	100	N/A	N/A	N/A

2.3.3 Distribution of coverage per gene in design one

The coverage per gene is also illustrated in Figures 2.7 and 2.8 below and shown per design and strain for comparison. In the first design, you can see from both Table 2.6 and Figures 2.7 and 2.8, that the coverage profile is similar between the two strains. The coverage slopes off, as is expected and the majority of genes have a coverage of less than 1000x. Due to the lower frequency of these extreme outliers, with in excess of 5000x coverage, only those with a coverage of less than 1000x are shown in Figures 2.7A and C. There are also a high frequency of genes that have a much lower coverage in the data for this design. This illustrates that although the majority of the genes have a depth of coverage of less than 1000x, targets that perform poorly either have a low coverage of less than 5 x or over enrich excessively (over 5000x coverage).

Figure 2.7A and B show that the majority of coverage is between 0-200x in target one. This is quite a high level of coverage considering the data was generated from a single Miseq run, and the initial percentage of target DNA was so low. More importantly, this is consistent between the two strains. However there is still quite a lot of variability in performance between targets within the same design, and ideally this would be reduced by improvements to the design. In Figures 2.7C and D, only the quality filtered data is shown, and this removes the outlying data, with the coverage shown primarily 1-250x.

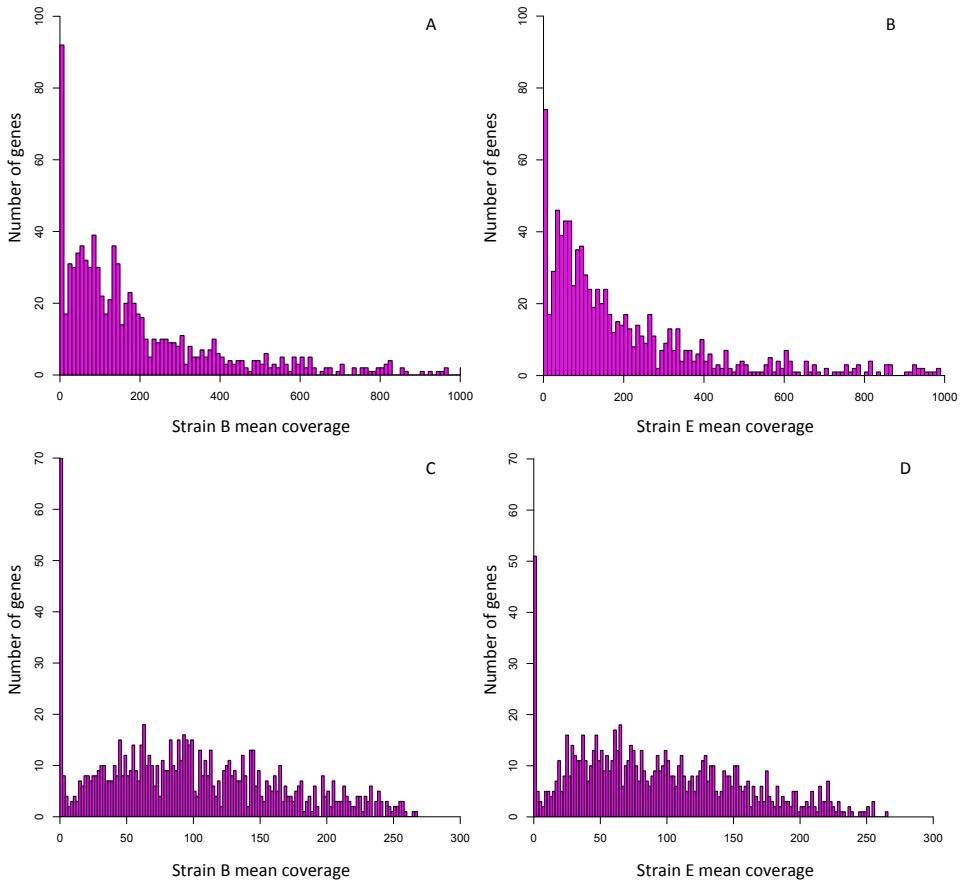

Figure 2.7: There are several genes with excessive coverage of greater than 5000 x. These are not shown because the majority of the genes have a coverage less than 1000x, and including these outliers reduces the ability to see the gene coverage distribution. A and B show the mean coverage for genes with a coverage between 1-1000x. The tailing of coverage is observed in both, strain B is shown in A, strain E in B. C and D show only the quality filtered data, and despite a peak of low coverage genes, the majority of the data falls within 1-250x, this peak is more exaggerated in strain B.

2.3.4 Distribution of coverage per gene in design two

Figure 2.8A shows the coverage for all genes in the second design for strain B,with no filtering of outliers. The distribution of coverage is much tighter than seen in Figures 2.7A for the first design. Figure 2.8B illustrates that strain E had much lower overall coverage compared to strain B in the second design, and so only coverage up to 1000x is shown. The target region for the second design was generated based on the genes that performed the best in the first design, and so it is expected that there would be less variability in the performance of the probes. Strain E shows a distribution very similar to that seen in the first design, with a tailing off the of the coverage, and a significant number (~30% in this library) having a much lower coverage than the rest of the genes in the target region. However as mentioned before, this is more to do with individual library performance than the design region, and strain B performs well, with tightly distributed high coverage across all the genes within the target region. A similar performance was shown in the other strains that were also sequenced with this design. Figure 2.8C and 2.8D show the filtered data for design two.

The redesign of the target region has obviously had an effect on the overall success of the enrichment, because unlike in the first design, there are no extreme outliers in the filtered data, or targets that excessively over enrich. Instead for both libraries, gene coverage lies between 0-200x for all targets. Even though there is low coverage for strain E, the distribution is quite tight between genes, and similarly with strain B, where the majority of coverage is between 150-200x.

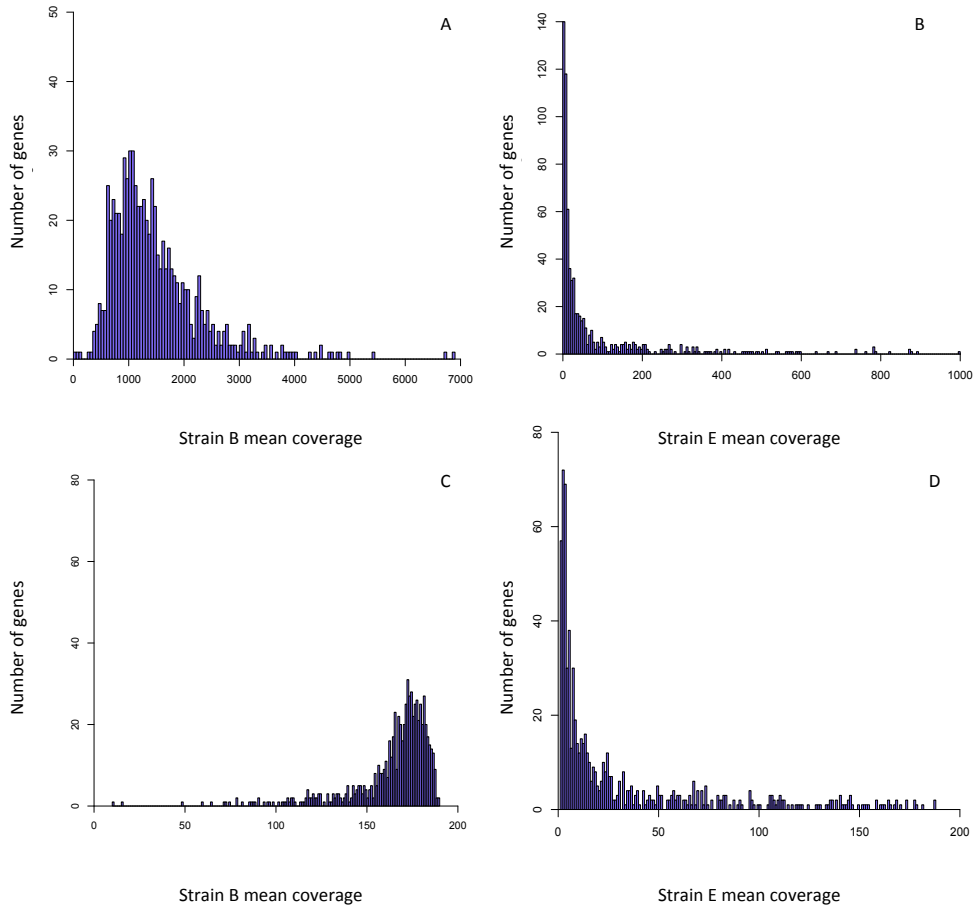

Figure 2.8: A shows the unfiltered coverage for genes in the second design for strain B and B shows the unfiltered coverage for genes in the second design in strain E. Due to much lower overall coverage; only genes up to a coverage of 1000x are shown in B. C and D show just the filtered data for srains B and E respectively. The majority of the coverage lies between 0-200x coverage. Strain E in B and D show a profile similar to that seen in the first design, with more variability to the coverage, and a tailing off in the data. However overall the data has a much tighter distribution and no extreme outliers are seen. In C, unlike in strain E, the coverage is much higher, and the distribution of coverage is much tighter than is seen in the first design. For the majority of genes, the coverage falls between 150-200x.

The lower degree of inter-target variability in design two is shown in Figure 2.9. It shows the distribution of coverage across each design for the two strains, which is more variable in both strains in the first design, than observed in the second design. This shows that the redesign of the target region has had a significant effect on the performance of the probes, with all of the probes enriching relatively uniformly in the second design. The decrease in variability between designs is reflected by the comparative length of the whiskers and the inter quartile range (IQR), which are both greater in the first design for both strains. In Figure 2.9, the whiskers represent the furthest data point within 1.5x the IQR. In the second design, the IQR is greatly reduced

for both strains, but several points are still outliers of the whiskers, and these are shown above and below in strain B and E respectively.

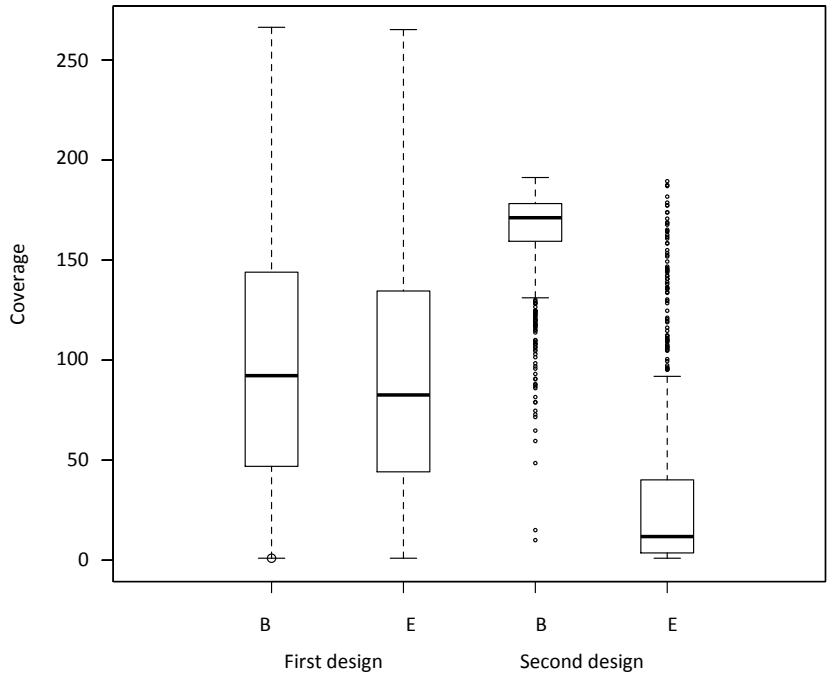

Figure 2.9: This shows the coverage for all of the genes within the filtered data for both strains in both designs. As discussed above, the distribution of coverage is tighter in the second design than the first design. This is reflected in the inter quartile range (IQR), which is much greater in the first design than second. The whiskers represent the largest and smallest data points up to 1.5 x the IQR. Outliers of this are shown in the second design, above and below strains E and B respectively.

Despite a very variable coverage across targets, the majority of the targets for design one have a good level of coverage, with the 1st quartile being in excess of 50x coverage for both strains. However the high degree of variability is evident from the large difference in between the mean and 3rd quartile.

In the second design, the enrichment is more uniform, despite the outliers indicated in Figure 2.9. The poor enrichment of strain E is evident from the low coverage compared to B in the second design. Strain B acts comparably to the other strains used in the second design, which are discussed in Chapter 3, and the tight distribution in coverage is reflected in the considerably smaller IQR and whiskers.

2.3.5 Individual target performance across libraries

From the difference in performance in terms of more even coverage and greater on-target percentage of reads, it is evident that probe design has a significant impact on the performance of the design. The target region for the second design was based on the targets which had a coverage of between 5-500 in both these strains and additional strains discussed in Chapter 3 in design one. However, in the first design, how comparable was the performance of individual targets between samples? Figure 2.10 demonstrates how the targets where enriched in Strain E and B in the first design. Each point represents a gene target, and the targets are ranked based on their coverage. There is a strong correlation between how the targets performed in both strains. In particular there is a cluster of targets at either end of the ranking, where poorly/over-enriched probes perform the same across the strains. This shows that the performance is reproducible and can be controlled by design.

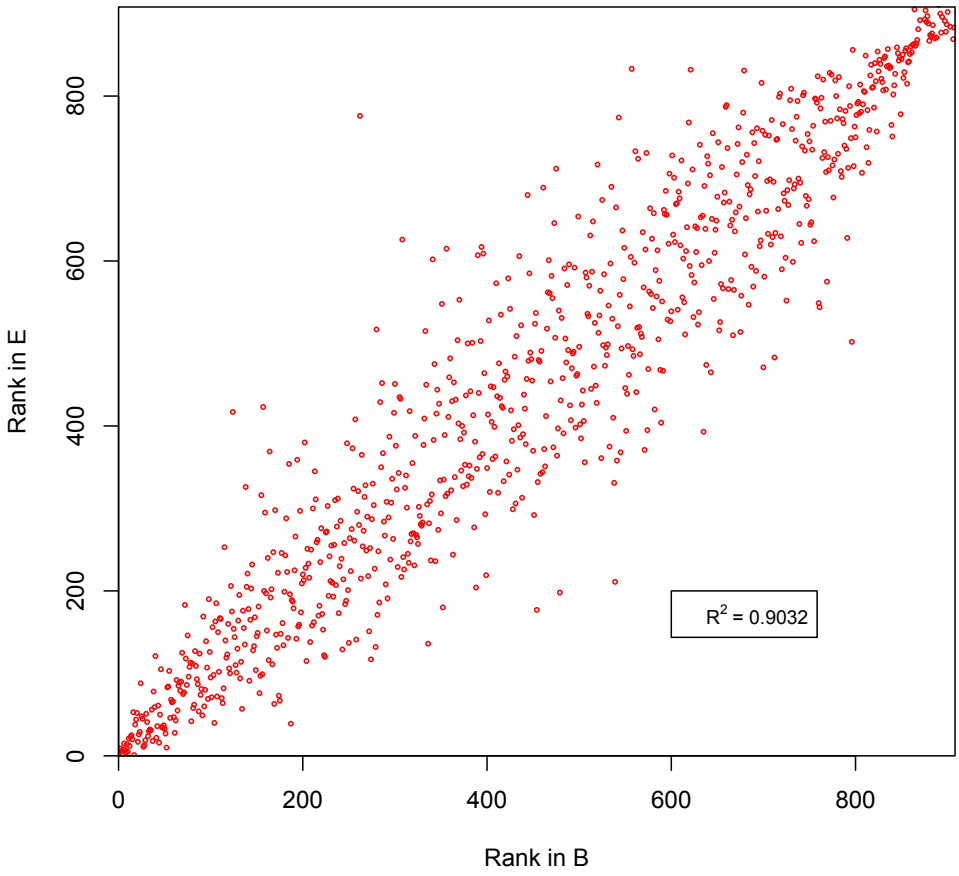

Figure 2.10: The performance of the gene targets in design one were ranked in both strain B and E and compared to see whether performance was reproducible between samples. As shown by the R2 value, there is a strong correlation between the performance of targets between samples.

In design two, due to less variability in the coverage between targets, this correlation is not observed. In the second design, there is very little difference in the coverage between genes ranking between 1 and 419 in Strain B in the second design (190- 170x coverage) and so large fluctuations between the ranked performance in the second design equate to very little actual difference in the coverage. In contrast, in the first design, although some of the probes had a higher coverage, this was far less uniform, with only 10 gene targets out of 985 having between 280-260 x coverage, both example intervals of 20x coverage being for the top ranked targets in each design.

2.3.6 SNP analysis of whole genome sequence and enriched data

Both the whole genome sequence SOLiD data and enrichment data were mapped using BWA version 0.5.9-r16. SNPs were subsequently called using GATK version 3.0. The data was mapped to both the whole reference Tb927 version 8.1, and the subsequent BAM files filtered to call SNPs within and outside of the target region. Strains E and B are the two strains used for whole genome sequencing and the enrichment array, and represent the B17 and Z310 zymodeme groups respectively.

2.3.7 SNP analysis against entire Tb927v8.1 reference

2.3.7.1 Analysis of SNPs within the whole genome in the WGS data

Due to the high degree of similarity between *T. brucei* subspecies, and because Strain E and B are both Ugandan *T.b. rhodesiense* strains, a low number of SNPs and a high percentage SNP similarity was expected , which is observed in Table 2.7. Over the entire genome, a total of 126,586 SNPs were found. The following SNP calls show the positions of the SNPs and do not reflect the genotype. Between the two strains 98,564 SNP positions were identical, which accounts for 86%/89% of the total SNPs called in strains B and E, respectively. These SNPs were generated using the GATK pipeline mentioned previously, and are the filtered SNPs, which had a depth between 5-100. As shown in Figure 2.6, the mean coverage in the WGS data is 30x; no SNPs were discovered above the 100x threshold. Similarly, the numbers of SNP positions unique to each strain were relatively similar, with strain B having 3.2% more unique SNP positions than strain E.

Table 2.7 Summary of SNP calls generated in GATK and then analysed in vcftools compare. This data demonstrates the very high degree of similarity between these two strains. SNPs were generated after mapping to the Tb927 version 8.1 reference, available from (www.tritrypdb.org). This was done in VCFcompare, the SNPs were called from the filtered data using the pipeline mentioned in Figure 2.5. These SNPs had a depth between 5-100

	SNP call summary			
	Sites unique to isolate	% Unique to isolate	Sites shared with other isolate	% Sites shared with other isolate
B17	11,939	10.8	98,564	89.2
Z310	16,083	14.0	98,564	86.0

The VarieatFiltration walker in GATK was used to determine the zygosity of the SNPs, and so this could be used to determine whether the SNPs were the same zygosity at the same position. This is shown in Table 2.8 and done using VCFcompare. A high proportion of these SNPs have the same zygosity at the same position, and with 93% (91,546 SNPs) out of the 98,564 SNPs shared between the strains, see Table 2.7, have heterozygous and homozygous SNPs at identical positions. Table 2.8 also shows the degree of uniqueness in the heterozygous/homozygous SNPs, and in both strains there are a greater percentage of heterozygous SNPs unique to one strain (~30%) compared to the homozygous SNPs (~11%). The ratio of heterozygous to homozygous SNPs is near identical also.

Table 2.8: This shows the number of heterozygous and homozygous SNPs per strain. It also illustrates that the percentage of unique SNPs per zygosity is relatively uniform. The degree of uniqueness is higher for the heterozygous SNPs (~30%) than for the homozygous SNPs (~11%). Table 2.7 showed that there were 98,564 SNPs shared between the two strains. 91,546 of these (93%) have the same zygosity at the same position.

	B17	**Z310**		
Homozygous AA count	72,398 (66%)	73,891 (64%)		
Heterozygous SNP count	38,029 (34%)	40,675 (36%)		
			Shared	Total
Homozygous AA sites unique to strains	7,949 (11%)	9,442 (12.8%)	64,449 (79%)	81,840
Heterozygous SNP sites unique to strains	11,008 (28.9%)	13,659 (33.5%)	27,097 (50%)	51,764
Total SNPs	110,427	114,566	91,546 (93%)	

2.3.7.2 Analysis of SNPs within the whole genome in the enriched data

SNPs were subsequently generated for the enrichment data in order to see whether the SNPs seen in the enrichment data correlated with the whole genome sequence data. SNPs were generated using the same BWA and GATK parameters as for the whole genome sequence data. Unlike the WGS data, the variants were filtered for SNPs with a depth between 5-300 for both design one samples and strain B design two, and 2-300 for design two strain E data. The lower parameter for design two strain E was set to 2 to find low coverage SNPs and still estimate their zygosity. Mean coverage is higher in the enrichment data, and so the higher threshold accounts for this. Table 2.9 shows the number of SNP positions identical and unique between the two strains for both enrichment designs. These are SNPs called against the entire Tb927 reference, not just the target region.

Unlike in the WGS data, Table 2.9 shows that in the enrichment data there appears to be less conservation between the SNPs for the two strains, with between 17-32% unique SNP positions in the first design. This is seen in the second design also, with the SNPs unique much higher ~90%, however this difference in design two is compounded by both low level coverage in some of the genes in strain E, but also by a much greater number of off target reads in strain B. If strain E in design two had performed equally to B, the percentage uniqueness would be much closer to that observed in the first design (20-30%). This shows that off target enrichment is very non-uniform across samples, because we would expect the degree of similarity seen with the WGS data.

Table 2.9: Breakdown of the SNPs that are unique to each strain, and shared. Whereas in the WGS data the two zymodeme groups appeared to have approximately equal numbers of homozygous and heterozygous SNPS, here the two strains used in the WGS data, E and B, appear to have very different numbers of SNPs, indicated non-uniform enrichment for non target regions.

		SNP call summary			
Design		Sites unique to isolate	% Unique to isolate	Sites shared with other isolate	% Sites shared with other isolate
One	B	8109	31.5	17,667	68.5
	E	3733	17.4	17,667	82.6
Two	B	71,655	87.6	10,138	12.4
	E	1,427	12.3	10,138	87.7

As before, VarientFiltration was used to determine the zygosity of the SNPs, and to compare whether the zygosity was preserved between strains at the same position. Similarly to Table 2.8, Table 2.10 shows very similar homozygous to heterozygous

ratios in the enrichment data. Interestingly, in the first design, the inclusion of the non-target data shows more than double the number of both heterozygous and homozygous SNPs compared to strain E for the same design, however the ratio is approximately equal. In the second design for strain E, the percentage of heterozygous SNPs is much lower, ~15% of the SNPs compared to the ~35% seen in the other strains/design. This is either due to low coverage making it harder to successfully determine the zygosity of SNPs, or poor hybridization resulting in an underrepresentation of heterozygous SNPs for this library. These inter-sample differences suggest unevenness of enrichment across non-target regions. Despite this, conserved zygosity in SNPs between strains is not much lower than in the WGS data (~88-93%).

Table 2.10: Shows the SNPs unique and shared between strains for both designs in relation to their zygosity, in the enrichment data. Unlike before, there is a much higher degree of uniqueness between these strains compared to the WGS data. Only ~40% of the SNP positions are shared. The disparity between the WGS data and the enrichment data suggests unevenness in the enrichment across non-target regions.

	First Design				Second design			
	B	E			B	E		
Homozygous AA count	17,120 (66%)	8,651 (65%)			54,235 (66%)	9,889 (86%)		
Heterozygous SNP count	8,656 (34%)	4,718 (35%)			27,564 (34%)	1,676 (14%)		
			Shared	Total			Shared	Total
Homozygous AA sites unique to strains	5,406 (31.6%	2,071 (15.0%)	11,714	19,191	46,383 (85.5%)	2,037 (20.6%)	7,852	56,272
Heterozygous SNP sites unique to strains	4,004 (46.3%)	2,963 (38.9%)	4,652	11,619	26,530 (96.2%)	642 (38.3%)	1,034	28,206
Total SNPs	25,776	13,369	16,366 (93%)		81,799	11,565	8,886 (88%)	

2.3.7.3 Representation of SNPs within the data

When including the non-target data, there appears to be unequal enrichment across non-target regions of the genome, and this resulted in an overrepresentation of heterozygous SNPs in strain B for design two. However it is also important to see whether the enrichment data is causing overrepresentations in the number of SNP varieties found, including the transversion/transition ratio. In Figure 2.11, the SNPs are plotted based on the genotypic change they cause.

Transversions, A>C, C>A, G>T, T>G, A>T and T>A, and transitions C>T, T>C G>A, A>G, generally occur unevenly in a SNP population, with transitions being more prevalent. However, despite slight differences, the ratios of each type of SNP are near identical between the enrichment and WGS data

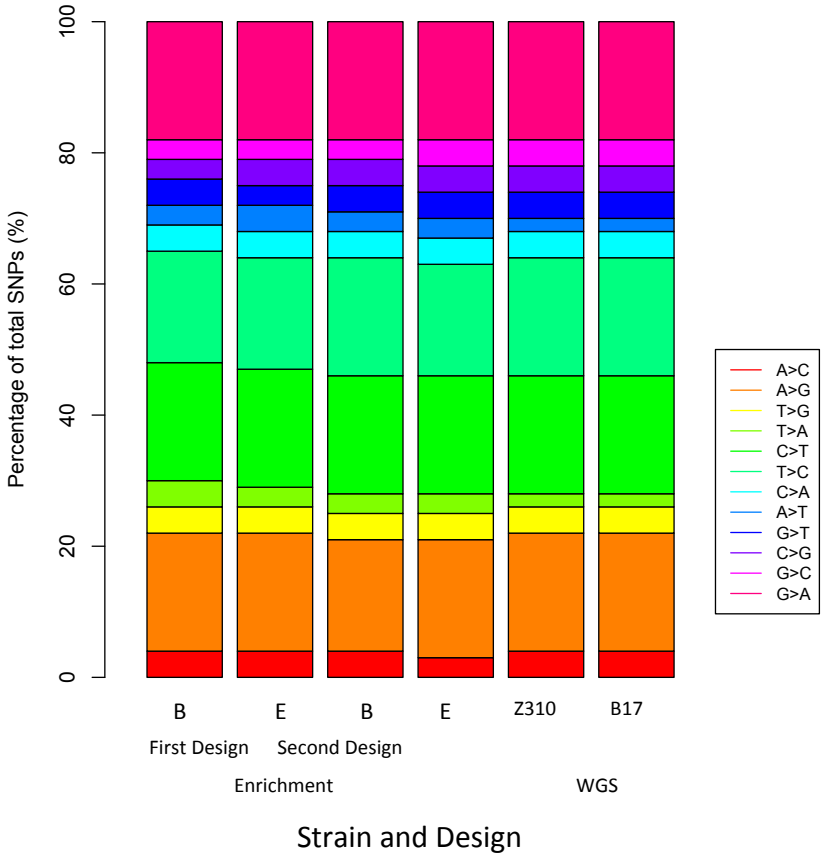

Figure 2.11: Shows the percentage of SNPs and their corresponding mutation, this shows that despite small differences in the numbers of SNPs called between the two strains, the percentage of SNPs that are transversions and transitions remain the approximately equal between the two sequencing methods.

2.3.8 Analysis of SNPs found within the target regions

As before, SNPs were generated using the same parameters in BWA and GATK, but were instead mapped against a custom Tb927 reference, for both WGS and enrichment data, in order to look just at SNPs within the target region for design one and two. SNPs shared between these strains could indicate *T.b. rhodesiense* specific differences, and those unique to a strain will be examined alongside other strains within the same zymodeme group to determine what defines a zymodeme group in Chapter 3.

2.3.8.1 SNPs found within the target region

As shown in Table 2.11, the high percentage of conserved SNP positions in the enrichment data ~90%, agrees with the degree of SNP conservation observed in the WGS data. Illumina's sequencing chemistry is more sensitive than SOLiD's chemistry, and so you would expect some SNP positions conserved within the Illumina data to be novel to those observed within the same region of the WGS data generated by SOLiD sequencing. However in Table 2.11, the number of SNPs discovered in the first design region is slightly lower in the enrichment data than observed in the WGS data. There are fewer conserved sites in the second design, however this is another artefact of strain E's lower coverage. However despite this lower coverage, the similar percentage identity, and the raw number of conserved and unique sites observed, compared to the WGS data, suggests that a high proportion of the SNPs within this target are still identified in spite of the low coverage.

Table 2.11: Summary of SNP calls generated in GATK and then analysed in vcftools compare. This data demonstrates the very high degree of similarity between these two strains. SNPs were generated after mapping to the custom Tb927 reference. SNPs with a depth of 5-100 are shown in the WGS data, and 5-300 for the enrichment libraries, except for strain E second design, these have a depth of 2-300.

Data			**Targeted SNP call summary**			
			Sites unique to isolate	% Unique to isolate	Sites shared with other isolate	% Sites shared with other isolate
WGS	First Design	Z310	400	7.9	4,656	92.1
		B17	389	7.7	4,656	92.3
	Second Design	Z310	941	10.1	8,419	91.5
		B17	783	8.5	8,419	89.9
First design		B	445	6.5	6,453	93.5
		E	685	9.6	6,453	90.4
Second design		B	5846	40.9	8,461	59.1
		E	588	6.5	8,461	93.5

Compared to the SNPs called over the entire reference, the homozygote to heterozygote ratio is higher within the target region by approximately ~7%, and this was observed in both WGS and enrichment sequencing data, see Table 2.12. As expected from the second design strain E data and Table 2.10, the homozygous SNP percentage was higher than seen within the rest of the data (85%). Despite the homozygous SNP count of this strain being only 30% lower than in strain B for the same design, the heterozygous SNP count was 69% lower than observed in strain B. In comparison, the WGS data mapped for the same strain over the second target region had 14% less homozygous SNPs than strain E, but still had 46% more heterozygous SNPs than seen in the enrichment data.

Conservation in the zygosity of the SNPs was comparable between the WGS and first design data (92-93%) and there were a greater number of total SNPs called compared to the WGS data. However strain E in the second design, which had a much lower overall coverage, had much lower conservation and fewer SNPs than in the WGS data.

Table 2.12: Shows the SNPs unique and shared between strains for both designs in relation to their zygosity, in the enrichment and WGS data. The heterozygote to homozygote ratio observed between strains is approximately equal, except in strain E second design, where the heterozygous SNPs appear to be underrepresented in the data. There is also a high level of congruence in the WGS and enrichment data in the percentage of SNPs that have conserved zygosity at a conserved site.

	First Design				Second design				WGS							
									First design				Second design			
	B	E			B	E			Z310	B17			Z310	B17		
Homozygous AA count	4,946 (72%)	5,125 (72%)			10,019 (70%)	7,703 (85%)			3,669 (73%)	3,578 (71%)			6,723 (73%)	6,702 (73%)		
Heterozygous SNP count	1,952 (28%)	2,013 (28%)			4,288 (30%)	1,346 (15%)			1,387 (29%)	1,467 (29%)			2,637 (27%)	2,500 (27%)		
			Shared	Total			Shared	Total			Shared	Total			Shared	Total
Homozygous AA sites unique to strains	295 (6.0%)	474 (9.2%)	4,651	5,420	3,285 (32.%)	969 (12.%)	6,734	10,988	357 (9.7%)	266 (7.4%)	3,312	3,935	693 (10.3%)	672 (10.0%)	6,030	7,395
Heterozygous SNP sites unique to strains	603 (30.9 %)	664 (33.0 %)	1,349	2,616	3,419 (80%)	477 (35%)	869	4,765	379 (27.3 %)	459(31.3%)	1,008	1,846	953 (36.1%)	816 (32.6%)	1,684	3,453
Total SNPs	6,898	7,138	6,000 (93%)		14,307	9,049	7,603 (90%)		5,086	5,045	4,320 (93%)		9,360	9,202	7,714 (91%)	

2.3.8.2 Representation of SNPs within the data

Removing the non-target regions from the data eliminates the majority of the unevenly enriched data. However, it was still important to both look at whether the zygosity was preserved within the target region, and also the different SNP varieties were not being over/under-represented in the enrichment data.

Figure 2.11 showed that over the entire dataset, there was no significant difference between the relative abundance of different SNP varieties, and this is also true for within the target region, as seen in Figure 2.12. The WGS data mapped to the target region acted as an approximation of the expected SNP proportions, and despite small differences, neither enrichment design deviated far relative to the WGS data.

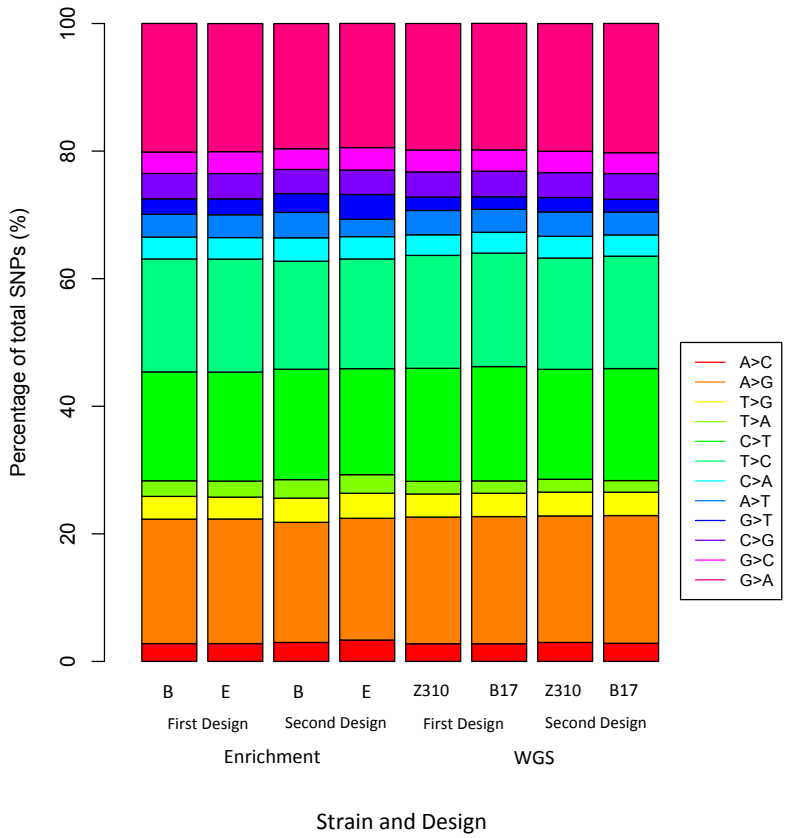

Figure 2.12: Shows the percentage of SNPs and their corresponding mutation. As previously seen in the data including the off target SNPs, this shows that despite small differences in the numbers of SNPs called, the percentage of SNPs that are transversions and transitions remain the approximately equal between the two sequencing methods. The WGS data for the target region can be used as an approximation to the expected SNP ratios.

2.3.9 Comparison of whole genome sequence data to enriched data

Table 2.13 provides an overview of the similarity/key differences found in the comparison of WGS and enrichment data. Each column shows the comparison of the performance in each strain/design combination of enrichment data to the WGS data. There is an increase in the total number of SNPs called within both target regions compared to the same region in the WGS data. In the second design, despite much lower coverage for strain E compared to the strain B, an increase in the total number of SNPs is still seen, with a 13% increase in the number of homozygous calls made compared to the WGS, but a decrease in the number of heterozygous calls made. This suggests that a high level of coverage is not necessary to capture the majority of SNPs within a region, however the number of heterozygous confidently called will suffer as a result. For the other three libraries, an approximate 30% in the total number of SNP calls was observed. An increase in coverage in genomic data should not increase the number of calls, only increase the validity and confirm the zygosity of the SNPs called, so this suggests these SNPs are unique to the enrichment data because of the increased sensitivity of the Illumina data compared to SOLiD.

More importantly, heterozygous/ homozygous SNPs do not appear to be highly over or underrepresented in the data compared to the WGS data, with the exception of strain E second design, which does have 12% decrease in the percentage of heterozygous SNPs called. This may be as a result of low coverage at positions of heterozygous SNPs, and so incorrect assignment of zygosity. The concordance between WGS and enrichment is greater than 80%, which was calculated by the percentage of SNPs remaining after remapping the enrichment data to the reference after SNPs found in the WGS data had already been incorporated into the reference, and removing SNPs known to be unique to the enrichment data. This is lower as expected in strain E second design, however as previously mentioned, despite poor coverage, 64% of the SNPs found in the WGS data were still found in the enrichment data.

Table 2.13: An overview of the differences found in the WGS and enrichment data in terms of total SNP calls and the zygosity of SNPs. Z310 refers to strain B, B17 to strain E. These data used to derive this table is provided in the appendices

	Percentage similarity (%) compared to WGS			
	First Design		Second Design	
	Z310	B17	Z310	B17
SNPs called in target region	26% increase	29% increase	35% increase	17% decrease
Number of heterozygous SNPs within the target region	29% increase	27% increase	39% increase	14% decrease
Percentage of heterozygous SNPs of total SNPs called (%)	1% decrease	1% decrease	3% increase	12% decrease
Number of homozygous SNPs within the target region	26% increase	30% increase	33% increase	13% increase
Percentage of homozygous SNPs of total SNPs called (%)	1% decrease	1% increase	3% decrease	12% increase
SNPs found in both enriched and WGS data	82% of WGS SNPs found in enrichment data	82% of WGS SNPs found in enrichment data	94% of WGS SNPs found in enrichment data	64% of WGS SNPs found in enrichment data
SNPs found in only the enriched data	2,754 (40%)	3,015 (42%)	5,475 (38%)	3,131 (35%)
SNPs after remapping	3,918	3,876	6,562	6,826
False positives, position conserved, genotype not conserved	1,164	861	1087	3,695
Concordance between WGS and enrichment	83%	88%	93%	60%

2.3.10 Illustrated examples of SNPs unique to enrichment sequence data

As shown above, more SNPs are called within the enrichment data than within the WGS data. In order to show these variants are not artefacts of allelic drop out, representative SNPs are shown in Figures 2.13 and 2.14 and show heterozygous and homozygous SNPs found in regions where the depth of the WGS and enrichment data exceeds 20x. The identification of heterozygous SNPs is particularly important because these can be observed following allelic drop out, rather than being true variants. Heterozygotes can also falsely be called within regions of low coverage.

The examples shown in Figure 2.13 and 2.14 were visualized using IGV and using the alignment of only quality filtered reads from strain B from the second design, and the alignment of the corresponding WGS data from the Z310 strain. As is indicated in Figure 2.13, the coverage is greater at this SNP position in the enrichment data, and the SNP in the center is heterozygous, with approximately half of the reads having the reference allele, C, and the remainder the alternative allele, A. The coloured bars on the coverage track show the ratio of C/A alleles by the percentage of the bar coloured in blue/green respectively. Due to the high depth of coverage, not all reads aligned are shown in the enrichment data. The same is shown for the heterozygous SNP to the right. The depth in the WGS data was 23x and 42x at the positions of the SNPs in the enrichment data, left to right respectively. The same positions within the enriched data had a depth of 342x and 348x respectively. Although two of the reads in the left most SNP are shown to have the alternative allele, this is far lower than the threshold for a heterozygous SNP and would be discarded based on the WGS data alone. Similarly, only one read aligned to the right most SNP had the alternative allele, and this would also not be considered a valid SNP based on the WGS data alone.

Figure 2.14 shows homozygous SNPs called in the enrichment data but not found in the WGS data. Three homozygous SNPs are shown in Figure 2.14, with 279, 273 and 214 reads unambiguously aligned across left to right and 91 %, 91% and 86% of the reads contained the alternative allele, respectively. The coverage in the WGS data at these positions was 33x. Due to their homozygous nature, you would expect these SNPs to at least be represented in a proportion of the WGS reads. Both the homozygous and heterozygous examples suggest that the enrichment sequence data is more sensitive to calling SNPs, and that the increased rates of detection are not artefacts or miscalls related to allelic dropout.

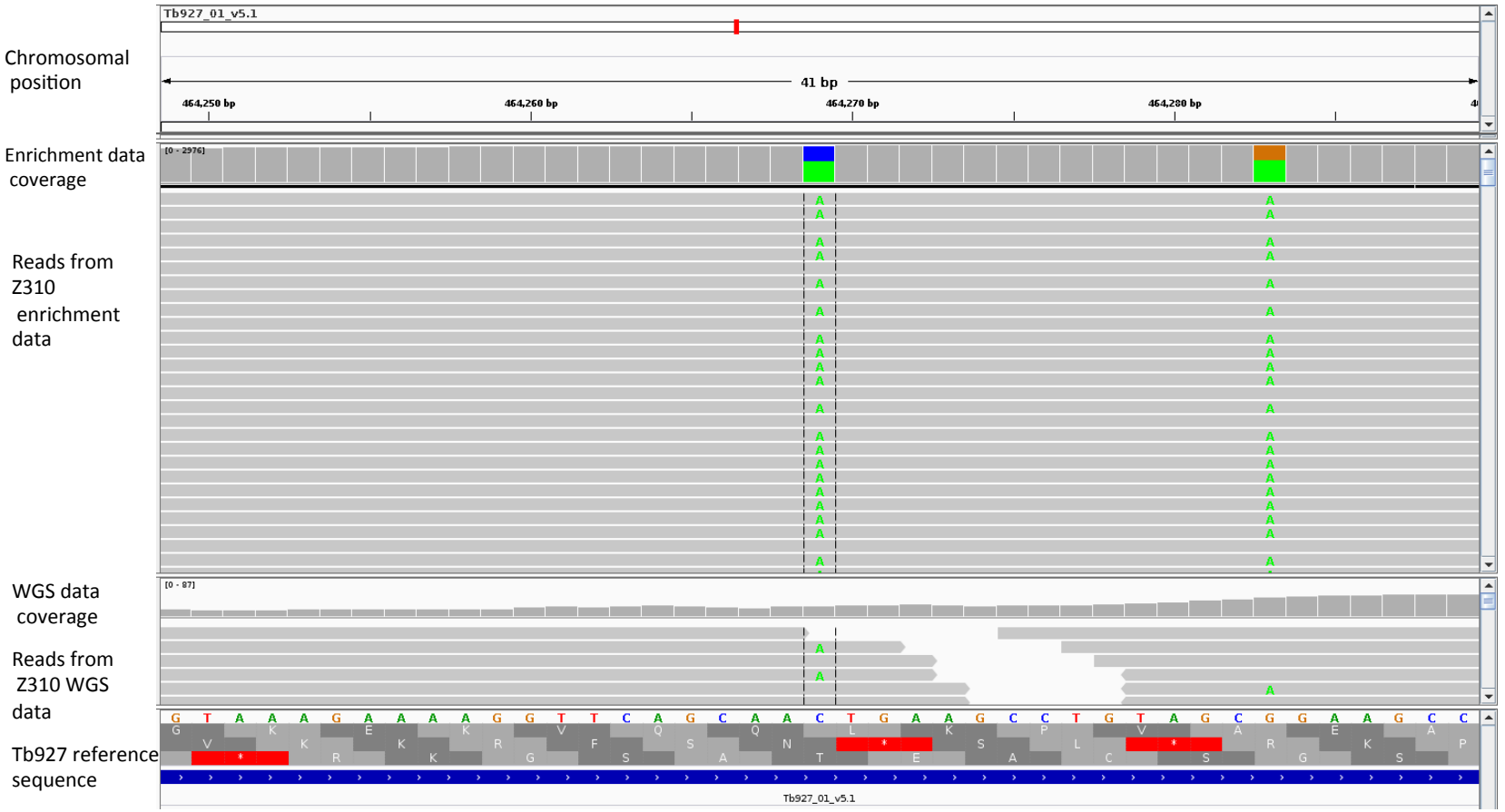

Figure 2.13: Image created in IGV. Two SNPs are shown in the reads aligned from the enrichment data, as indicated. The second track down, which shows the level of coverage in the enrichment data, shows two colours representing the different alleles in this heterozygous SNP. The ratio of the colours shown represents the relative ratios of the alleles called against this position. If we look at the aligned reads from the WGS data, this alternative allele is seen in a couple of reads, however insufficiently to be called a heterozygous SNP.

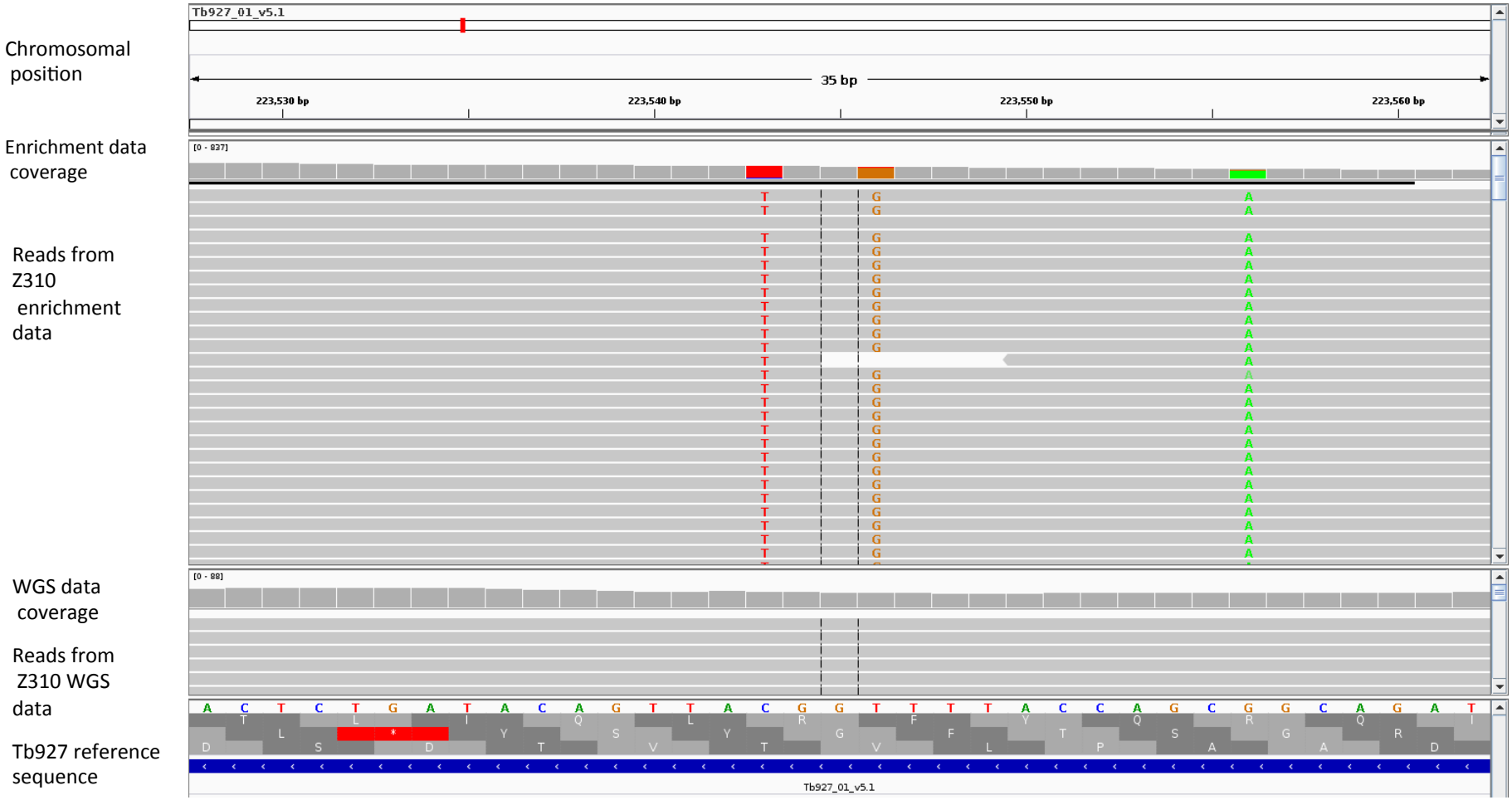

Figure 2.14: Image created in IGV. This shows three homozygous SNPs found in the enrichment data but not found in the WGS data, despite being homozygous variants.

2.4 Conclusion

The analysis of the data in this chapter has highlighted several things. One of these is that it is possible to generate high quality sequence data from minute quantities of starting DNA, and this means that this method of library preparation would be amenable to clinical or field samples, where sample quantity may be a hindrance to collecting enough DNA to prepare the sample in a traditional library preparation manner. It has also shown that MDA and Whatman FTA™ cards can be used in conjunction with target enrichment without adversely affecting the quality of the data.

The comparisons between the designs show that design is very important in determining the success of the enrichment. Good design can improve the percent of on target DNA sequenced, reduce variability in terms of coverage between targets in a design, and is reproducible between samples and designs. Data from the enrichment sequencing also does not appear to highly over or underrepresent either certain varieties of SNPs nor heterozygous/homozygous SNPs, compared to the WGS. Despite only ~80% of the SNPs found in the WGS data being found in the enrichment data, more total SNPs are called within the target region unique to the enrichment data and conserved between designs and strains, suggesting an increased sensitivity for SNP detection, as we would expect with newer sequencing chemistry. Figures 2.13 and 2.14 demonstrated that these SNPs were indeed valid and not an artefact of poor coverage and/or allelic dropout.

Despite poor performance of one of the libraries in the second design, a high percentage of the SNPs were still identified despite low coverage. Considering less than 0.01% of the DNA in the original sample contained parasite DNA, and ~60% initially mapped to the parasite, this a an enrichment of ~6000x, which makes it affordable to sequence to a high depth mixed samples, without discarding a high proportion of data. Overall, the enrichment technology performs comparably to the WGS data, and would be suitable in place of WGS where WGS is not a viable option, particularly in mixed samples or in unculturable conditions.

CHAPTER 3

Using multiple strains sequenced using enrichment sequencing to define variants contributing to phenotypic changes

3.1 Introduction

3.1.1 Genetic diversity in a population

Genetic diversity in a population can be measured in many ways. One way of defining diversity using sequence data is to look at SNP patterns and their functional consequences, in order to determine variations causing phenotypic differences. In parasites, SNPs can be used as genetic markers for identifying potential resistance in strains and looking at loci potentially under positive selection. In extreme examples, this can lead to mutations conferring an advantage dominating in a population; this is known as a selective sweep. Often this results in a greatly reduced level of diversity surrounding these highly positively selected SNPs (Kim and Stephan, 2002; Fay and Wu, 2000). In *Plasmodium,* variants called following sequencing were used to determine SNPs associated with artemisinin (ART) resistance (Cheeseman et al., 2014). Similar studies have also used SNP variants to investigate diversity arising under selection in *Plasmodium* following widespread use of antimalarial agents (Wootton et al., 2002), and in *Leishmania* to look at mechanisms of drug resistance (Downing et al., 2011).

This chapter will primarily look at the genetic diversity amongst three zymodeme groups within Ugandan *T.b. rhodesiense* strains. Seven strains belonging to three zymodeme groups B17, Z366 and Z310 are compared across two enrichment target regions. Additional strains as discussed in the methods were also used to check the validity of the enrichment method discussed in Chapter 2, for use on very low parasitaemia human *T. brucei* infections. However primarily the analysis will be focused on the seven strains used in both sequence captures, particularly as the most metadata is readily available for these strains, WGS data is available for one

representative B17 and Z310 strain and the phenotypes of these infections have been reproduced in experimentally infected mice. The clinical manifestation of these strains has already been discussed in (Smith and Bailey, 2000).

This chapter aims to look at variation, in this case SNPs, looking at both SNP conservation across strains and regions of unique SNPs, in order to decipher the potential functional effect of this variation, and whether this can be correlated with a difference in phenotype.

3.1.2 Aims of the chapter

In this chapter, the data produced using the methods outlined in Chapter 2, will be used against multiple strains per zymodeme group in order to try and elucidate potential genetic variants causing the phenotypic differences seen in these sets of strains. First the localization of unique and conserved variants will be used to observe any SNP dense regions of the genome. Secondly these variants will be annotated to determine the functional effect of these mutations, in order to generate candidate genes, which through mutation, are involved in the generation of these phenotypes. The genomic location of low-high impact SNPs will also be investigated to see whether there is any clustering of any particular type of variant. Unique variants identified through this process will be assigned GO terms and undergo GO term enrichment analysis in order to decipher whether particular pathways correlate with an abundance of deleterious SNPs, as these may be under selection. The proportion of non-proliferative stages has been postulated to determine infection outcomes, and so the proportion of different bloodstream stage forms has also been investigated in strains B and E.

3.1.3 Defining virulence in *T. brucei*

Defining virulence in trypanosomes is not straightforward (Morrison, 2011). Typically the three subspecies are described as having distinct phenotypes, with *T.b. rhodesiense* considered to give rise to the most acute infections, and *T.b. gambiense* typically causing more chronic infections. However within these subspecies there is great inter-strain variation, as demonstrated by the difference in strains within the same subspecies but different zymodeme groups. Classification by zymodeme group in this

instance is effective at grouping strains with similar infection profiles. Classification by iso-enzyme banding patterns (IBP), was previously used extensively in other parasites including *E. histolytica, T. cruzi* and *Giardia* (Sargeaunt and Williams, 1979; Mebrahtu et al., 1992; Bertram et al., 1983). However significant intra-zymodeme variation has also been observed in *Leishmania* and *T. cruzi* (Baptista-Fernandes et al., 2007; Mendonça et al., 2002).

The virulence of a parasite can be defined in multiple ways. This can be defined by characteristics in the clinical manifestation, as has been done with these *T.b. rhodesiense* strains, in which presence or absence of chancre is one of the defining characteristics of a virulent infection. Differences in clinical manifestation can also include variations in the length of the prepatent period, the parasitaemia at the first peak of parasitaemia, and the progression from early to late stage of the disease, as signified by parasites present in the cerebrospinal fluid (CSF) (Smith and Bailey, 2000; Morrison, 2011). In these terms, a parasite would be classed as highly virulent if its prepatent period was short and its first peak of parasitaemia was high (Morrison, 2011).

However the virulence of a parasite can also be defined by it's ability to infect a host and be transmitted to its vector. In *T. brucei*, only short stumpy forms are capable of being transmitted to the vector (Matthews et al., 2004). These transmissible parasites can also be divided into two discrete populations, the older short stumpy forms, which are leading towards apoptosis, and the younger short stumpy forms, which are still infective to the vector (Reuner et al., 1997; Seed and Wenck, 2003). As discussed later, some of the strains described in this chapter have different abundances of various life cycle stages, including the transmissible stumpy stage. Variance in this abundance can alter the capability of the parasite to be uptaken by the vector (MacGregor et al., 2011). By defining virulence by transmissibility, strains with the optimal number of transmissible forms are the most virulent.

Proliferative life stages often differentiate into the stumpy stage as a quorum sensing measure, but also in altruism, to allow the infection to be maintained within the host for a longer duration, and optimize younger short stumpy stages to be transmitted (Reuner et al., 1997; Seed and Wenck, 2003). The idea of this self-sacrifice to maintain a longer sustained infection in the host is reviewed in more depth by Duzsenko *et al*, 2006 (Duszenko et al., 2006). It is the immune reaction to the apoptotic events in the

older short stumpy population, which causes the most immunogenic response, and so a high population of these non-proliferative forms could cause the greatest immune response (Seed and Wenck, 2003). Once strains lose their ability to differentiate into these short stumpy stages and become monomorphic, this alters their virulence, and this is one of the reasons that stumpy abundance throughout an infection has been postulated to be one of the main determining factors for a strains virulence (MacGregor et al., 2011).

Host tolerance to the parasite is also considered a feature of virulence. The strains used within this chapter are initially used to infect a trypanosusceptible mouse strain, A/J, and are subsequently passaged into a more trypanotolerant mouse strain, C57BL/6. Primarily studies on host resistance in trypanosomes have been focused on the effect in cattle, with studies on trypanotolerant breeds N'Dama and the trypanosuspectible Boran breeds (Naessens, 2006; Orenge et al., 2012). However the interplay between host resistance and trypanosome infection in strains with differing levels of susceptibility, has also been explored (Morrison et al., 2010).

Using the aforementioned methods for the determination of virulence, the strains discussed in this chapter can be divided into three distinct phenotypic groups. Strains from zymodeme group Z310 typically present with a chronic infection. Patients infected with these strains are either asymptomatic, or present with a low grade infection after a long prepatent period. Initial peaks of parasitaemia are higher than the B17 zymodeme group strains, however these then lapse into very low parasitaemia chronic infections. In contrast, patients infected with strains from the B17 zymodeme group often present with a late stage infection, as confirmed by parasites within the cerebrospinal fluid. Infections with this parasite have a short prepatent period, and infection with these strains results in a more severe clinical manifestation, with a chancre almost consistently present, and a quick transition for the haemolymphatic stage of the disease to the meningoencephalitic stage of the disease. Z366 strains have an intermediate phenotype between these two zymodeme groups.

3.1.4 Differentiation in *T. brucei*

The parasites used in this chapter are bloodstream forms, generated through experimental infection in mice. Due to this, only the mammalian stages in Figure 3.1, are considered. In natural infections, in order to complete its life cycle, the parasite

must complete both stages in its vector host, the tsetse fly, and in it's mammalian host, which can either be human or more commonly in *T.b. rhodesiense* infections, cattle (Njiru et al., 2004). Information on the entire life cycle is provided within Chapter 1.

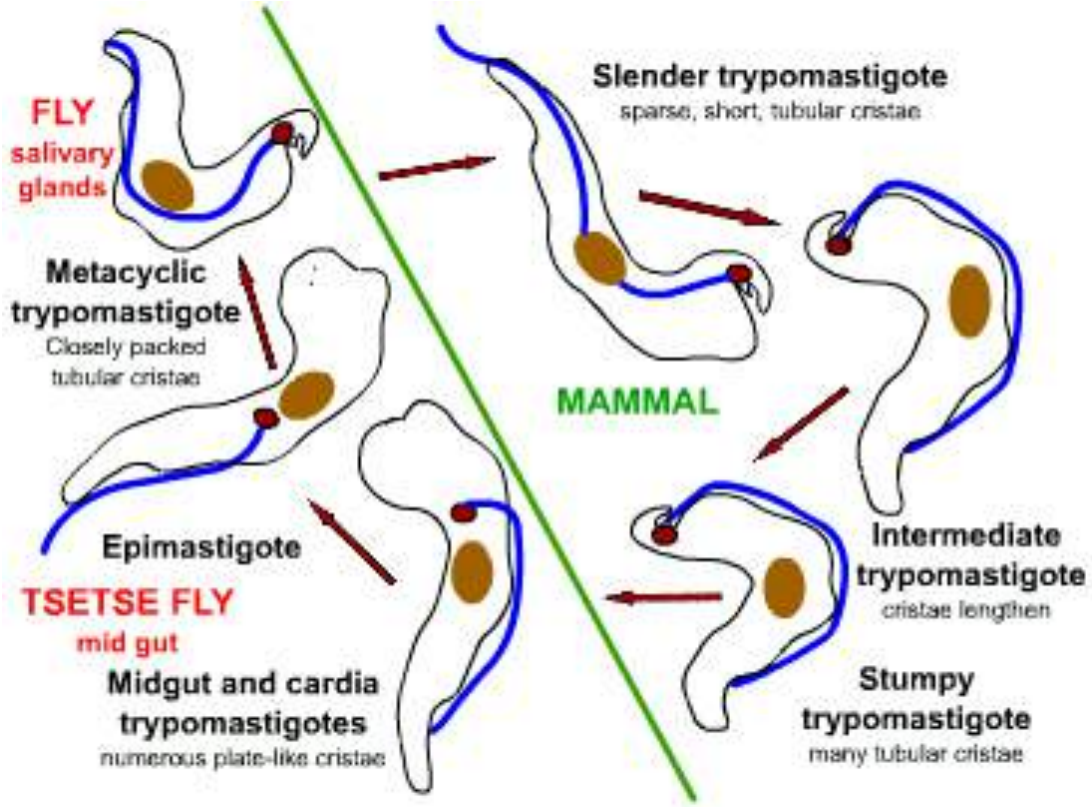

Figure 3.1: Shown is the lifecycle of T. brucei. This is the complete life cycle showing the major stages of the life cycle, with mammalian stages on the right hand side, vector stages on the left. The two stages important to this chapter are the stumpy trypomastigote, which shall be referred to as the short stumpy stage, which is required for uptake by the vector, and the slender trypomastigote, which will be referred to as the long slender stage. This image was taken with permission from www.microbiologyonline.org

When bitten by an infected tsetse fly, metacyclic trypomastigotes present in the tsetse's salivary glands are injected into the mammalian host. These trypomastigotes then differentiate into bloodstream forms. The initial bloodstream forms are the highly proliferative slender forms, however as the burden on the host increases, and the parasitaemia increases, these stages undergo extensive structural reorganization and become the short stumpy forms shown in Figure 3.1 (Matthews, 1999). These are incapable of proliferation and are adapted for transmission into the vector, as previously mentioned (Rico et al., 2013). This slender to stumpy transition varies between strains, and in this chapter will be used in association with other phenotypic

traits to identify whether the relative abundance of life cycle stages present in the infection correlates with the virulence observed.

Included within this chapter is microscopy data, which is used to calculate the relative abundance of different bloodstream form stages. Figure 3.2 is provided as a reference for how these stages are identified. Within Figure 3.2 are two parasites, the parasite to the left is a long slender parasite, and to the right is the differentiated form, the short stumpy stage parasite. There are multiple physical differences between the two, however as expected, in an infection, these parasites fall within a spectrum, with a variety of intermediates also present.

3.1.5 Long slender (LS) to short stumpy (SS) transition

As aforementioned, the ratio of LS to SS forms is a key determinant in the progression of an infection. Both stages elicit different immunological responses, with apoptosis in the older SS forms partly responsible for the host response (Rico et al., 2013). They also show reduced antigen switching, and so the host is more capable of mounting a host response (Barry and McCulloch, 2001). During the transition from slender to stumpy form, the parasite undergoes four distinct stages, the initial slender stage, in which the mitochondrial activity is repressed, followed by an intermediate stage, followed by an early stumpy stage, then a mature stumpy stage, in which the mitochondria activity is upregulated and ready for transmission to vector (Tyler et al., 1997).

In the stumpy form of the parasite, the parasite has a shortened rotund appearance, with a short flagellum and a more posterior kinetoplast position compared to what is observed in the long slender form. Mitochondrial biogenesis begins, with the slender mitochondria becomes enlarged and restructured in the stumpy form (Reuner et al., 1997; Tyler et al., 1997). In the slender form, the flagellum is long, the body of the parasite is considerably thinner compared to the stumpy stage, and the kinetoplast has a more anterior position (Fenn and Matthews, 2007).

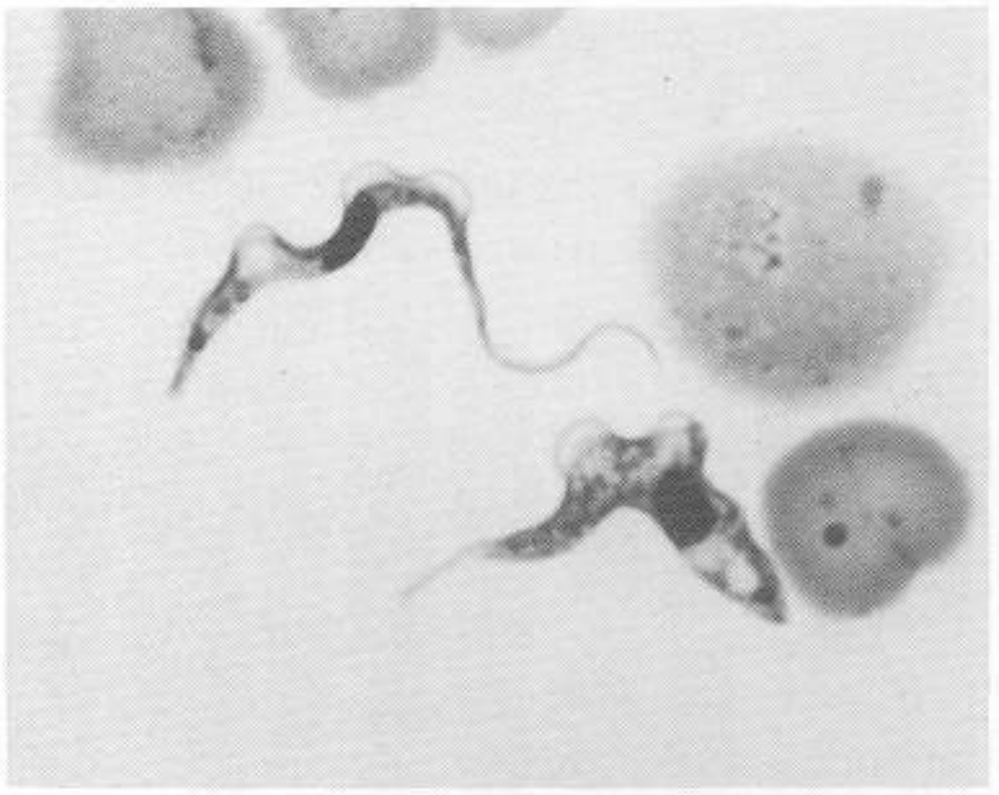

Figure 3.2: A giemsa strained micrograph showing the two major life stages in the bloodstream forms. The long slender is shown on the left, the short stumpy stage on the right. Image was taken with permission from the International Livestock Research Institute's website (ILRI).

3.1.6 Functional annotation

Following the advancement in short read mappers, as mentioned in Chapter 2, multiple tools have been developed for the downstream processing of variants within these large datasets. The three main variant annotators are ANNOVAR, SNPeff and ensembl's variant effect predictor (VEP) (Wang et al., 2010; McLaren et al., 2010; Cingolani et al., 2012b). Annotation by all three pieces of software is largely concordant, with McCarthy et al finding 85% of all annotations consistent between all software (McCarthy et al., 2014). However this consistency is largely restricted to the exonic regions, with concordance between annotations dropping to 44% in non-coding regions. These inconsistences are primarily due to how the software deals with loss of function effects, which are the most damaging. In part this is because ANNOVAR does not annotate stop/start lost or gained effects, which are considered in the SNPeff software to be the highest impact SNPs (Wang et al., 2010; Cingolani et al., 2012b).

In view of this, and due to the greater flexibility to annotate variants using custom user built databases for non-model organisms, SNPeff was used for the analysis in this chapter. SNPeff is compatible with GATK, which was used to generate these variants, however it also supports Samtools mpileup (Cingolani et al., 2012b). It calculates annotations by using an interval forest, which is a hash of interval trees that are indexed by chromosome (Cingolani et al., 2012b).

3.1.7 GO term annotation

GO terms are used to assign functions to genes and were constructed by the Gene Ontology Consortium to fit within three types, biological processes, molecular functions and cellular components. Biological process refers to the pathway or process the gene contributes to, molecular function describes the biochemical activity of the gene product and cellular components describes the location where the gene product is active (Ashburner et al., 2000). GO terms were assigned using tritrypdb's GO enrichment tool, as explained later in detail (Aslett et al., 2010).

3.1.8 REVIGO

REVIGO software uses a simple algorithm to reduce a set of GO terms to visualize the pathways enriched in a dataset. It clusters GO terms based on semantic similarity, and at a user set threshold, collapses GO terms into a broader GO on the basis of uniqueness (Supek et al., 2011). It assigns uniqueness using the simRel method in order to choose which terms are redundant and can be collapsed (Schlicker et al., 2006).

3.1.9 SNPRelate

SNPRelate is a R package which can be used to generate phylogenetic analysis from SNP data (Zheng et al., 2012) . It does using another R package, gdsfmt, which is used to generate a genomic data structure (GDS) file, instead of relying on multiple alignment methods of fasta sequences, such as those employed by software such as MUSCLE (Edgar, 2004; Zheng et al., 2012). It also "prunes" the dataset by only comparing variable sites, so conserved SNPs are removed.

3.2 Methods

3.2.1 Strain selection

All of the samples used in the initial design were blood samples taken from experimentally infected mice. These parasites are *T.b. rhodesiense* parasites isolated originally by Wendy Bailey in 1997 and the phenotypes described in Smith & Bailey, 1997 (Smith and Bailey, 2000). These strains originate from Uganda, the importance of this being that it is one of the few locations where *T.b. rhodesiense* and *T.b. gambiense* infections co-exist. This presents an unusual circumstance in that these usually geographically isolated have the potential to recombine.

The strains used in this study were picked on the basis of their phenotype in natural human infections. The strains picked were from three zymodeme groups, Z366, Z310 and B17, produce an intermediate, chronic and acute phenotype in human infections respectively (Smith and Bailey, 2000). In humans, B17 infections were commonly associated with a chancre, and patients presented with an acute early stage infection, whereas in Z310 infections, chancres were rarely present, patients had an asymptomatic early stage, and presented with late stage disease (Smith and Bailey, 2000). In murine infections, the Z310 infected individuals presented with higher parasitaemias than those seen in the B17 infections, and showed more severe symptoms. In the Z310 infections in mice, several individuals also had to be humanely culled prior to schedule. The presentation of these symptoms following infection with these strains has been observed previously (Goodhead et al., 2013).

The zymodeme group is determined by what iso-enzymes are present and was determined by MLEE electrophoresis as explained in Stevens & Tibayrenc (Stevens and Tibayrenc, 1995). Initially 2 strains, strain E from the B17 zymodeme group and strain B from the Z310 were used in a pilot study. Following this a total of 7 strains were then used, 3 of these were from the Z310 zymodeme, 3 from the B17 zymodeme, and one was intermediate Z366 phenotype. Further details of these samples are included in Table 3.1. These strains were chosen to look at similarities in inter and intra zymodeme groups, and attribute this to the phenotype.

These seven strains were used in the second design and an additional 9 strains were kindly provided by the University of Glasgow and included. These were primarily *T.b. gambiense* strains, and had a range of phenotypes. Unlike the other samples, these samples arrived as DNA extracted from whole blood.

Table 3.1: Shows the available metadata for each of the strains included within this chapter. The top seven were sampled from experimental mouse infections, the additional samples were provided by the University of Glasgow and were from natural human infections.

Source of sample	Region of origin	T. brucei subspecies	Zymodeme group	Strain	Phenotype in human infection	Sample preparation method	Used in first design
Experimental mouse infection	Uganda	*T. brucei.rhodesiense*	Z310	B	Chronic	Lysed PBMC applied to Whatman FTA™ classic card	Yes
				M			
				T			
			Z366	O	Intermediate		
			B17	K	Acute		
				E			
				N			
Clinical sample	Guinnea	*T. brucei.gambiense*	Unknown	G1	Unknown	DNA extracted from whole blood	No
				G2			
				G3			
				G4			
				G5	Symptomatic but no parasites in blood. Positive CAT and PCR for parasites		
	Cote d"Ivoire			G6	Unknown		
	Sierra Leone			G7	Very acute		
	Soroti, Uganda	*T. brucei.rhodesiense*		G9	Unknown		
	Tororo, Uganda			G10			

3.2.2 Sample collection

Sample collection from the mice used for enrichment sequencing were taken as described in Chapter 2. This outlines the sample collection for the QPCR, and microscopy data shown in this chapter. Samples were collected by initially infecting 2 female A/J mice intraperitoneally with 10^4 parasites from a blood stabilate of B17 and 2 female A/J mice with Z310. This mouse strain is particularly susceptible to trypanosome infection, which is important in order to obtain a high parasitaemia blood sample with which to infect subsequent mice. They were subsequently humanely sacrificed following positive results for parasites from microscopy screening.

Blood was then collected from these mice and 10^4 parasites were passaged into C57BL/6 mice, 5 for each isolate. Mice were bled prior to infection and 25µl of blood was then taken twice weekly for metabolomic analysis, 10µl of blood was also taken every other day for qPCR analysis, and daily spots were used for recording the presence of the trypanosomes, and used to create thin films for staining purposes. The mice used for metabolomic and QPCR analysis were culled prior to schedule due to ill health. Sample collection was initially intended for up till two weeks post infection, however two of the five mice infected with the Z310 strain were moribund by day nine post infection. The remaining mice showed signs of anemia, and so all individuals were culled.

3.2.3 Using microscopy to observe the relative abundance of bloodstream forms

Differences in the presentation of symptoms following from infection with these two strains are potentially related to a difference in cell cycle progression. This was tested *in vivo* using mice in order to obtain blood samples for screening using both microscopy and qPCR. This was done to determine the relevant abundance of both the slender and stumpy stages of the parasite.

3.2.4 Reverse field's stain of thin films

Thin films were made from approximately 5µl of blood collected by daily tail snip for parasite detection. Once dry, these were subsequently fixed in absolute methanol for 8 seconds, and stained using a reverse Field's Stain protocol as follows (Prolab

diagnostics™, UK). Slides were fixed in methanol using 8 dips at a rate of 1sec/dip, then stained with Field's stain B (Eosin), rinsed with water, stained with Field's stain A (methylene blue) before then rinsing in fresh water before being left to air dry upright. Each stage involved 8 dips at the same speed of 1 dip/sec.

Once dry, the films were then used to determine the number of parasites in either the stumpy, slender or intermediate form. For each thin film this ratio was determined by counting the parasites and ensuring, where possible, at least 20 parasites were counted. Due to the fluctuation of parasite numbers during infection, occasions when this was not possible are indicated in Figure 3.3. For highly parasitized slides 20 parasites or more were counted from multiple fields.

3.2.5 Using QPCR to validate the differential bloodstream form abundances observed in microscopy

QPCR was used to validate the microscopy data for frequency of slender and stumpy forms, (see Table 3.2), because although the microscopy data illustrates a very striking difference between strains, microscopy is limited because of its qualititative nature. Positive parasitaemia was confirmed by using a constitutive marker TbZFP, and the percentage of these parasites in the stumpy stage was determined by using a stage specific marker PAD1 (MacGregor et al., 2011).

3.2.6 Sample preparation for QPCR

10µl of blood was collected by tail snip into a microtube containing 2µl of 22.5mmol EDTA, which was centrifuged at 6000g to remove the plasma. Following the removal of the plasma, an equal volume of PBS (phosphate buffered saline) was added to the pelleted cells and resuspended. An equal volume of ABI purification lysis solution (ABI 4305895) was then added to the resuspended cells and stored at -80 degrees until use. Samples were taken from 5 B17 and 5 Z310 infected mice at days 6, 8 and 9-post infection, with day 6 representing the first peak of parasitaemia.

3.2.7 RNA extraction

RNA was extracted, using samples processed as above, using the protocol from the Purelink™ RNA Micro Kit, using the RNA extraction from suspended cells protocol (ABI 12183016). The only deviations from the protocol were that only 5µl of the sample was used and diluted into 35µl of PBS prior to the first step. This was due to the low sample volume obtained from tail snips, and samples were subject to 6000g in the initial centrifugation step. Samples were DNase treated on column using Purelink™'s on-column DNase treatment. Following extraction, the samples were resuspended in 20µl nuclease free H_2O. The quality was checked using A260/280 and A260/230 values from the Nanodrop™ and quantity using RNA Qubit™ values (Fisher Scientific, UK; Invitrogen, UK).

3.2.8 cDNA preparation and RT-PCR

Complementary DNA (cDNA) was produced with the ABI High Capacity RNA to cDNA kit (ABI 4387406) according to manufacturer's instructions and then amplified on an ABI RT-PCR machine. 5µl of total RNA was used in a 20µl reaction and reverse transcription was performed using once cycle of 37°C for 60 minutes and 95°C for 5minutes. The posttranscriptional regulator TbZFP3 was used as a constitutively expressed control, and a ΔΔCT method was used to compare the expression of the constitutively expressed posttranscriptional regulator TbZFP3 against the stumpy marker PAD1 (Paterou et al., 2006). Primers used for the amplification of PAD1 cDNA were 5'-GACCAAAGGAACCTTCTTCCT-3' and 5'-CACTGGCTCCCCTAAGCT-3'. For TbZFP3, the primers 5'- CAGGGGAAACGCAAAACTAA-3' and 5'-TGTCACCCCAAC TGCATTCT-3' were used (MacGregor et al., 2011).

Power SYBR™™ green PCR master mix was used in 25µl reactions (ABI 4367659). Typical reaction mixes were as follows. For PAD1 QPCR reactions 12.5µl of SYBR™ green PCR mix, 0.75ul of 10µM for each primer, 4µl of dH_2O and 7µl of cDNA (diluted 4µl of cDNA prepared as above, in 45µl of H_2O). For ZFP3 reactions, 12.5µl of SYBR™ green PCR mix was used, 2.25µl of 10µM of each primer, 1µl of dH_2O and 7µl of cDNA prepared as with the PAD1 reactions. Cycle conditions as follows: initial denaturation of 95°C for 10minutes, followed by 40 cycles of 95°C for 15 seconds then 60°C for a minute, then a melt curve. Reactions and NTCs (non template controls) were set up in

triplicate. Melt curve analysis and a 1.5% agarose gel were used with the PCR product to verify that there was only one amplification product.

3.2.9 Library preparation

The libraries analyzed in this chapter were prepared as described in Chapter 2. Samples donated by the University of Glasgow already had the DNA extracted. However these still required WGA. This process and the QC involved are described in Chapter 2.

3.2.10 Bioinformatic analysis

3.2.10.1 Read alignment to Tb927 reference

The enrichment data was mapped to the Tb927 version 8.1 reference, available at www.Tritrypdb.org, indexed with bwa is and mapped with algorithm aln with default settings. The data was mapped as paired end using default settings, see Chapter 2 for more information. The SOLiD data, which is used for comparison purposes in parts, was mapped as described in Chapter 2. Enrichment data was also mapped to the *T.b. gambiense* reference DAL972 v8.1 using bwa's default settings as above.

3.2.10.2 Variant calling

Variant calling was done in GATK version 3.0. This was done as described in Chapter 2, however in addition variants were filtered to make sure they fell within the target region using GATK's select variants walker (DePristo et al., 2011). The zygosity of these SNPs was determined as mentioned in Chapter 2, and in conjunction with vcf-isec, was used to look at the zygosity of conserved and unique SNPs across different strain combinations (Danecek et al., 2011).

3.2.10.3 Use of VCFtools to generate SNP intersections

Variants within the target region were bgzipped, tabix indexed and used with vcf-isec to generate intersections between different strains. For SNPs present in x number of files, vcf-isec -f -n = x was used. Unique gene sets were derived using vcf-isec -f -c –a, to find positions unique to the first file given.

Seven of the strains were used in both designs. In order to generate the largest SNP dataset for these strains, the SNPs were compiled from those that were conserved between designs, and those that were unique to one design. This increases the total number of SNPs because some of the genes in design one were reannotated, and the second design probe set was extended into these newly annotated regions. In comparisons to the G7 strain, SNPs from both designs were used, but only those within the second design target region were used for comparison. This was because G7 was not used in the first design. Vcf-compare was also used for across strain SNP comparisons. Picard tool's sortvcf tool was also used to enable parsing between different software (http://picard.sourceforge.net.). Only SNPs with a coverage depth of at least 5 were included.

3.2.10.4 SNP functional annotation

SNP datasets were analyzed using SNPeff/SNPsift software to annotate target regions. A custom database was created using version 8.1 of the Tb927 reference, and its associated GFF3 file, both are accessible at www.tritrypdb.org. This was converted to GTF format using the software maker and used with the fasta to annotate the variants (Cantarel et al., 2008). Snpeff generates multiple predictions per gene, which reflects the many potential impacts the SNP can have, dependent on the reading frame or splicing. A variant can affect more than one gene if it is in a regulatory region, and genes can also harbor multiple transcripts (Cingolani et al., 2012b). In SNPeff, a canonical transcript is determined to be the longest protein-coding transcript. In cases of not protein coding genes, this is the longest cDNA (Cingolani et al., 2012b).

These were assigned to four impact groups, low, moderate, modifier and high depending on the predicted SNP effect. They were also annotated for the particular SNP type for instance stop gained, or stop lost, the range of potential SNP effects within this data set is shown in the results. These effects were also grouped into synonymous, missense and nonsense effects. Snpsift was used to calculate the abundance of different SNP types/effects using the vcfEffOnePerLine.pl script and the Snpsift extract fields (Cingolani et al., 2012a).

3.2.10.5 GO term analysis

Genes containing SNPs that were unique to one of these zymodeme groups (B,E or O) were separated into impact groups (low,moderate,modifier and high), and were used with Tritrypdb's GO term analysis package to assign GO terms to the genes with these SNPs (Aslett et al., 2010). This was an attempt to understand if certain pathways were more enriched with different sets of SNPs. For instance if the more virulent strains from the B17 group had multiple genes with deleterious effects in a particular pathway, it would suggest these pathways are under selective pressure, and may correlate with the altered phenotype.

There are three types of ontology analysis, which assigns GO terms based on either the cellular component they are associated with, the molecular function the gene has, or the biological process/pathway the gene is involved in. For these, the genes were assigned to their biological process GO term. GO terms were sources from InterPro's database and annotations from the Tritrypdb and Genedb's databases. Only GO terms with a P value of less than 0.05 were assigned.

For each set of genes GO terms were associated, and the P-value, Bonferroni adjusted p-value and the Benjamini-Hochberg scores were calculated. The fold of enrichment was calculated by counting the number of genes with this term within the dataset, and the total number within this pathway i.e. the number of genes within this background, to give the percentage of pathway included within the dataset. The percentage of the total number of genes assigned GO terms within this pathway was then divided by the percentage of this background of genes to give the fold enrichment.

3.2.10.6 Fold enrichment

The fold enrichment per GO term was visualized per strain for unique SNPs by each impact group using REVIGO software (Supek et al., 2011). The axes shown in the REVIGO plots are irrelevant; it is the distance between these GO terms which shows the relationship. The degree of uniqueness per GO term is determined by 1-(average semantic similarity) over a list of GO terms. GO terms are collapsed when they aren't considered to be unique. Any terms with a greater than 0.7 degree similarity were collapsed into broader GO terms, and uniprot's database was used to determine uniqueness between GO terms. Closely related pathways cluster and overlap each

other, a large distance between points demonstrates a great degree of uniqueness between GO terms. Plotted GO terms are coloured according to their log10 P value. P values were taken from the bonferroni adjusted values because multiple comparisons were done. Due to a cut off of 0.05, all plotted terms are significant, however those coloured blue have the greatest significance, and those plotted towards the red end of the spectrum have the least significance.

The GO terms are given a dispensability score, this uses the adjusted P-value and semantic similarity between GO terms to determine how indispensible a GO term is, and whether this can be collapsed into a larger GO category (Supek et al., 2011). GO slims could also have been used as alternatives, however the broadness of these categories very often reduces the power of the data by obscuring biologically interesting data points (Supek et al., 2011).

3.2.10.7 Generating a dendrogram from SNPRelate

A dendrogram was generated using the snpgdsVCF2GDS function to generate a GDS file from a VCF file containing SNPs from all seven of the strains used in both designs. This multi-sample VCF was generated from SNPs called in GATK, and subsequently merged using VCFtools merge function. Non bi-allelic SNPs and conserved SNPs between samples were removed. The resulting file was then transformed into a matrix using the snpgdsDiss function of the package, and subsequently snpHCluster, to generate z-scores. snpgdsCutTree and snpgdsDrawTree functions were used to generate a dendrogram based on z-scores, using default settings.

3.3 Result and discussion

It was suspected that the reason for the differences in virulence between Z310 and B17 was due to a difference in their progression through the life cycle. Whole blood samples from B17 and Z310 mice infected as aforementioned did exhibit the symptoms previously associated with these strains. Striking characteristics include a difference in the level of parasitaemia, the prepatent period, and the symptoms exhibited by the infected host. Both strains were passaged at the first peak of parasitaemia. As

previously demonstrated, Z310 enters the first peak of parasitaemia prior to B17. The parasitaemia at this first peak is also much greater than that of B17.

3.3.1 Microscopy demonstrates key differences in the relative abundances of bloodstream forms in B17 and Z310 infections

Microscopy was used to ascertain the ratio of stumpy to slender forms in order to observe if there was a significant difference between the two isolates. Figure 3.4 demonstrates that there is a significant difference in terms of the relative abundance of the two stages both during the infection and between the two isolates, which may account for the difference in virulence. Throughout both infections, it is evident that this ratio is maintained, with Z310 parasites primarily being in the slender form for the duration of the infection and B17 primarily stumpy.

Figure 3.3 shows the mean parasitaemia through days three to nine post infection, based on count data from daily tail snips. Z310 infections have a short prepatent period, with the first peak in parasitaemia observed at day four, and a high parasitaemia from day three. In contrast, B17 infections do not peak until day seven post infection, and the parasitaemia is both lower than the Z310 peak, and falls quicker post peak compared to the Z310 infections. Both strains have the highest degree of variability between individuals at day six, just as the B17 infection parasitaemias are entering into their first peak and the parasitaemia in Z310 infections are lowering following their first peak. Low parasitaemia periods during the infection are represented by troughs in Figure 3.4.

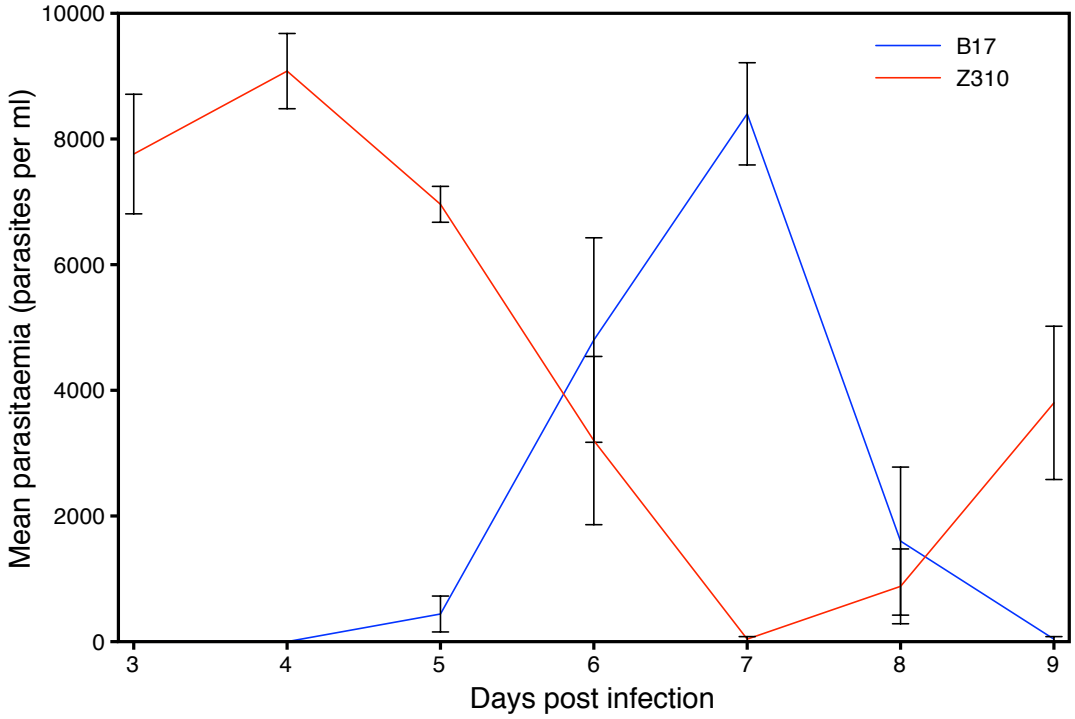

Figure 3.3: This shows the mean parasitaemia (parasites per ml) and is derived from counts from thin films of blood collected by daily tail snip. The mean parasitaemia is taken from five individuals for each strain. Standard error bars show the degree of variation in the parasitaemia between individuals, which is greatest at day six post infection. Mean parasitaemia is shown for Z310 infections in red and B17 infections in blue. The first peak of parasitaemia for B17 infections is seen at day seven, and for Z310 infections at day three. The peak of parasitaemia is higher in Z310 infections, and occurs after a much shorter prepatent period, and takes longer to lower the parasitaemia compared to B17 infections post the first peak of parasitaemia.

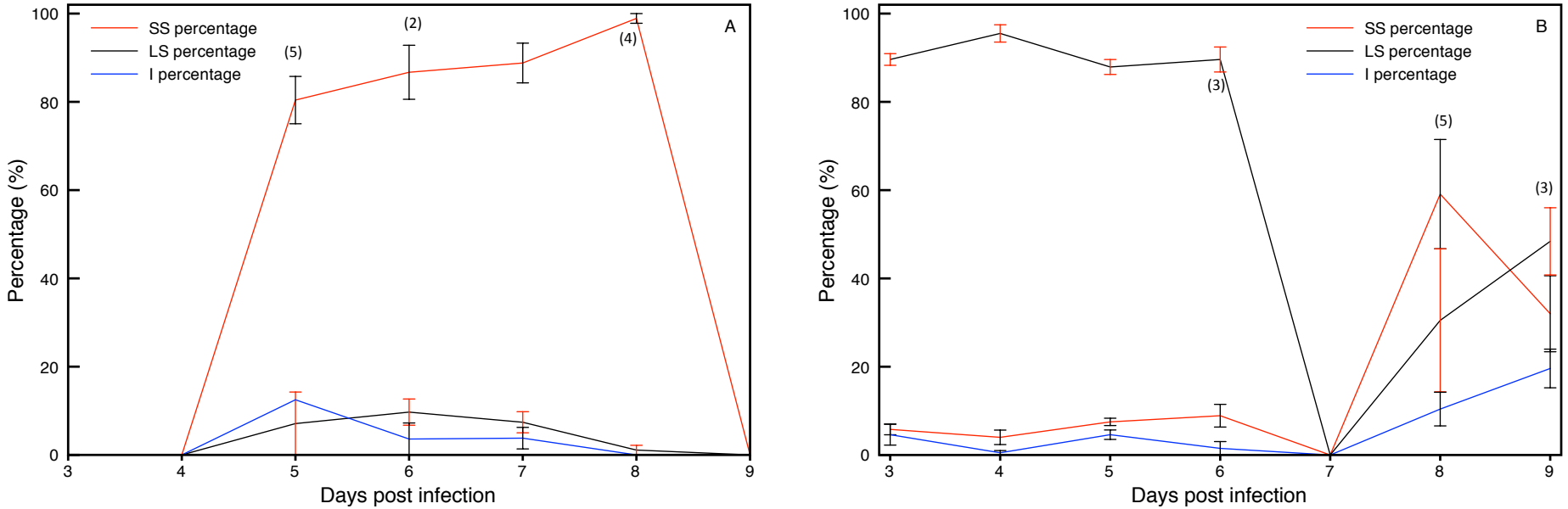

Figure 3.3: A and B illustrate changes in the relative abundance of short stumpy (SS), long slender (LS) and intermediate (I) stages during the course of infection. A shows this in B17 infected mice, B in Z310 infected mice. The percentage of parasites counted in each stage is shown in red, black and blue for SS, LS and I stages respectively. The standard error for each is shown in black for SS and I stages, and in red for the LS stages to discriminate between LS and I. In instances where fewer than 20 parasites could be counted in total, the number of mice this occurred in is shown in brackets. Throughout the infection the parasites in the Z310 strain are primarily slender forms, a stage, which can still proliferate unlike the non-dividing stumpy stage. This may account for the much higher parasitic load compared to the B17 strain, which primarily consists of stumpy stage parasites Following a great decline after the first peak of parasitaemia is it interesting to note that for the first time during infection the percentage of stumpy parasites is greater than that of slender forms, however as the parasitaemia begins to rise again the stumpy percentage begins to fall. B17 enters its first peak of parasitaemia on the 6th day post infection, whereas the Z310 infection is already in its peak by the day 3. Throughout the course of the B17 infection the stumpy forms predominate.

3.3.2 QPCR data correlates with microscopy data and shows B17 infections consist of predominantly short stumpy forms

Samples from the five infected mice for each strain were collected once the mice tested positive for parasites. The mice in this experiment had to be sacrificed earlier than anticipated due to the sudden decline in the Z310 infected mice, so only three time points, days 6, 8 and 9 post infection were available for QPCR analysis. Sacrifice prior to planned time point has been reported in previous experimental infections with this strain (Beament, 2002).

PAD1 has been previously described as a stumpy stage marker (MacGregor et al., 2011; MacGregor and Matthews, 2012), and given that the difference in phenotype in B17 and Z310 strains is suspected to be related to a difference in the ability of the parasites present to differentiate or the proportion of non-differentiating/non proliferative stages present, it is unsurprising that infections with the more acute isolate, Z310, contained a higher percentage of proliferative stage parasites, compared to B17. These observations were also confirmed by the QPCR data (Table 3.2). B17 infections contained fewer highly proliferative stages, and caused a chronic infection in these mice.

Trypanosomes differentiate into the stumpy form in response to population density (Reuner et al., 1997). Table 3.2 shows the level of expression of PAD1 relative to the housekeeping post-transcriptional regulator, TbZFP3. B17 infected patients commonly present with severe late stage disease, in our infections Z310 enters its first peak of parasitaemia earlier than the more chronic strain B17, and the parasitic load is greater than that of B17 and this is maintained for longer compared to the B17 infections.

Due to the greater number of parasites, and the propensity of trypanosomes to differentiate into the stumpy form in response to population density, the proportion of stumpy forms in the Z310 strain was expected to increase following the peak of parasitaemia, and as a result the expression of PAD1 to increase. In fact the opposite is observed, PAD1 levels in Z310 are maintained following the first peak of parasitaemia, and are then lowered. In contrast, the B17 infection contained a far greater number of stumpy forms, as is reflected in the relative levels of PAD1. This is seen particularly at day 6, with a 10^4 fold higher level of expression than Z310.

Table 3.2- This table shows the relative expression of PAD1 compared to a housekeeping gene TbZFP3 from CT scores. These are based on blood samples taken at days 6,8,9 post infection as indicated, and the values represent the fold increase as average over the five mice infected with the same strain. As suspected, B17 has a much higher level of expression compared to the Z310 strain, with a 10^4 difference in fold increase during the first peak of parasitaemia for Z310 (day 6). Following the peak of parasitaemia, particularly with such an acute isolate, we would expect an increase in the relative levels of PAD1 expression as the parasites differentiate to the stumpy form in response to increased population numbers, however levels remain the same and soon falter after the peak of parasitaemia in the Z310 isolate.

Mean	**Days post infection**		
	Day 6	Day 8	Day 9
Z310	$5.50e^{10}$	$9.64e^{10}$	$2.51e^{3}$
B17	$5.25e^{14}$	$5.61e^{7}$	$4.71e^{6}$

3.3.3 PAD1 regulation

The 3' UTR of PAD1, which is in the intergenic region between PAD1 and PAD2 is thought to control the expression of PAD1 (MacGregor and Matthews, 2012). Given that PAD1 is differentially expressed in B17 and Z310, this would suggest that perhaps that this may be the reason for a difference in regulation between these two strains. However sequence data for both of these two strains in the 3' UTR of PAD1 and the intergenic region between PAD1 and PAD2 contained no SNPs and had a 100% identity. This is unsurprising given the high percentage of similarity between *T. brucei rhodesiense* strains. This is consistent with the theory that although PAD1 is present on the surface of stumpy forms and not slender forms, it is not a factor important for the induction of differentiation into the stumpy form, but a product of differentiation (MacGregor and Matthews, 2012).

3.3.4 Mapping of enrichment data

Sixteen samples were used in the second enrichment design, seven of which were strains also used in the first design. These were mapped using BWA aln on its default settings against the Tb927 reference version 8.1. Where mentioned, the data was also mapped to the DAL972 reference version 8.1. The associated metadata is included in Table 3.1 for these strains. The strains used in both designs were enriched from DNA extracted from infected mouse blood applied to classic FTA™ cards. Samples used solely in the second design were kindly donated from the University of Glasgow, DNA was already extracted from these, and these were used directly in WGA reactions and

subsequently prepared as described for the other libraries in Chapter 2. These samples were from natural human infections instead of experimentally infected mice, and so the parasitaemia was very low compared to the parasitaemia in the infected mouse samples, excluding sample G7, which was a very acute *T.b. gambiense* infection.

Figure 3.4 shows the percentage of reads mapped, shown in red and purple, with the percentage of uniquely mapped reads within this shown only in red. Unmapped reads are shown in blue. Performance in terms of percentage mapping, was more variable across the first design, as shown in Chapter 2, however performance across the second design was more uniform, excluding strain E.

Only 1-2% of the reads in the human infections mapped to *T. brucei*, with the exception of the heavily parasitized sample, G7. This equates to between 120,000-880,000 reads. Although a high percentage of data would need to be discarded for subsequent analysis, the majority of these strains had concentrations of less than 12ngμl^{-1}, G9 and G10 had less than 1ngμl^{-1} prior to amplification. In murine infections, the parasite DNA only represented less than 0.1% of the sample prior to enrichment, in a parasitaemia of 10^6 parasites ml^{-1}. The pre-enrichment percentage in the human infections is much lower than this, due to lower parasitaemia and increased host genome size. Despite only a small percentage of reads mapping to Tb927, this still correlates with an approximate 100-fold enrichment. This shows that even very low parasitaemia samples are successfully enriched, however a higher depth of sequencing may be required to generate sufficient coverage for analysis. G7's original concentration was less than 10ngml^{-1}, showing the proportion of the target DNA takes precedence over starting DNA concentrations.

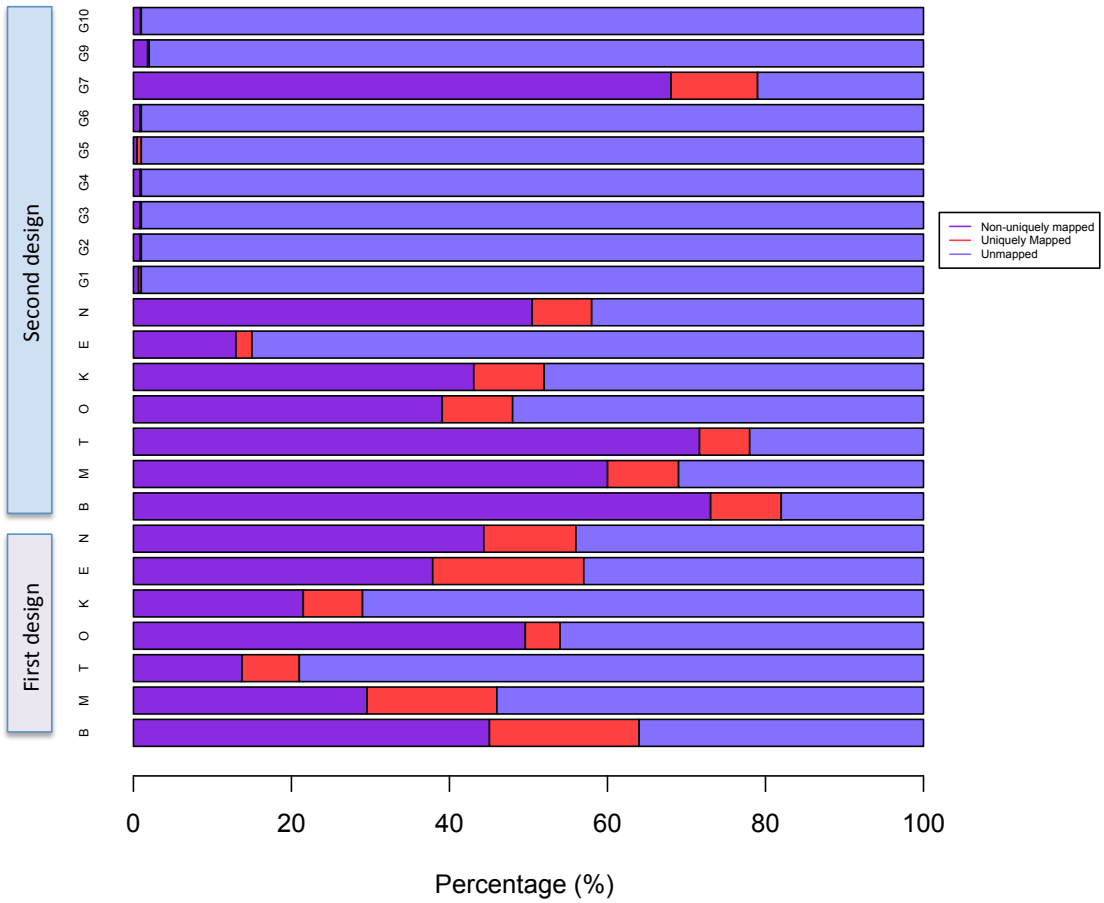

Figure 3.4: This shows the mapping percentages for the strains used in both enrichment designs. With the second design including all the of the strains used in the first design. The percentages of non-uniquely mapped are shown in purple and red, uniquely mapped in red, and unmapped in blue. The samples derived from experimental mouse infections (B,M,T,O,K,E,N), which were used in both designs, enriched well, but more uniformally in design two compared to design one, with the exception of strain E. Samples from human infections had much lower parasitaemias, and even though they were successfully enriched, the lower starting percentage means that only a small percentage of the sample maps to the Tb927 reference.

3.3.5 SNP analysis

SNP conservation between different groups of strains can be used to indicate the degree of overlap between strains. Only the strains used in both designs with an assigned zymodeme group were used in the subsequent analysis. Strain B, which is assigned to zymodeme group Z310, and strain E, which is assigned to zymodeme group B17, are compared to each other, and to strains within their zymodeme group. As mentioned in Chapter 2, only these strains have corresponding WGS data.

3.3.6 Inter-zymodeme variation

Figure 3.5 shows the degree of variation as determined by SNP diversity, between zymodeme groups. The strains shown in Figure 3.5A and B are B,E and O and represent zymodeme groups Z310,B17 and Z366 respectively.

When comparing just Z310 and B17 strains, as discussed in Chapter 2 and shown in Figure 3.5A, there is a high degree of similarity, with 15,982 SNPs conserved, which accounts for 84 and 92% of the total SNPs in Z310 and B17 respectively. However Z310 has 3,022 SNPs unique to this strain, compared with the 1,405 found unique to B17.

In a three-way comparison, a high percentage of SNPs were conserved in all three strains (62-75%), as shown in Figure 3.5B. Despite a relatively equal proportion of SNPs being conserved between all three strains, there is a higher degree of similarity between zymodeme groups Z310 and B17 than to Z366. This is reflected in the lower percentage of SNPs unique to each strain, which was ~6% in both Z310 and B17, compared to the 14.3% of unique SNPs in strain Z366. The overlap between Z310 and B17 in the three-way comparison was also higher, with 4,211 SNPs conserved just between Z310 and B17, in addition to the 11,771 SNPs conserved between all three strains. However there is a greater degree of overlap between Z310 and Z366 than Z366 and B17, with only 258 SNPs conserved between B17 and Z366, compared to the 1,721 conserved between Z310 and Z366 strains. Z366 and Z310 result in the manifestation of intermediate and acute murine infections, which indicates distinct mechanistic differences in the manifestation of the disease in B17 infections, which are chronic in murine infections.

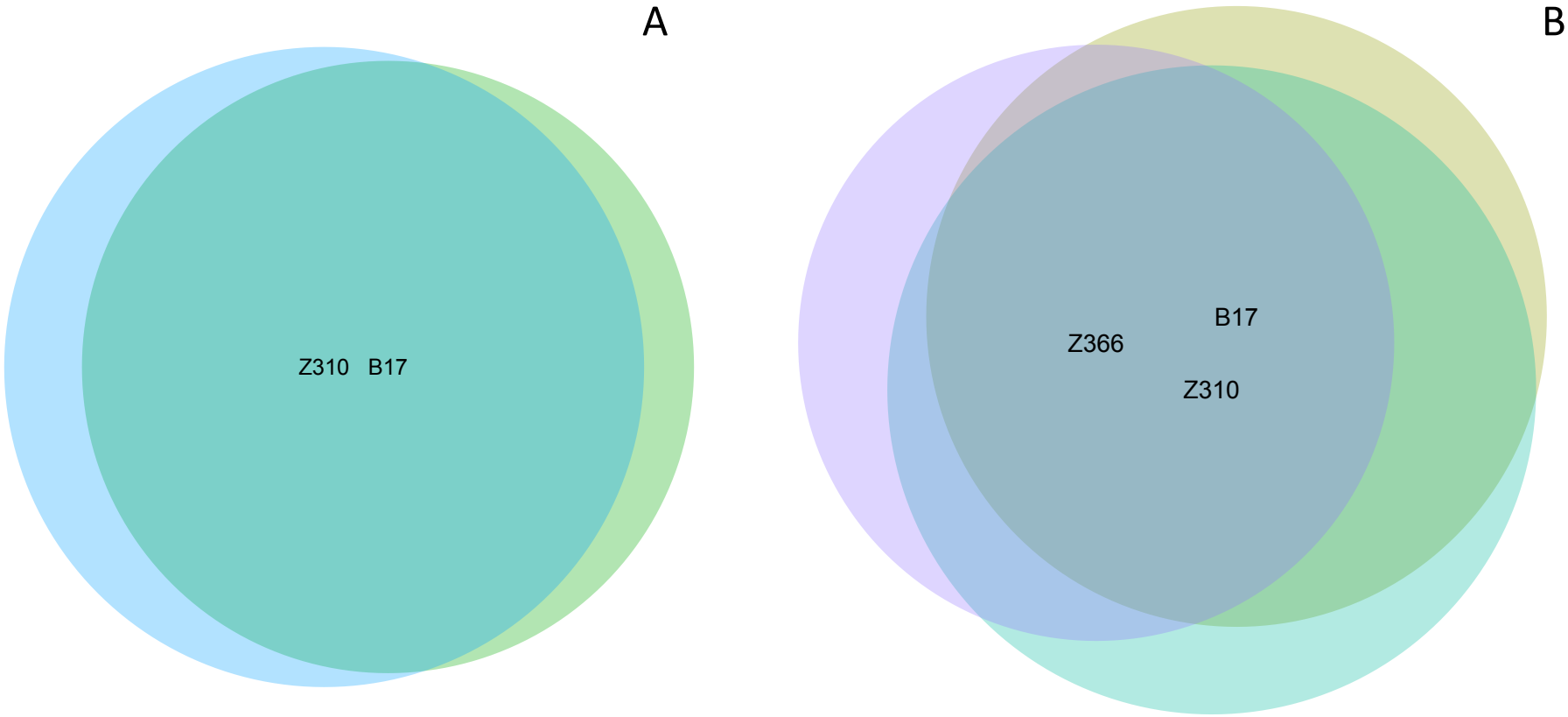

Figure 3.5: A and B are weighted Venn diagrams showing the overlap between zymodeme groups based on conserved SNP positions. In A unique SNP positions in Z310 are shown in blue, and green for the unique SNP positions in B17, and conserved SNP positions are shown in the overlap. These represent 1,405, 3,022, and 15,982 SNPs respectively. In B all three zymodeme groups, those with chronic, acute and moderate phenotype are shown. In B unique SNP positions in B17 are shown in yellow, purple for unique SNP positions in Z366 and pale green in Z310. Overlaps between these represent the conserved SNPs between either two or three of the zymodeme groups. 11,771 SNP positions were conserved between al three zymodeme groups, 1,721 were conserved between Z310 and Z366, 257 between B17 and Z366 and 4,211 between Z310 and B17. 1,147, 1301 and 2,289 were unique positions to zymodeme groups B17, Z310 and Z366 respectively.

3.3.7 Intra-zymodeme variation

Figure 3.6 shows the within zymodeme group variation in Z310 in Figure 3.6A and B17 in Figure 3.6B respectively. Figure 3.6A uses strains B,M and T for Z310 comparison and 6B uses K,E, and N for B17 comparison. A greater number of SNPs were conserved between all Z310 strains, with 16,545 SNPs conserved, compared to the 13,822 SNPs conserved between all B17 strains. Figure 3.6A illustrates a lower degree of variability between Z310 strains, with not only a higher number of SNPs common to all strains, but a low percentage of uniqueness per strain (3-7%). The strain with the highest degree of uniqueness is strain T with 1,257 unique SNPs (6.7%), and less overlap with strain B, with 399 SNPs shared, and strain M, with 457 SNPs shared. In comparison, strains B and M share 1,291 SNPs and have only 769 and 643 SNPs unique respectively.

There is a higher degree of variability within the B17 zymodeme group, and this is observed in Figure 3.6B. The degree of uniqueness per strain is higher than observed in Z310 (5-12%). As seen in Z310, one strain has far fewer conserved SNPs; in B17 this is strain N, with 2,342 unique SNPs, compared to the 818 in strain K and 1,337 in strain E. Positions conserved between all strains of each zymodeme group account for only 70-82% of each strains total SNPs in B17, compared to Z310, where this accounts for 87-88%.

The SNPs conserved between all three strains within a zymodeme group could be considered a core genome for that particular group. SNPs additional to this 'core' set could potentially have very little impact on the virulence of the strain, as there is a high degree of similarity between all three zymodeme groups, and despite marked differences in disease manifestation between zymodeme groups, there aren't between strains belonging to the same zymodeme group. This suggests the genetic differences responsible for these phenotypic differences seen are those conserved between strains within a zymodeme group.

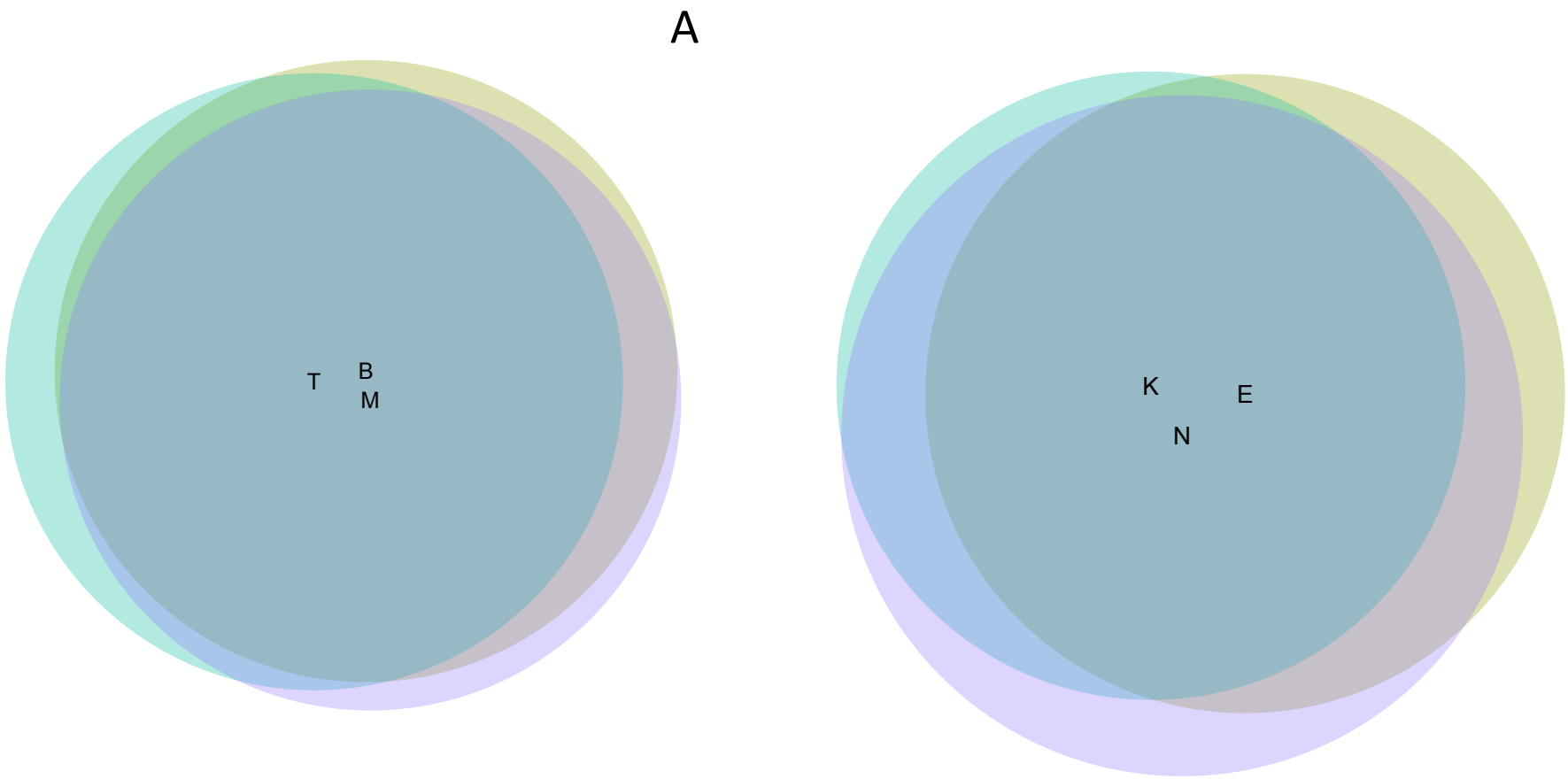

Figure 3.6: A and B are weighted Venn diagrams showing within-zymodeme group variation based on conserved SNP positions. A shows conservation within the Z310 zymodeme group in strains B,M and T. B shows conservation between strains K,E, and N in the B17 zymodeme group. The unique SNP positions are shown in green, purple and yellow for strains T,M and B in A and K,N and E in B respectively. Conserved positions are shown in the overlapped regions. In A, 769, 1,257 and 643 SNPs were unique to strains B,T and M. 457 SNPs were conserved between M and T, 399 between B and T and 1,291 between B and M. 16,545 SNPs were common to all Z310 strains. In B, 818, 1,337 and 2,342 SNPs were unique to strains K,E and N. 410 SNPs were conserved between K and E, 1,755 between K and N, and 1,818 between E and N. 13,822 SNPs SNPs were common to all B17 strains.

3.3.8 Inter-zymodeme variation with multiple strains

The variability between zymodeme groups when looking at multiple strains per group is shown in Figure 3.7. The strains used for comparison in Z310 and B17 groups were identical to those used in the intra-zymodeme comparison in Figure 6. Zymodeme group Z366 is represented by only one strain, strain O. Only SNPs conserved within the zymodeme group were used for the comparison. As was seen in Figure 3.4, there is a higher degree of similarity between zymodeme groups Z310 and B17. As expected, because only one strain represents zymodeme group Z366, there is a higher degree of uniqueness (17.6%) compared to the number of unique SNPs in groups Z310 and B17. Given the variability seen in Figure 3.5B, where only strains B,E, and O were compared, this degree of uniqueness in strain O is lower than expected.

11,182 SNPs were conserved between all 3 zymodeme groups, and this represented 70, 68 and 81% in groups Z366, Z310 and B17 respectively. A large proportion of the remaining SNPs were overlapped between B17 and Z310, with 1,862 SNPs present in both of these, and 1,784 SNPs present in both Z310 and Z366. Only 247 SNPs were found in both Z366 and B17. Despite to the comparison being derived from SNPs conserved within each zymodeme group, the number of SNPs present in all three groups is not significantly lower than seen in the 3-strain comparison in Figure 3.5, in which 11,771 SNPs were conserved.

In Figure 3.7B, strain G7 was compared in addition to the three zymodeme groups. G7 was from an acute *T.b. gambiense* infection, and so we would expect the majority of these SNP differences to arise because all of the other strains belong to separate sub-species. G7 was only used in the second design, and so the SNPs used for comparison were only those found within the second design for all strains. A high degree of similarity is known amongst *T. brucei's* three subspecies, and so it is unsurprising that 62% of G7's SNPs were identical to those found in all the *T.b. rhodesiense* strains compared. In total there were 9,226 SNPs conserved between G7 and the 7 strains represented in these three zymodeme groups. G7 did have a higher number of SNPs unique compared to the other strains as expected, with 2,638 SNPs representing 17.6% of the total SNPs for that strain. However this is comparable to the number of SNPs unique to strain O in the cross zymodeme comparison in Figure 3.5A, and so this

demonstrates a level of variation across subspecies similar to that seen between zymodeme groups.

In this comparison, the number of SNPs unique to Z366 is actually 1,607 SNPs and represents 10.4% of the total SNPs called for that strain. The number of SNPs overlapping between Z366 and G7 is substantially higher than seen between G7 and either Z310 or B17, with 1,051, 229 and 178 SNPs conserved between them respectively. This suggests as seen in Figure 3.7B, that B17 and Z310 are more closely related, and G7 is more closely related to Z366 than to Z310 and B17.

Figure 3.8 shows this variation between zymodeme groups using a rooted tree generating using R package SNPRelate. The SNPs used in this analysis are the same as those used in Figure 3.7. SNPs were taken from both designs for the seven strains assigned zymodeme groups. As evident from Figures 3.5-7, there is a high degree of similarity between all seven of these strains, and conserved SNPs between strains are redundant in constructing a phylogeny. Very few SNPs are unique to each strain, and inferring a relationship from less than a thousand SNPs over seven strains, as has been done here, is ill advised. This could be resolved if WGS data was available for all of the seven strains because a larger number of SNPs could be observed than from across one tenth of the genome. As is observed in Figure 3.4 and 3.6, zymodeme groups B17 and Z310 have a higher degree of similarity compared to the Z366 strain, strain O. This is seen by the branching of strain O separate to the remaining *T.b. rhodesiense* strains. Strains K and E, and strains B and M also cluster, which you would expect from the higher degree of conserved SNPs, but strains T and N, which have a higher degree of uniqueness, do not cluster with their zymodeme group.

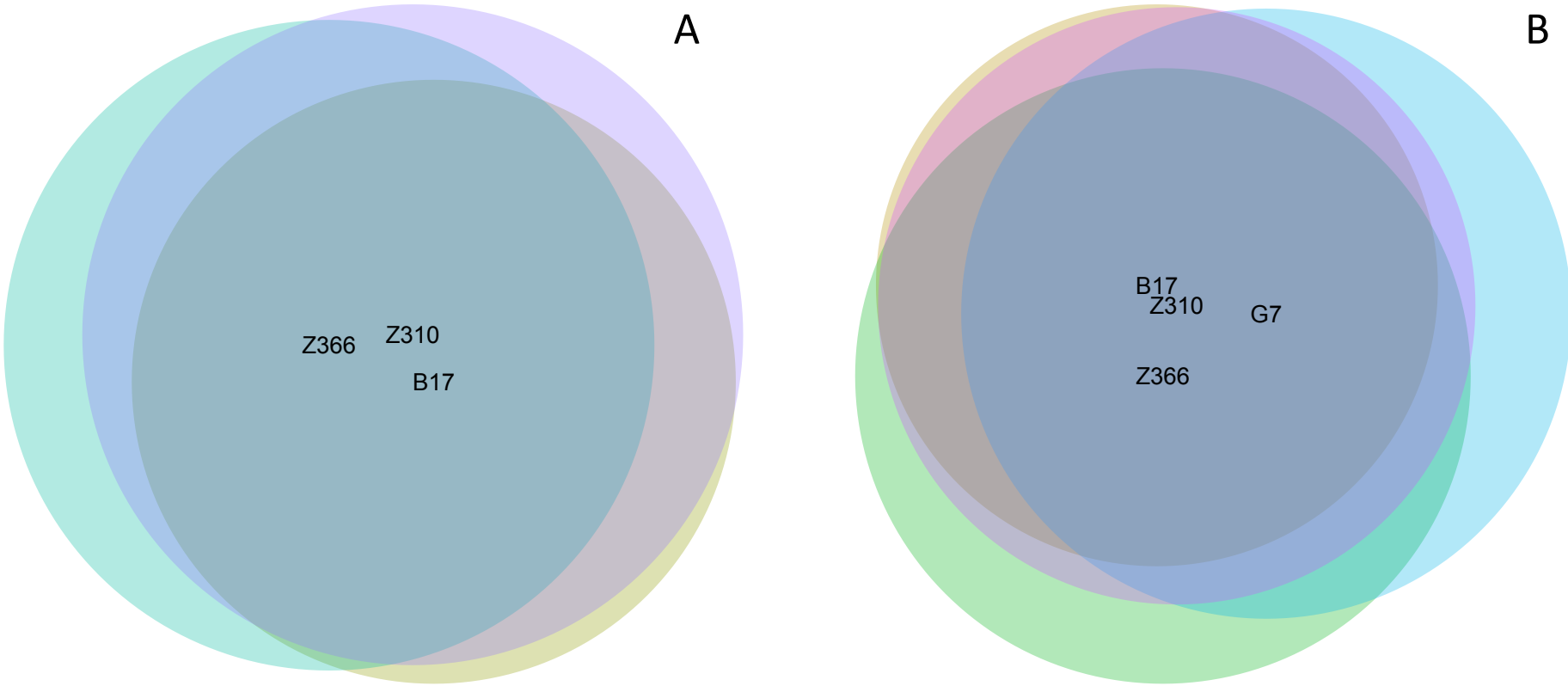

Figure 3.7:

A shows across zymodeme variation when including each of the 3 strains per zymodeme group for groups Z310, B17 and Z366. B shows these groups and T.b. gambiense strain G7 in addition. In A, 11,182 SNPs were conserved between all three zymodeme groups, and 531, 1,717 and 2,826 SNPs were unique to zymodeme groups B17,Z310 and Z366 respectively. 247 SNPs were overlapped between B17 and Z366, 1,784 between Z366 and Z310, and 1,862 between B17 and Z310. In B, 9,226 SNPs were conserved between all T.b. rhodesiense strains and G7, 160, 454, 1,607 and 2638 SNPs were unique to B17,, Z310, Z366 and G7 respectively. There were 142 SNPs conserved between B17, Z366, and G7, 178 between B17 and G7, 229 between G7 and Z310, 555 between B17,G7 and Z310, and 587 between Z366 and Z310, 623 between B17 and Z310, 952 between G7, Z366 and Z310, 1,051 between G7 and Z366, and 1,720 between all T.b. rhodesiense strains .

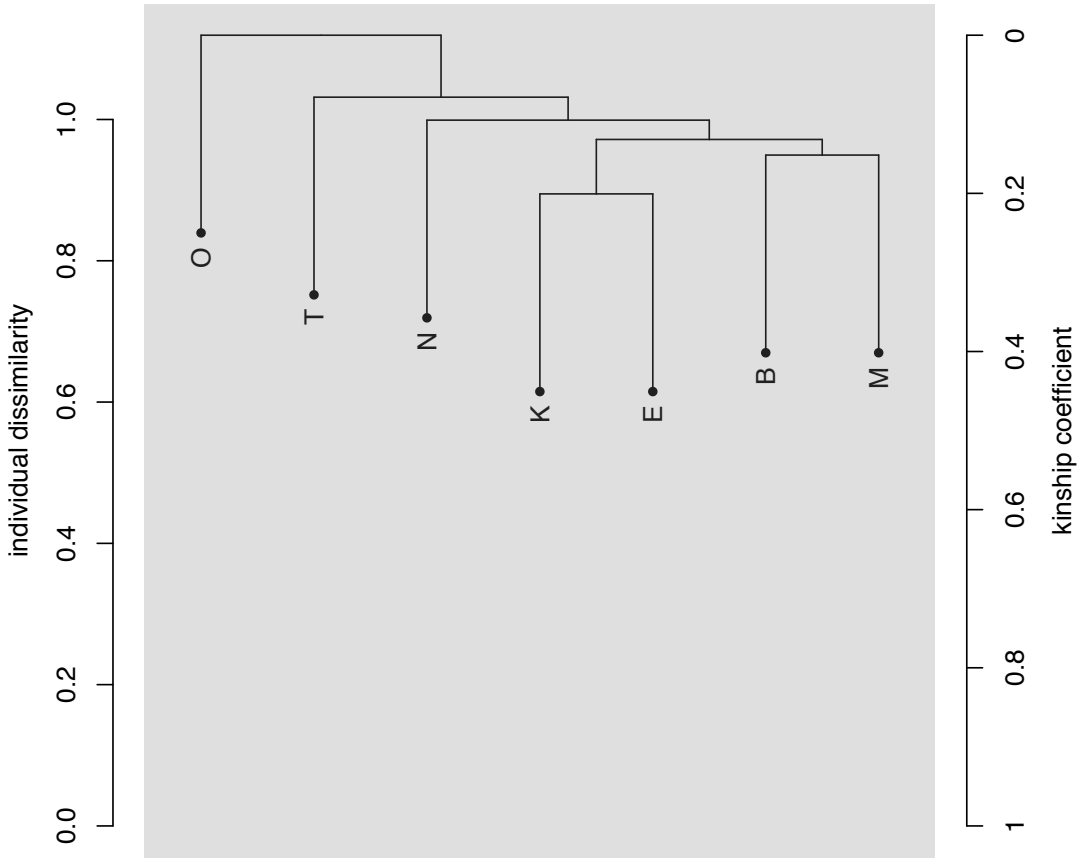

Figure 3.8 Rooted tree generating using R package SNPRelate. This was done using SNPs taken from both designs for the 7 strains used in both designs. Due to the high degree of conserved SNPs between strains, which are redundant in generating a phylogeny, very few SNPs are unique to each strain, and inferring a relationship from less than a thousand SNPs over seven strains is ill advised. This could be resolved if WGS data was available for all of the seven strains. As is observed in Figure 3.4 and 3.6, zymodeme groups B17 and Z310 have a higher degree of similarity compared to the Z366 strain, strain O. This is seen above. Strains K and E, and strains B and M also cluster, which you would expect from the higher degree of conserved SNPs , but strains T and N, which have a higher degree of uniqueness, do not cluster with their zymodeme group.

3.3.9 SNPs unique to a zymodeme group but shared between strains

13,044 SNPs were shared between all Z310 and B17 strains. However the genetic differences resulting in this difference in phenotype should be in the SNPs not conserved to both group, and so Figure 3.9 shows the number of SNPs unique to each zymodeme group per chromosome. Actual SNP counts are shown in bold in each corresponding bar, and the relative homozygous to heterozygous ratio is shown. A SNPs uniqueness to a zymodeme group was determined by SNPs that were common to all of the three strains within the zymodeme group, and the absent on this conserved

SNP in the other zymodeme group. As previously discussed, the B17 strains are more diverse and so less SNPs are conserved between strains. In comparison, there is less diversity between Z310 strains, and this is evident from the 3,501 SNPs conserved within the zymodeme group, but unique to B17, compared to the 778 SNPs in B17.

The SNPs per chromosome are plotted side by side for comparison in Figure 3.9, and so differences in zygosity can be observed. Large deviations in this ratio are seen on chromosome 5,6 and 8. In chromosome 5, 35% of the SNPs unique to B17 are heterozygous, compared to 75% in Z310. Similarly in chromosome 6, 14% of unique SNPs were heterozygous compared to 42% in Z310. However in chromosome 8, the heterozygotes found in B17 greatly outnumber those in Z310, and here is the largest difference, with a 68% increase in the heterozygote to homozygote ratio. The SNPs within this chromosome also contribute to 40% of the total SNPs within B17.

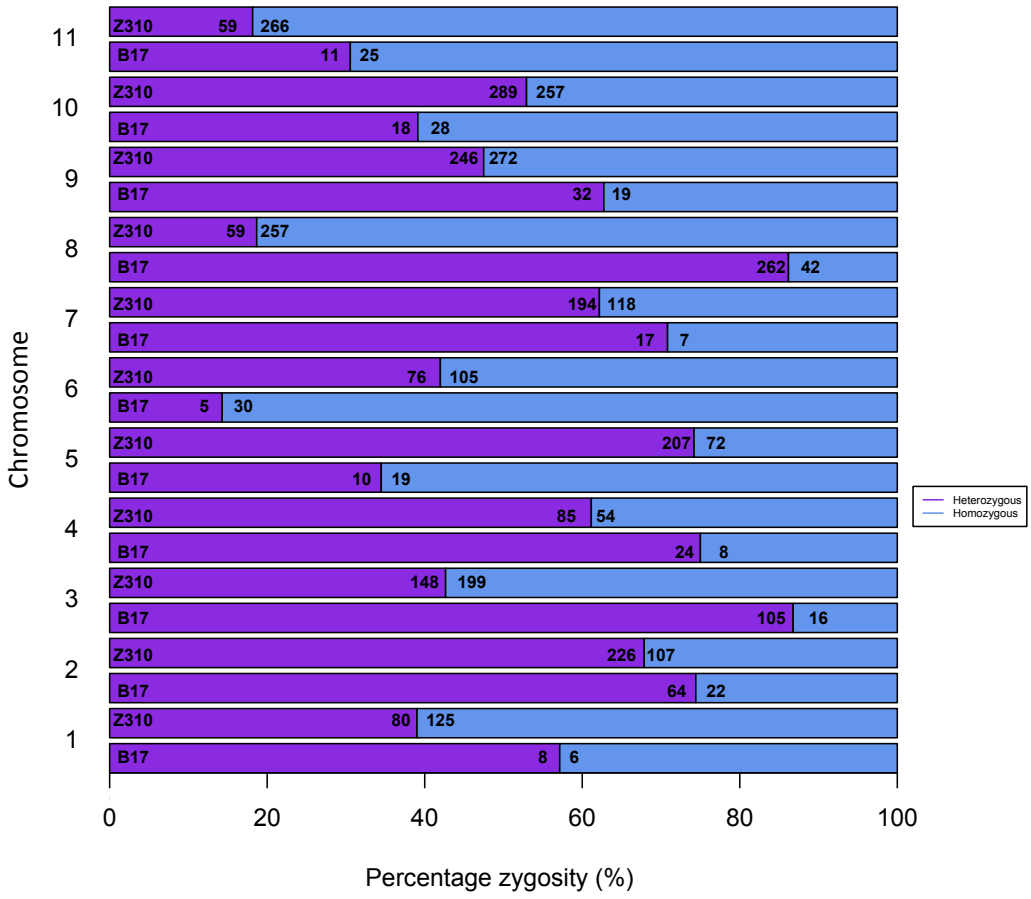

Figure 3.9: Shows the percentage of homozygous and heterozygous SNPs unique to one zymodeme group, but conserved between all strains of that zymodeme group. Heterozygous SNPs are shown in purple, homozygous in blue. The actual SNPs counts are shown in bold on each corresponding bar.

3.3.10 SNP frequency within the genes that contain these unique SNPs

In order to ascertain whether the genes with these unique SNPs are under selective pressure, the frequency of SNPs per gene were calculated. Figure 3.10 shows both zymodeme groups follow the same trend, with a tendency for a large number of gene containing less than 5 SNPs per gene, irrespective of the difference in total number of SNPs between groups. Several genes do include greater than 20 SNPs, however these are lower impact SNPs.

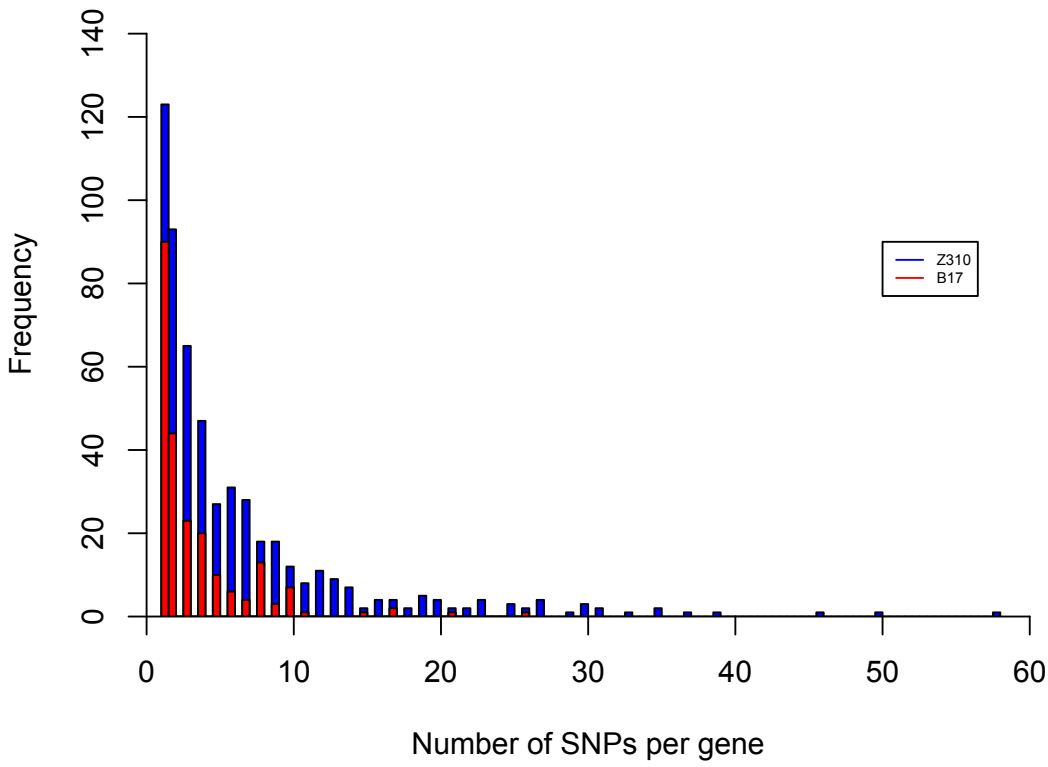

Figure 3.10: Shows the number of frequency of SNPs within genes with SNPs unique to each zymodeme group. SNPs from Z310 are shown in blue, B17 in red. B17 had less unique SNPs, but both groups have less than 5 SNPs in the majority of genes. Genes with a higher number of SNPs have generally lower impact or synonymous SNPs.

The genes with the highest number of unique SNPs are shown in Table 3.3. Only the genes with the five highest unique SNP frequencies are included. In Z310, all of these genes include greater than 35 unique SNPs, whereas the gene with the highest number of unique SNPs in B17 has 26. The majority of these genes have no assigned function and so it is hard to predict the potential effect. The genes with the highest frequency of SNPs in both zymodeme groups are both hypothetical. In each zymodeme group, there

are a high number of SNPs in a leucine rich repeat protein (LRRP) and these are important in the facilitation of protein-protein interactions and are commonly present at expression sites. However Z310 also has 50 SNPs present in the transmembrane component of a ferric reductase, and iron transport has previously been studied in trypanosoma and other species in relation to its importance in virulence. Iron is essential for parasite growth, and targets to disrupt iron transport pathways have previously been suggested for potential drug targets (Dean et al., 2014).

Due to the ability for kinases to activate and repress expression, it is interesting to note a high number of SNPs (39) conserved between Z310 strains in a protein kinase. Kinases have been extensively studied in other organisms and are often key virulence factors (Galyov et al., 1993; Saeij et al., 2006; Canduri et al., 2007). Due to the chronic nature of these strains, altered kinase activity could lead to reduced transcription and subsequent down-regulation of pathways and effect virulence. Interestingly, a high proportion of the genes containing the greatest number of SNPs in B17 are located on chromosome 8, suggesting phenotype altering effects could be located on this chromosome.

Table 3.3: Shown are the five genes within each zymodeme group, which have the greatest number of SNPs unique to its zymodeme group. Fewer SNPs are conserved between B17 strains but unique to that group, and so the number of SNPs seen in these genes is much lower.

Zymodeme group	Chromosome	Gene ID	Function	Number of SNPs
Z310	3	Tb927.3.600	Hypothetical protein	58
	11	Tb927.11.4430	Ferric reductase, transmembrane component	50
	2	Tb927.2.1380	LRRP	46
	7	Tb927.7.5220	Protein kinase	39
	5	Tb927.5.2510	Hypothetical protein	37
B17	8	Tb927.8.2390		29
	3	Tb927.3.580	LRRP	26
	8	Tb927.8.7200	Hypothetical protein	21
	8	Tb927.8.5050		17
	6	Tb927.6.900		17

3.3.11 SNPs conserved between zymodeme groups

SNPs that are conserved between zymodeme groups indicate that these are part of a "core" rhodesiense genome, rather than responsible for the differences in clinical manifestation. *T.b. rhodesiense* is often considered a host variant of *T. brucei*, with the transfection of *T. brucei* strains with SRA enough to confer resistance to human serum (Xong et al., 1998). All three subspecies of *T. brucei* are closely related however *T.b.brucei* and *T.b. rhodesiense* are more closely related than *T.b. rhodesiense* to *T.b. gambiense*. Due to the accumulation of random mutations following *T.b. rhodesiense's* divergence from *T.b. brucei*, although these SNPs may form a "core" rhodesiense genome, the majority of these SNPs are unlikely to have a functional impact.

In Figure 3.11A, the ratio of heterozygous to homozygous SNPs, and their associated counts per chromosome are shown for SNPs conserved in all Z310 and B17 strains. Their chromosomal location is shown in Figure 3.11B. As seen in Figure 3.11A, there is little variation in the zygosity ratio of the SNPs conserved, although there is a greatest relative number of heterozygotes in chromosomes 8, 9 and 10. These chromosomes also have the lowest numbers of conserved SNPs, suggesting key phenotype defining differences may be found within these chromosomes.

Figure 3.11B suggests that these conserved SNPs are not confined to one region of the genome; they are located uniformly across chromosomes, as we would expect.

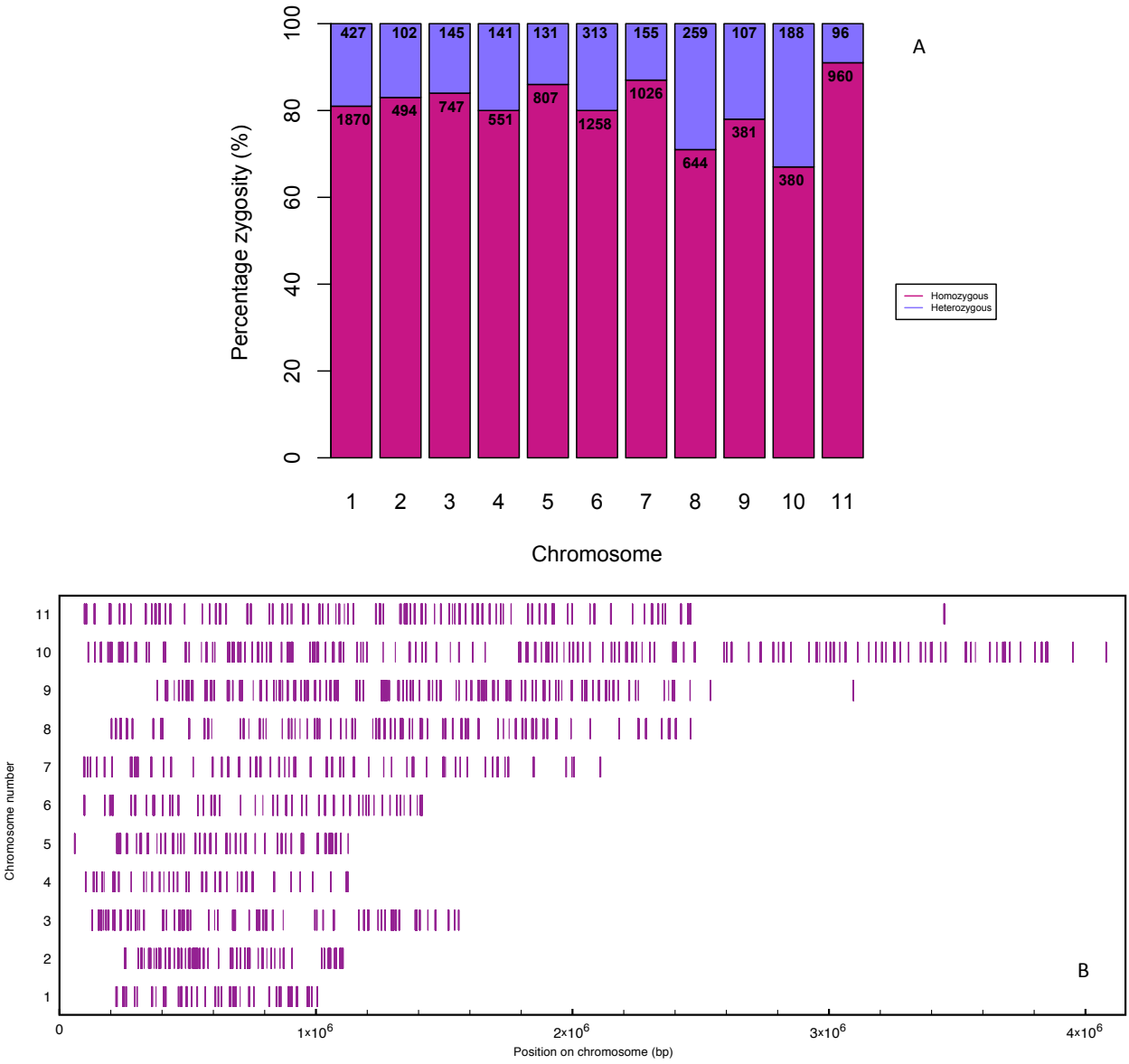

Figure 3.11: A and B both show the SNPs conserved between both the B17 and Z310 zymodeme groups in all strains. A shows the heterozygote to homozygote ratio per chromosome, and the number of SNPs per chromosome is shown. B shows the location of these conserved SNPs along each chromosome

3.3.12 Localisation of SNPs unique to a zymodeme group

In Figure 3.11B, it is evident that SNPs are being conserved genome-wide within all these rhodesiense isolates. Due to this, and the high degree of similarity between all three subspecies of *T. brucei*, small changes are likely to be responsible for this phenotypic disparity. In Figure 3.12A-C, the chromosomal location of SNPs unique to one zymodeme group are shown, in order to determine regions dense with unique SNPs.

SNPs unique to each set of strains are shown in Figure 3.12A. Those SNPs present in only B17 strains are shown in blue, those in Z310 only shown in red. Due to a greater number of unique SNPs in Z310, these unique SNPs are spread evenly throughout the genome, however the SNPs unique to B17 are more restricted to specific regions. In particular, there is a high density of SNPs unique to B17 on chromosome 8. Clusters of unique SNPs in B17 are also present on the end of chromosome 3 and chromosome 2. The conservation of these SNPs within all three B17 strains suggest these have an impact on virulence rather than individual strain differences.

WGS data is only available for strain B and E, and so a comparison of just these strains is shown in Figure 3.12B, in order to see any strain specific differences. As observed in Figure 3.12B, the SNPs are localized similarly to the pattern seen when using multiple strains from the same zymodeme group. Additional unique SNPs are found throughout the genome, but these don't form dense regions of SNPs like those observed in chromosome 8. The occurrence of low density additional unique SNPs dispersed throughout the genome can be accounted for by individual strain diversity.

Strain O is used in Figure 3.12C for three-way zymodeme comparison. As previously discussed, Strain O has a greater number of unique SNPs, and appears to be more closely related to Z310. As with Z310, these unique SNPs are not localized to one region of the genome. However there is a high density of SNPs unique to Z366 within chromosome 6, which is seen in neither Z310 nor B17.

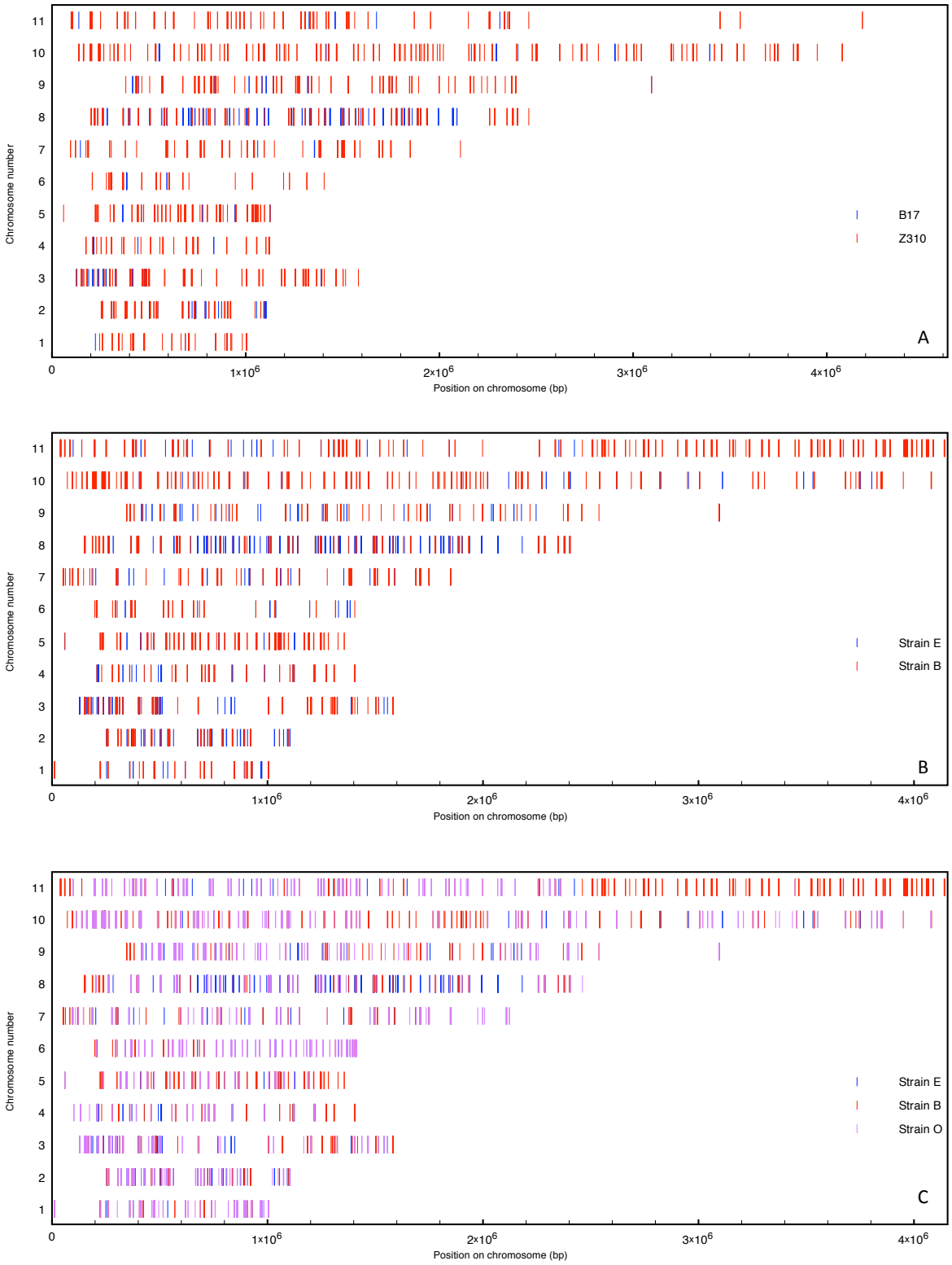

Figure 3.12: A-C show the chromosomal locations of SNPs . A shows the SNPs that are shared between all strains within a zymodeme group, so strain B,M and T for Z310, and K,E and N for B17, but not present in the other respective zymodeme group. B shows the SNPs unique to strain B and E compared to each other. C shows the positions of SNPs unique to strain B, E and O compared to each other, which represent zymodeme groups Z310, B17 and Z366 respectively.

3.3.13 Mapping chromosome 8 to DAL972

Due to the abundance of SNPs along chromosome 8, strains B, E and O were mapped to DAL972 version 8.1 alongside gambiense strain G7. SNPs found for each strain from both designs were combined as before. From these only SNPs from genes included in both designs on chromosome 8 were extracted. This was due to G7 not being used in the first design. The chromosomal position of SNPs unique to either strains B,E or O are shown in Figure 3.13. The SNPs shown in G7 were compared against the SNPs found in in all three rhodesiense strains, in order to determine the level of "gambiense specific" SNPs expected along chromosome 8. In comparison to the number of SNPs seen in Figure 3.13, there are significantly fewer SNPs in chromosome 8 for all three rhodesiense strains, particularly in strain B. This suggests chromosome 8 is more "gambiense-like".

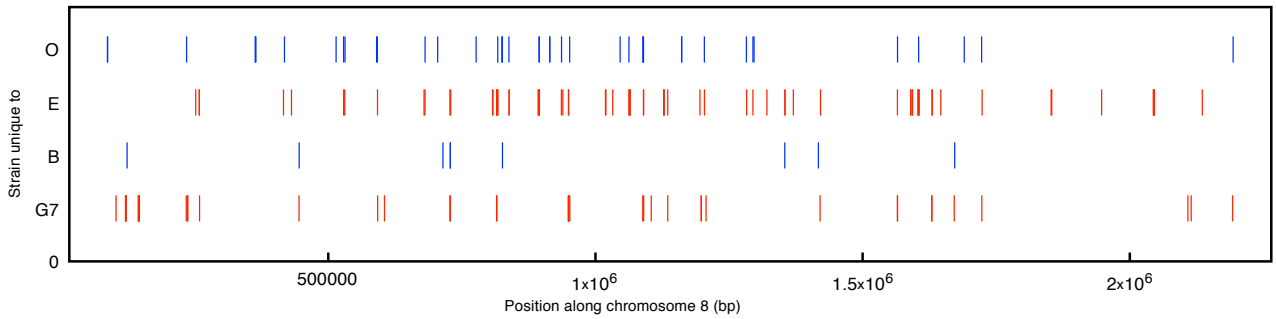

Figure 3.13: This shows the SNPs present when reads mapping to chromosome 8 are mapped to the DAL972 reference. Shown are SNPs unique to strains B,E,O and G7. The Z310 strain, strain B, has markedly fewer SNPs to DAL972 than the other strains. This includes G7, which is a T.b. gambiense strain.

3.3.14 Functional implications for SNPs

In order to appreciate the functional effect of these SNPs they were annotated using SNPeff and SNPsift. This was first used to ascertain whether the SNPs were synonymous or non-synonymous and whether the non-synonymous SNPs were likely to have a deleterious effect. These SNPs were then divided into 4 impact groups, low, moderate, modifier and high. Synonymous variants were considered to be low impact because they won't change the protein product. SNPS considered moderate were those that were unlikely to significantly change the protein produced, but might alter its

interactions for instance with other proteins. Examples of moderate impacts include mutations causing missense variants, or deletions that do not alter the reading frame. Variants labeled modifiers were generally restricted to non-coding regions, and have an impact on regulatory regions such as un-translated regions (UTRs). The SNPs with the greatest impact were annotated high impact, and these were presumed to have a greatly disruptive effect on the protein product. Variants causing a high impact include but are not restricted to stop gained and lost effects.

Variants were also classified as missense, nonsense or silent mutations. Silent mutations were synonymous, and missense mutations accounted for the majority of the moderate and modifying mutations. Nonsense mutations are comparatively rare to the other categories, as these cause higher impact effects. These are both relatively broad ways of categorizing SNPs, and so the individual predicted SNP effects are also shown in Figure 3.14.

Figure 3.14A-C shows the percentage of SNPs annotated as described above. The combination of strains used to generate the SNP set are shown on the left hand side of the figure. The first two bars represent the SNPs that are unique to just these strains. The following bars represent the SNPs unique to an individual zymodeme group, for Z310 and B17 this includes all three strains, for Z366 this includes just strain O. The SNPs shared between zymodeme groups B17 and Z310 and between all three zymodeme groups are also shown.

The majority of SNPs annotated across all strain combinations caused a modifying effect, and this can be seen in Figure 3.14A. This accounted for greater than 80% of the SNPs in each strain combination. The next largest category were the low impact SNPs, this was the most variable category, but accounted for no greater than 10% of all of the SNPs annotated across strains. This was followed by the moderate impact SNPs, and lastly by high impact SNPs. High impact SNPs are shown on the Figure 3.14A, however they represent such a small proportion of the SNPs identified, that they are not apparent.

In Figure 3.14B, it appears that the ratio of missense, nonsense and silent mutations does not alter greatly between strains, with just greater than 40% of the SNPs being missense mutations, the majority of the remaining SNPs being silent mutations, and a small percentage nonsense. However strain O (Z366) and strain E do deviate from this

slightly, and have an increased number of missense mutations, with approximately 50% missense, and a greater number of nonsense mutations, particularly in Z366.
Individual SNP types are shown in Figure 3.14C and were predominantly from regulatory regions causing downstream and upstream effects across all strain combinations. These accounted for approximately 80% of the SNPs annotated. The majority of the other SNPs were predicted to have an effect in an intergenic region, whether missense or synonymous. Out of these categories the number of intergenic SNPs differed the most between different strain combinations. The greatest number of these were unique to strain B, seconded by those in the Z310 strains. There were other SNP outcome types within the data which only represent a very small proportion of the data and so aren't evident in Figure 3.14C, these were the stop lost/gained and start lost events. Figures 3.14A-C demonstrate no significant deviations in the SNP effects seen, and so this change in virulence can not be associated with the change in abundance of a specific SNP effect.

3.3.15 Localization of different predicted SNP effects

In order to determine whether more deleterious SNPs were clustering at certain regions of the genome, the different SNP effect groups, low, moderate, modifier and high, were plotted for strain B, E and O to represent groups Z310, B17 and Z366 respectively in Figure 3.15. As before, the SNPs plotted are those unique to that strain. Figure 3.15A contains SNPs unique to strain B, and all low, high and modifying SNPs are dispersed evenly throughout the genome. However when comparing this to Figure 3.15B, in which strain E's unique SNPs are shown, it is evident that a greatly increased number of SNPs with predicted moderate effects are located after 2.5 $x10^6$ base pairs on chromosome 11. A higher abundance of moderate SNPs are also positioned between positions 1- 0.5 x 10^6 on chromosome 10 in strain B, which is not observed in strain E, but is also seen in strain O, which is shown in Figure 3.15C. As with strain B, the SNPs in strain O are not particularly clustered by effect, and are shown in Figure 3.15C. In contrast, strain E SNPs which are shown in Figure 3.15B, consist of fewer moderate SNPs, however these are primarily clustered around chromosome 8, and a smaller cluster is seen between positions 0-0.5 x 10^6 on chromosome 3. The other SNP effects appear equally spaced around the genome.

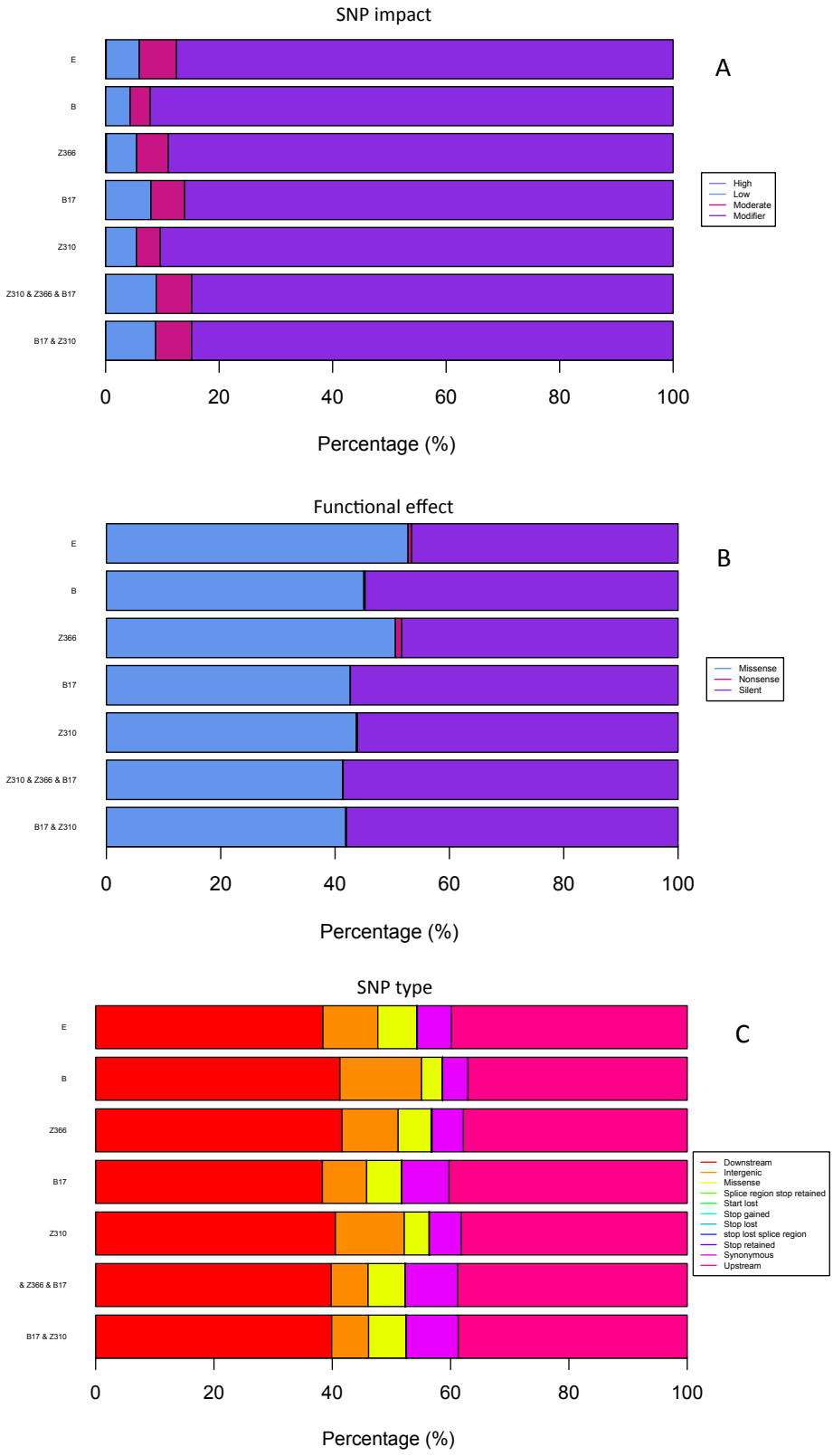

Figure 3.14: A-C show the relative percentages of the SNPs annotated using SNPeff and SNPsift. A categorises the SNPs into severity groups, low, moderate, modifying and high. B categorises whether the SNPs are synonymous/silent or if they result in a missense or nonsense effect. C shows a breakdown of these effects into the specific types.

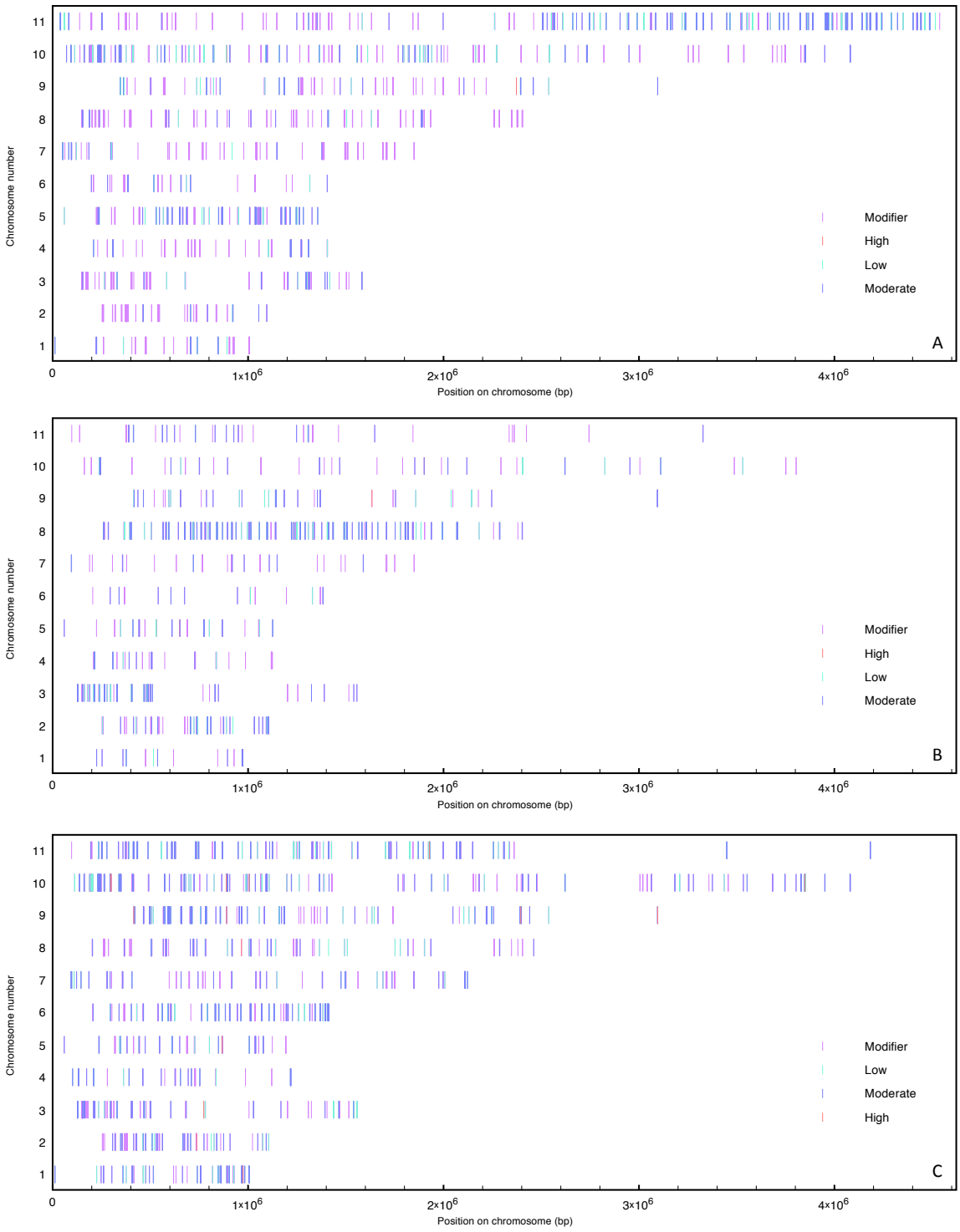

Figure 3.15: A-C show the localisation of the four different SNP effect groups, low, moderate, modifying and high. A shows the unique SNPs for strain B, B for strain E, and C for strain O, and these collectively represent zymodeme groups Z310, B17 and Z366.

3.3.16 High impact SNPs unique to specific zymodeme groups

Fewer high impact SNPs were unique to an individual strain or zymodeme group than were conserved across zymodeme groups. No high impact SNPs were seen across all B17 strains that weren't observed in Z310 strains, however two were found conserved just between Z310 strains, as shown in Table 3.4.

Table 3.4: High impact SNPs, and the combination of strains they are conserved to. *The SNP in gene Tb927.10.16400 is in a different position to that shown conserved between all seven strains. **There are also two high impact SNPs within gene Tb927.10.1110, both have stop gained effects. ***The SNP position in Tb927.10.3820 is different to the one shown conserved between Z310 strains.

SNP conserved in	Chromosome	Gene ID	Functional effect	Gene function
Z310 strains	2 high impact SNPs			
	10	*Tb927.10.16400	Stop gained	VSG
	10	Tb927.10.3430		Hypothetical protein
B17 strains	No high impact SNPs were conserved between the B17 strains			
Z366 strain	15 high impact SNPs			
	1	Tb927.1.4740	Stop gained	Hypothetical protein, conserved
	2	Tb927.2.4200		Protein kinase
	3	Tb927.3.3020		Actin-like protein
	5	Tb927.5.2510		Hypothetical protein, conserved
	8	Tb927.8.3240		Lathosterol oxidase
	9	Tb927.9.1770		Hypothetical protein, conserved
	9	Tb927.9.4900		Condensin subunit 1
	9	Tb927.9.15400		Ankyrin-repeat protein
	9	Tb927.9.18100		VSG fragment
	10	**Tb927.10.1110		Kinesin
	10	Tb927.10.3430		Hypothetical protein, conserved
	10	***Tb927.10.3820		
	10	Tb927.10.15680		
	11	Tb927.11.6830		
Strain B	2 high impact SNPs were unique to strain B			
	9	Tb927.9.15290	Stop gained	Hypothetical protein, conserved
	10	Tb927.10.3430	Stop gained	
Strain E	5 high impact SNPs unique to strain E			
	1	Tb927.1.4740	Stop gained	Hypothetical protein, conserved
	5	Tb927.5.1490		Eukaryotic translation initiation factor 4 gamma type 1 (EIF4G1)
	5	Tb927.5.1910		Hypothetical protein, conserved
	9	Tb927.9.10440		DNA polymerase epsilon catalytic subunit
	10	Tb927.10.7550	Start lost	Hypothetical protein, conserved

However the Z366 strain had 15 high impact SNPs that were unique just to that strain, of these 7 had no assigned function. One gene, Tb927.10.1110 had two high impact SNPs. Gene Tb927.10.4820 had a high impact SNP in both the Z310 strains and Z366, however these were at different positions. Similarly gene Tb927.10.16400 contained a high impact SNP unique to the Z310 strains in addition to the one conserved between all rhodesiense strains seen in Table 3.4. When looking at SNPs unique to one strain, strain B has fewer high impact SNPs, with two in hypothetical proteins, compared to the five seen in strain E, three of which are also hypothetical proteins.

Compared to the SNPs conserved between all *T.b. rhodesiense* strains, which allude to differences which might define differences in virulence between *T. brucei* subspecies, SNPs specific to particular rhodesiense strains can elucidate the phenotypic differences seen between the different zymodeme groups. The functions of the genes containing these high impact SNPs, which weren't covered previously, are briefly described below.

Actin is not differentially expressed in procyclic and bloodstream forms, however it does localize to a different region of the parasite depending on its stage in the life cycle. In procyclics it is present throughout the cell, in bloodstream forms, it co-localizes with the flagella pocket, which is highly polarized and the site of all endocytic activity within the parasite. Bloodstream forms are heavily reliant on endocytosis whereas this is severely reduced in procyclics (Morgan et al., 2002; Garcia-Salcedo et al., 2004). Actin is essential for bloodstream forms, rapidly leading from cell cycle arrest to cell death when depleted (Garcia-Salcedo et al., 2004). Procyclic forms do undergo gross malformations in the golgi following depletion, but this does not arrest growth or kill the parasite (Garcia-Salcedo et al., 2004).

The SNP residing in a gene which produces an actin-like product was seen in the Z366 strain, which has a moderately virulent phenotype, however the abundance of different cell cycle stages, which was significantly different in B17 and Z310, is not known for this strain. The effect of this mutation on stumpy forms may be similar to that seen in the procyclics, and this indicates that perhaps Z366 virulence is determined by stumpy forms predominating the infection, as was seen in B17 strains.

Bloodstream forms of *T. brucei* cannot synthesize sterols de novo, whereas procyclic forms can (Coppens and Courtoy, 2000). Lathosterol oxidase is involved in sterol production and causes the formation of 7-dehydrocholesterol. Depletion of sterols can

lead to aberrant growth of the parasite in the procyclic forms, and so although mutations in lathosterol oxidase (as seen in Z366) may not directly affect the normal function of the bloodstream forms, it may impact on the insect stages that the bloodstream forms differentiate into, affecting infectivity (Pérez-Moreno et al., 2012). There is no available microscopy and QPCR data for Z366, and so the relative bloodstream form abundances are unknown. This makes judging the potential adverse effects of this mutation more difficult.

Ankyrin repeat proteins facilitate protein-protein interactions and so are important in a wide variety of cellular mechanisms (Al-Khodor et al., 2010). Despite the unknown function of this gene, proteins with ankyrin repeats in them have been known to be involved in the regulation of the cell cycle, transport within the cell and the stability of the cytoskeleton. However without a known function, it is hard to predict the potential consequences to the parasite.

Z366 also has a high impact SNP in a kinesin gene. The kinesin heavy chain is essential for the colonization of the mammalian host because it causes upregulation of IL10, which leads to an increase in arginase activity, which is required for colonization and growth during the acute stages of infection (Beschin et al., 2014) Kinesins are known to have functions in multiple pathways however many of the kinetoplastid specific kinesins have not been characterized, although one has been found involved in the maintenance of normal cell morphology and cytokinesis (Hu et al., 2012). Z366 also has a mutation in a condensin gene, and these are responsible for the structural integrity and organization of chromosomes, suggesting mutations in multiple genes involved in the structural integrity of the parasite (Hirano, 2012).

However strain E also has high impact mutations within key genes. Both eukaryotic translation initiation factor 4, gamma type 1 (EIF4G1) and DNA polymerase epsilon catalytic subunit, are essential for parasite growth, with loss of function in either of these leaving the parasites unviable, because these genes are essential for transcriptional and translational control (Dhalia et al., 2005; Zinoviev and Shapira, 2012).

3.3.16 GO term analysis

GO term analysis was carried out on the annotated SNPs unique to each representative strain for each zymodeme group, strains B,E and O. This was done to look at whether genes related to particular pathways were enriched for. Due to the neutral model of evolution, the majority of the SNPs accrued are most likely to be from multiple pathways and be of little functional importance. This was observed particularly with the low and moderate annotated SNPs, with SNPs in genes from multiple pathways, but very few significantly enriched. Figure 3.16 and 3.17 show the GO term annotations for each impact group for each strain. Each circle represents a GO term, with the greater the P value assigned, the larger the circle. The least unique GO terms have been collapsed into single GO terms. The relative distance between the GO terms reflects the degree of similarity between GO terms. The greatest number of GO terms were assigned to modified and low effect groups. This analysis can be used to look at whether the same pathways are being enriched across strains and SNP impact categories.

3.3.16.1 Pathways seen in modifying SNPs

There were 21, 34, and 51 GO terms assigned to the genes with annotated as modifying in strains B, E and O respectively. In B the majority of these GO terms were related to general metabolic processes affecting many functions of the parasite. This is shown in Figure 3.16A, where the processes enriched do not appear closely related and cluster together. However the most enriched GO term was autophagy, which is interesting because an increase in autophagic activity has been linked to differentiation of the parasite during high parasite densities, and this strain's infections are dominated by the highly proliferative slender stages (Schmidt and Bütikofer, 2014). Base editing, post translation modifications and recombination processes were also enriched, as was golgi vesicle transport, although this is unsurprising considering there were conserved SNPs in the Rab proteins between *T.b. rhodesiense* strains.

Unlike in strain B, the pathways upregulated in strain E were more clustered, as seen in Figure 3.16B. In the uppermost left of the plot, multiple GO terms are collapsed into protein localization, with approximately a third of the GO terms from this category associated. In strain E, the non-proliferative stages dominate infection, and multiple pathways are required for the transition of long slender (LS) to short stumpy (SS)

differentiation, including cytoskeleton rearrangement and protein relocation (Rotureau et al., 2011; Matthews et al., 2004). Small GTPase activity is also enriched, as with strain B, this is expected due to the high impact SNP in a Rab escort protein, which suggests these pathways are affected.

In strain O, polyol metabolism and translational termination are the most enriched GO terms. Bloodstream forms rely on a constant source of glucose, and polyol metabolism is involved in the conversion of excess glucose (Uzcategui et al., 2004; Vertommen et al., 2008). Rapidly dividing forms will consume more glucose, and so an upregulation in this pathway could suggests there is a greater number of non-proliferative forms present in the infection. As seen in Figure 3.16C, genes from a wider variety of pathways contain modified SNPs compared to strains B and E, however pathways do appear to cluster similarly to strain E. Several GO terms are collapsed to cellular localization as was seen in strain E. Post translational protein modifications were also enriched as was seen in strain E.

Modifying effects were the most abundant SNP impact group, and cause regulatory effects. GO terms for various pathways were enriched, showing which pathways these modifying effects are impacting. Despite differences between each strain, similar pathways were enriched across strains including post translational protein modification GTPase activity and vesicle transport. Although not excessively enriched, with no pathway with a greater than five fold enrichment, many of these pathways suggest differences in regulation related to the relative abundance of life cycle stages present in the infection.

3.3.16.2 Pathways seen in moderate SNPs

There were fewer SNPs annotated moderate, and the genes containing these were assigned to 24,50 and 9 GO terms in strains B,E and O respectively and are shown in Figure 3.16D-F. These SNPs affect a greater number of pathways compared to those seen in the modifying SNPs in strains B and E, and this is shown by the number of GO terms associated.

As seen previously, the pathways of the genes containing these mutations do not cluster by function in strain B compared to strain E and O. However aspects of lipid and steroid metabolism are highly enriched, although this may be an artefact as these GO

terms have few genes allocated to them. However pyrimidine and pseudouridine biosynthesis is highly enriched. Bloodstream forms can synthesize pyrimidines *de novo*, however they are also capable of salvaging from exogenous pyrimidine sources, making them more reliant on the host (Ali et al., 2013b). Lack of pyrimidines is lethal for the parasite, but parasites can maintain an infection just through salvage. However parasites incapable of *de novo* synthesis have been shown to be less infective to mammalian hosts (Ali et al., 2013b) . Uracil uptake was also upregulated in these pyrimidine starved parasites, and interestingly synthesis of the glycosylated isomer, uradine, is upregulated in this strain (Ali et al., 2013a). Z310 strains, such as strain B do cause chronic symptoms, and the long slenders, which predominate in infections with this strain, need to differentiate into short stumpy forms before being infective to vectors. The morphology of the pyrimidine starved parasites is not discussed, however their propensity to not differentiate will affect their infectivity.

In strain E, as shown in Figure 3.16E, and seen in Figure 3.16B, the genes containing these SNPs cluster, with enrichment seen particularly in lipid metabolism, macromolecule localization, and in particular lipid transport. *T. brucei* is capable of synthesizing its lipids de novo, however due to the high demand for glycolipids for the VSG coat in bloodstream forms, *T. brucei* may scavenge from it's host when these lipid resources are low, and the parasites are rapidly proliferating (van Hellemond and Tielens, 2006). In strains E, non-proliferative forms are more abundant, and so the parasite may be able to facilitate more of its own lipid metabolism.

Transcription from the RNA polymerase II promoter is also enriched, and this is complex mechanism in *T. brucei*, partly due to the trans-splicing nature of the genome (Das et al., 2008). However again, the importance of this enrichment may be overestimated due to few genes being assigned to this GO term.

Fewer genes caused moderate effects in strain O, as is evident from Figure 3.16F. However the pathways that are affected by these SNPs do cluster. As is seen with strain B, pseudouridine synthesis is potentially impacted by the SNPs present within these genes. However the most enriched GO terms were cytoskeleton, microtubule and golgi vesicle transport. This is anticipated from the earlier analysis in which this strain had mutations in genes important in the maintenance of cell morphology and vesicle transport.

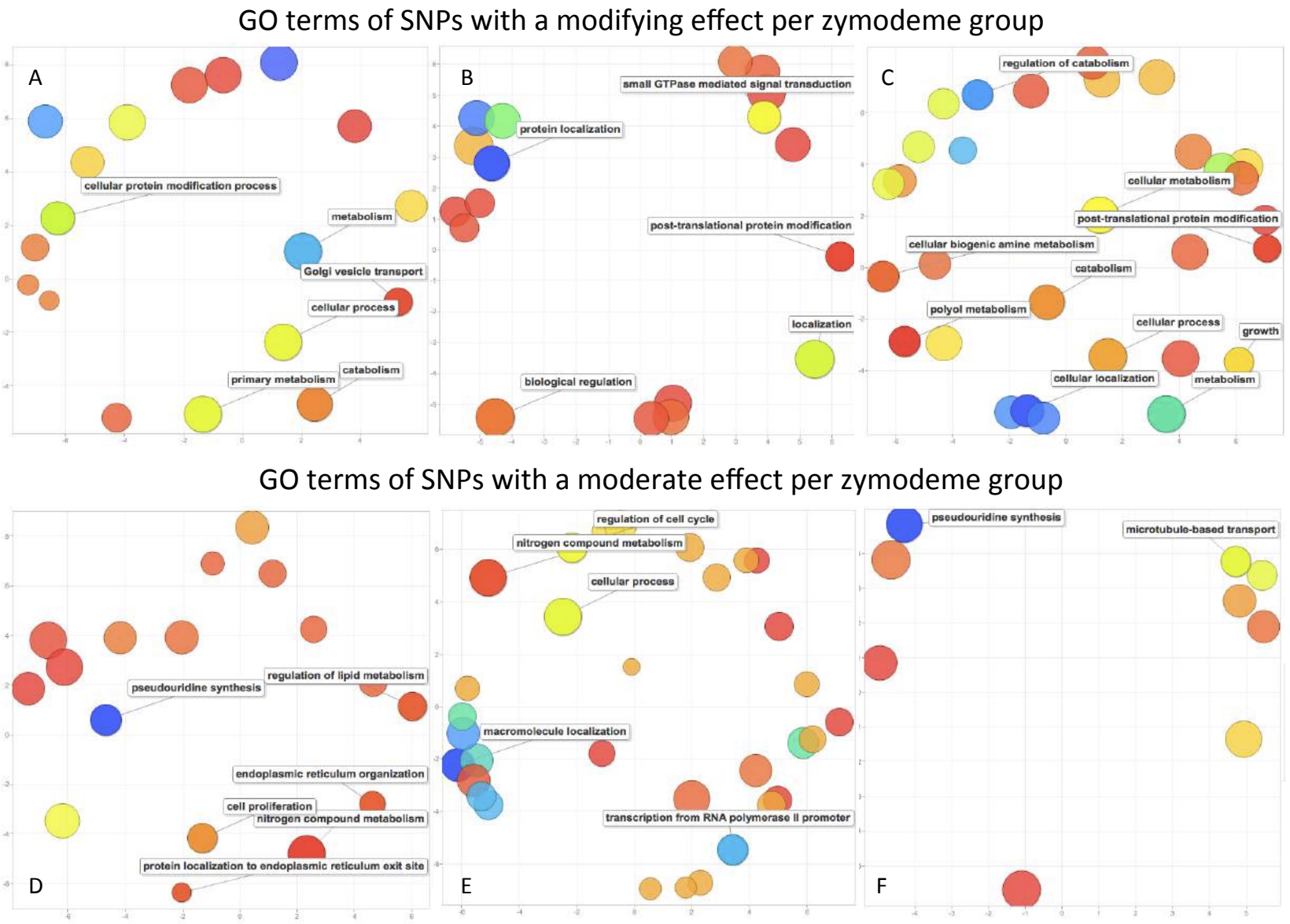

Figure 3.16: Shows plots made using REVIGO to reduce the number of GO terms seen within a sample. A-C show GO terms for the genes with modifying effects in them in strains B,E and O in A,B and C respectively. D-F show GO terms for the genes with moderate effects unique to each zymodeme group, ordered by strain as with A-C.

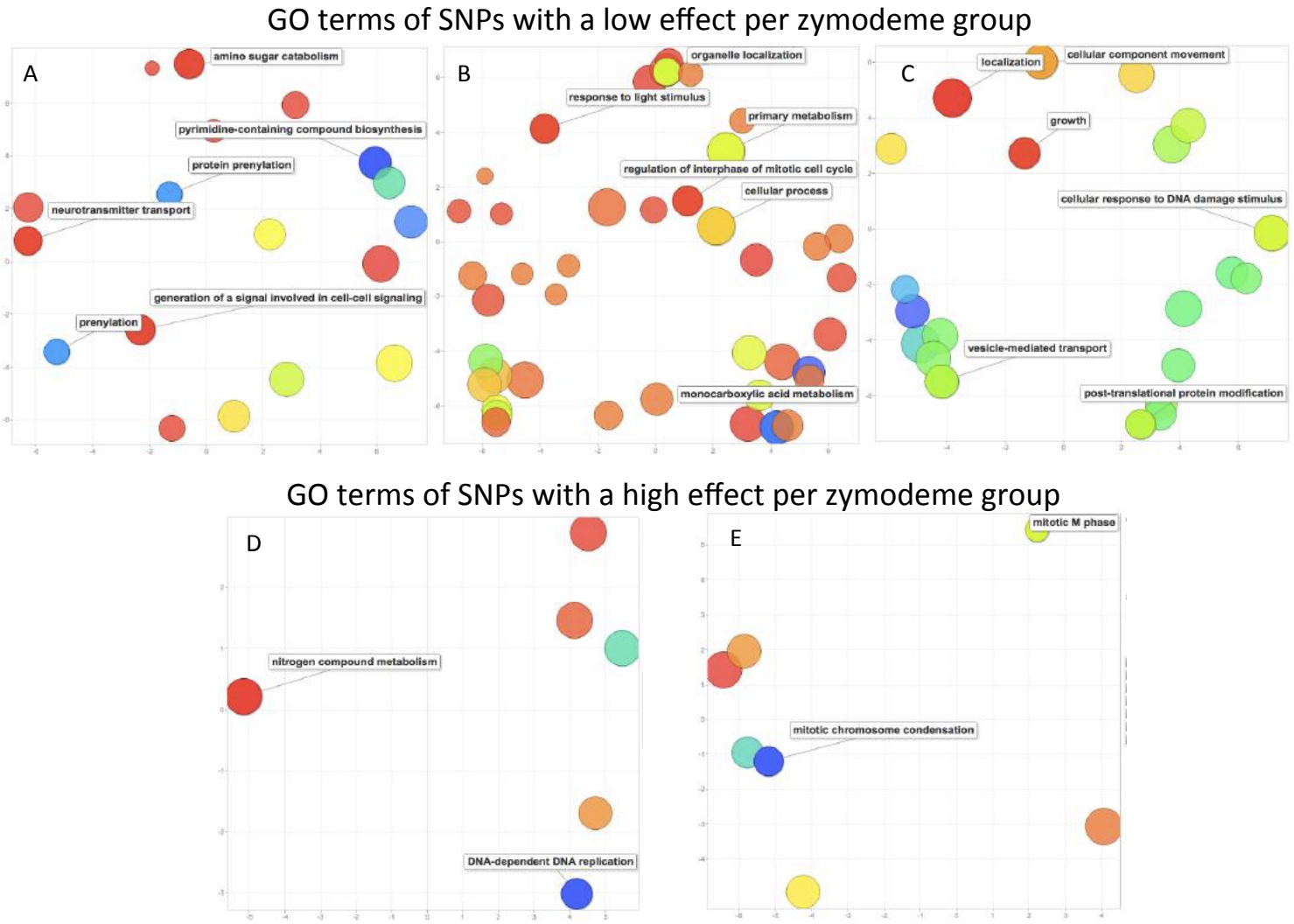

Figure 3.17: Shows plots made using REVIGO to reduce the number of GO terms seen within a sample. A-C show GO terms for the genes with low effects in them in strains B,E and O in A,B and C respectively. D-F show GO terms for the genes with high effects unique to each zymodeme group. Only strain E and O are shown, in D and E because B didn't have any annotated high effect SNPs

3.3.16.3 Pathways seen in low impact SNPs

This category has the greatest number of GO terms associated, with 33, 66,31 for strains B,E, and O respectively. However the pathways affected by SNPs in this category are of the least importance, because the majority of the SNPs contained within these genes are non-deleterious. However it is interesting to note that as before, more GO terms cluster in strains E and O compare to strain B. The majority of these GO terms relate to general house keeping metabolic functions, such as lipid and protein metabolism and pyrimidine biosynthesis. Neurotransmitter and cell-to-cell signaling, as shown in Figure 3.17A refers to a mutation in syntaxin, which is important for neurotransmission in vertebrates but is important for organelle stability and organization in protists (Dacks and Doolittle, 2002).

3.3.16.4 Pathways seen in high impact SNPs

Figure 3.17D-E show the GO term associated with high impact SNPs in strains E, in Figure 3.17D, and strain O in Figure 3.17E. This is because there were only two high impact SNPs in B, and these couldn't be assigned a GO term because they had no functional annotation, see Table 3.4. For both of these strains, the GO terms were all primarily related to cell cycle control and DNA replication, both of which are important for differentiation. The differences in life stage abundance and severity of disease between zymodeme groups do suggest differentiation impacting on virulence, and SNPs with potentially very severe functional effects within these pathways could account for this.

3.3.17 Mutations within known virulence factors

As previously discussed, virulence has been defined in trypanosomes in a variety of ways, however little attention has been given to the genes causing these phenotypic changes (Morrison, 2011). Beneath are three genes, which are implicated in the progression of a *T. brucei* infection, and their function. Included within the sequence capture design were two genes known to be important virulence factors, these were the haptoglobin haemoglobin receptor (HpHbr) and oligopeptidase B. The HpHbr gene is responsible for resistances to lysis. Mutations specific to a zymodeme group have been found in both of these genes and are discussed below.

3.3.17.1 Haptoglobin haemoglobin receptor (HpHbr)

The resistance *of T.b. gambiense* and *T.b. rhodesiense* parasites to lysis by the host is done through two different mechanisms. *T.b. rhodesiense's* resistance strategy is better understood, and is facilitated by a gene specific to *T.b. rhodesiense strains*, the serum resistance associated protein (SRA) (Njiru et al., 2004). In *T.b. gambiense,* lysis is averted by altering the activity of the haptoglobin haemoglobin receptor. Initially this was thought to be across all gambiense isolates, however as shown in 2012 by Symula and colleagues, the *T.b. gambiense* subspecies is further divided into two groups, *T.b. gambiense* group 1 containing a mutation causing this resistance, and group 2 strains resist lysis via an alternative mechanism (Symula et al., 2012).

In susceptible trypanosomes, a complex called the trypanosome lysis factor 1 (TLF1), containing the haptoglobin related protein (Hpr) and the lipoprotein, APOL1, bind to the surface of the HpHbr protein and are internalized (Symula et al., 2012). This internalization causes cell death. Due to its importance in resistance, this gene was included in the panel of genes used in the target enrichment design.

In the haptoglobin haemoglobin receptor, there were eighteen SNPs, seven unique to strain B, one to strain E, ten conserved between both zymodeme groups. The SNPs unique to strain B were in a non-coding region downstream in the gene, and had a modifying effect. Strain E had one unique SNP, and this was also in a non-coding region upstream. In the conserved SNPs, one was synonymous and seven of the ten were within the coding region and caused missense mutations.

3.3.17.2 Oligopeptidase B

Oligopeptidase B is a serine protease that has long been associated with virulence in *T. brucei* but also in *Leishmania* and *T.evansi* (Morty et al., 1999; 2005). It is responsible for the hormonal downregulation seen in these infections, by lysing lysine and arginine peptides (Kangethe et al., 2012). As a byproduct of parasite death from lysis, multiple immunogenic products are released into the host's blood and lymphatic system (Kangethe et al., 2012). These include a variety of lipids and also enzymes such as peptidases, and lead to the variety of symptoms associated with trypanosomiasis. One perpetrator in this process is oligopeptidase B. In particular this has been associated with causing neurological symptoms and assisting with the passing of parasites

through the blood brain barrier (Kangethe et al., 2012). However it important to note that oligopeptidase B is not an essential gene, and it is likely other peptidases take over its function in its absence (Kangethe et al., 2012).

In oligopeptidase B, three SNPs found were unique to strain B, three unique to strain O, and none to strain E. However there was conservation between strains B and E, with twenty SNPs in this gene present in both strains B and E. Of these, ten were identified as non-synonymous, six of which caused a missense mutation within the coding region, and the remaining were located within regulatory regions, with one downstream, and three upstream of the coding region.

3.3.17.3 Cathepsin B and cathepsin L (brucipain)

Cathepsin B and cathepsin L, also known as brucipain, are both cysteine proteases, which have been associated with helping to establish *T. brucei* infections within the mammalian host (Abdulla et al., 2008). Since 1999, it has been acknowledged that cysteine protease inhibitors could be used to prolong the life span of mice infected with a lethal dose of *T. brucei* (Scory et al., 1999; Troeberg et al., 1999). However it wasn't until 2004 that these cysteine proteases were characterized (Mackey et al., 2004). RNAi studies since have shown knockdown of cathepsin B clears infection, however knockdown of cathepsin L (brucipain) prolongs the lifespan of the infected mouse, but does not clear the infection (Abdulla et al., 2008). It has been suggested that brucipain's function is to aid the passing of the parasite through the blood-brain barrier, and initiate the late stage of the disease (Abdulla et al., 2008; Nikolskaia et al., 2006).

This gene was not included within the target regions, however because there is WGS data available for strains B and E, mutations within this gene could be observed. Within the WGS data, nine SNPs were conserved between B17 (strain E) and Z310 (strain B), two were specific to each strain. All of the SNPs had a modifying effect, except one, which was synonymous. These were all located in upstream and downstream regions of the gene.

3.4.1 Conclusion

This chapter has attempted to tie together phenomic data from different sources, using both traditional parasitology techniques and sequence data in order to try and elucidate mechanistic differences causing the difference in phenotype observed in zymodeme groups B17 and Z310. As previously discussed, the relative abundance of the short stumpy and long slender bloodstream forms is thought to be a key factor in determining the outcome of an infection. Short stumpy forms elicit a stronger host immune response due to the older stumpy forms heading towards apoptosis, and the immunogenic compounds they release into the host bloodstream/lymphatic system upon cell death. However, long slender stages are highly proliferative and a highly proliferative population, such as those observed in Z310 infections, can increase the burden on the host.

Due to the differences in parasitaemia throughout the course of infection with these strains, and the associated host symptoms, a difference in the abundance of the bloodstream forms was anticipated. B17's bloodstream forms were predominantly stumpy throughout the early stages of infection, and although in human infections, this zymodeme is considered more virulent than zymodeme group Z310, in mice, infections with B17 strains caused a more chronic infection. Conversely, the slender stages were predominant in the Z310 strain, which caused a very acute infection in mice, and caused the culling of individuals prior to schedule, suggesting a higher abundance of slender stage parasites correlates with a more virulent infection. The stage abundance observed in microscopy data was confirmed using qPCR, by monitoring PAD1 concentrations. Mutations in the PAD1-PAD2 intergenic region were ruled out as the cause of differential expression of PAD1 in these strains.

This phenotypic analysis alongside the metabolomics work shown in Chapter 4 suggests that the basis of virulence based on the presence of absence of a chancre perhaps isn't the most accurate way of determining the virulence of the strain. In human infections, following the first wave of parasitaemia, the Z310 strain is largely asymptomatic until the development of the severe late stage. Within host variation can also account for variation in the presentation of symptoms, as is often seen in regions endemic for trypanosomiasis (Bucheton et al., 2011).

Data within this chapter also illustrated how the methods used in Chapter 2 could be applied to natural infections, which typically have very low parasitaemias. Substantial enrichment was seen within these samples, however due to the original low parasitaemia, the number of parasite reads within the sample were still very low. However strain G7, which had a high parasitaemia but had a starting DNA concentration of 10ngµl^{-1} was enriched successfully, demonstrating the percentage of the total DNA the parasite represents, not the DNA concentration, determines the success of the sample. If sequencing from samples with very low parasitaemias, successful enrichment and the necessary amount of data could be achieved through deeper sequencing.

Analysis between all seven *T.b. rhodesiense* strains used showed a high degree similarity between all strains, as expected. It also suggested a relatively uniform degree of strain uniqueness, suggesting a "core rhodesiense genome". Analysis of SNPs unique to a zymodeme group showed regions with differences in zygosity between strains. SNP abundance per gene was also investigated, and showed the majority of genes had few SNPs, whereas a few genes appeared to be under greater selection and had greater than 20 SNPs per gene.

Visualising the unique SNPs showed clusters of SNPs on chromosome 8, however mapping to the DAL972 gambiense reference showed a great reduction in the number of SNPs. The most significant reduction was seen in strain B, which belongs to zymodeme group Z310, the more chronic of the zymodeme groups, suggesting a phenotype similar to that seen in *T.b. gambiense* strains through genetic exchange along chromosome 8.

Functional annotation of these unique SNPs, not restricted to chromosome 8, showed fourteen genes with high impact SNPs were conserved between all *T.b. rhodesiense* strains, implicating these were important in defining the "core" phenotype of *T.b. rhodesiense* strains. GO term analysis and functional annotation showed several pathways potentially differentially regulated between the two zymodeme groups which could be underlying causes to the phenotypic difference seen. In particular vesicle transport was highly enriched. Several GO terms suggested other potential pathways responsible, including differences in pyrmidine biosynthesis, which is essential for growth, and had mutations in genes within this pathway in strain B, which contained the most proliferative stages. Mutations in other proposed virulence factors, the

haemoglobin haptoglobin receptor, oligopeptidase B and brucipain were also seen.

Although the phenotypic differences observed in these strains cannot be explained by a definitive change, this chapter gives evidence of multiple genetic differences could contribute to this.

CHAPTER 4

Using transcriptomic and metabolomic analysis to further phenotype parasites and understand the mechanisms driving these differences in virulence

4.1 Introduction

Different kinds of phenotypic data can be used alongside higher throughput methods such as sequencing to understand how genetic factors regulate a phenotype. As mentioned in Chapter 3, strains can be phenotyped based on factors such as differences in the clinical manifestation, such as how quickly the parasite reaches its first peak of parasitaemia and whether it causes the development of a chancre at the site of bite, two factors which have been used to phenotype the strains that are being discussed in this chapter. Often these traits are used to define the virulence of the strain, in this case the strains from the zymodeme group B17 are considered more virulent because they cause a chancre around the bite site, and often patients infected with these strains present with a more acute infection which progress to the secondary stage far quicker. In comparison, strains from the Z310 zymodeme group are considered here to be less virulent because the patients manifest a more chronic infection, which is slower to develop into the second stage. Chancres are also absent in these chronic infections.

The evolution of virulence and the impact of the host-parasite relationship on the fitness of a strain and its virulence has been extensively studied (Tibayrenc, 2011; Black et al., 1983; Alizon and Lion, 2011). It is often a disadvantageous strategy for a parasite to be highly virulent because this reduces the lifespan of its host and the opportunities for the parasite to be transmitted. However differences in virulence, such are those seen in these *T.b.rhodesiense* strains have been observed in several heteroxenous parasites (Rigaud et al., 2010).

Genome sequencing is important in understanding the potential phenotype of a parasite in comparative studies, however it is a static, and does not indicate the regions of the genome that are actively expressed. Due to this, it is hard to identify genes potentially responsible for generating a phenotype from genomic data alone.

4.1.1 Aims of this chapter

In this chapter, infections with these two strains are going to be characterized using two key methods in an attempt to better under understand how transcriptional and metabolomic regulation influences disease manifestation This chapter will be making use of two high throughput methodologies, transcriptomics, and metabolomics in order to further identify phenotypic differences between the B17 and Z310 infections. This will be done through studying the metabolites from different stages of the infection, and RNAseq, and hopes to bolster the variation seen in Chapter 3 and understand the mechanisms that can give rise to the different phenotypes observed.

Chapter 3 showed that these two zymodeme groups have different phenotypes in part due to differences in the relative abundances of the bloodstream forms. By using transcriptomic data, which allows for investigation into the actively expressed regions of the genome, and metabolomics to look at the abundance of products produced as a result of active metabolic pathways, this chapter aims to further understand the regulatory differences in these strains.

Hopefully by combining the genomic data, which is useful for determining the virulence potential of a strain, transcriptomic data, which determines what transcripts are actually expressed, and metabolomic data, which shows the abundance of protein and non-protein metabolites, a more comprehensive picture of what is causing these differences in differentiation and the mechanisms involved can be elucidated. Combining analysis from both techniques will also be adding the potential to look at the host response from the upregulated metabolites, alongside the differentially regulated pathways of the parasite.

4.1.2 Metabolomics

The metabolome is transient like the transcriptome, and comprises of all metabolites present at the time of sampling. Metabolites are products of metabolism and generally less than 1.4kDa in size (Vincent and Barrett, 2015). Metabolomics is a high throughput technology that until recently hasn't been utilized to its full potential, particularly in parasitology. However with advances in NMR and HPLC, there have been studies in the metabolomes of multiple organisms and systems. Amongst others uses, metabolomics

has been used to implement new drugs and finding metabolomics profiles for diseases (Kaddurah-Daouk et al., 2014; Vincent and Barrett, 2015).

Metabolomic studies have also been carried out in protists and these have been used to develop effective drug regimens. In *Plasmodium*, metabolomics was used to better understand the function of the PfCRT protein (Fidock et al., 2000). In *T. brucei*, metabolomics led to elucidating the method of action (MOA) of eflornithine and subsequently how resistance to this develops (Vincent et al., 2010). Similarly, the MOA of miltefosine in *Leishmania* has also been partially discovered through metabolomics (Canuto et al., 2012).

Metabolomic studies look at all of the metabolites at a particular time, however metabonomic studies can also be used to study a signature change in metabolites (Vincent and Barrett, 2015). This often looks at just a targeted range of metabolites and the differences between metabolites at two different intervals. These intervals could represent different stages of an infection or a disease progression, and reproduced patterns of metabolite change can be useful in understanding a system for example in uninfected and infected hosts (Creek et al., 2012a; Vincent and Barrett, 2015).

4.1.3 Considerations in analyzing metabolomic data

One of the main difficulties in analyzing this metabolomic data is distinguishing between host and parasite metabolites because the samples were from infected hosts. Due to the nature of *T. brucei*, multiple organs within the host are affected, and so a whole variety of pathways are differentially regulated. Due to this, it is hard not only to identify the source of the metabolites, i.e. host or parasite, but also whether the metabolites are involved in the manifestation of the disease, or just a by-product of the disease. For instance, metabolites typically indicating liver damage show the parasite is virulent and has damaged this organ, however these metabolites are not important in causing the disease, they are a by-product of damage by the parasite.

Another consideration is that metabolites are identified by a combination of mass and charge, and due to this, metabolites cannot be confirmed with complete certainty (Lynn et al., 2015). In particular, this makes the identification of isomers more challenging. The metabolites identified are also within a range, and metabolites of interest may not be annotated by the analysis or be outside of this range of detection. An example in

trypanosomes is the elusive stumpy-inducing factor (SIF), which is speculated to cause the differentiation of long slender forms to short stumpy forms, and is believed to be of low molecular weight (MacGregor and Matthews, 2012). It is highly unlikely that SIF could be identified within the metabolomic data because its weight and structure are unknown, however it is believed to be a key metabolite. Although the detection of metabolites and subsequent analysis has improved in terms of the number of metabolites that can be assigned and the sensitivity, not all metabolites can be identified by one of the two current methods only (Lynn et al., 2015; Vincent and Barrett, 2015). The differences in these methods are outlined beneath, however the data in this chapter was produced using liquid chromatography and mass spectroscopy.

4.1.4 Methods of metabolite detection

4.1.4.1 Nuclear magnetic resonance (NMR) spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy was the first of the two methods developed for metabolite detection. It uses the spin properties of nuclei, most commonly ^{1}H or ^{13}C, to generate a magnetic field and identify the compounds within a sample (Dieterle et al., 2011). It is least favoured now for the analysis of complex metabolomic samples because it is far less sensitive, however unlike mass spectroscopy, it does provide you with exact and accurate quantification (Dieterle et al., 2011).

4.1.4.2 Mass spectroscopy (MS) and liquid chromatography (LC)

Mass spectroscopy was developed after NMR, however it has gained popularity due to its greater sensitivity, which is required for the resolution of complex metabolomic samples and its ability to measure down to femtomolar and attomolar quantities (Pan et al., 2007). The use of MS was limited until developments in liquid chromatography (LC), which have widened the capabilities of MS for metabolomics when using the two techniques in conjunction. Unlike NMR, sample preparation grossly effects signal intensity, which is used to determine metabolite concentration, and so samples are spiked with known concentrations of standards to aid quantification (Dettmer et al., 2007).

4.1.5 IDEOM software can be used to identify and analyse LC-MS data

IDEOM software is used for processing raw LC-MS data using XCMS and mzmatch.R tools to identify metabolites from peaks, and filtering raw data for noise (Smith et al., 2006; Scheltema et al., 2011). Metabolites are identified from their retention times and mass and compared against a metabolite database (Creek et al., 2012a). The relative abundances compared to a control and the confidence rating of the identified metabolite were both used to filter metabolites used for the analysis later in this chapter.

4.1.6 RNAseq considerations and procedure

RNA sequencing (RNAseq) allows for the sequencing of all transcripts present at the time of sampling, giving a snapshot of gene expression at a set point (Wang et al., 2009). Due to the more transient nature of the transcriptome, this means RNAseq lends itself to experiments investigating how different conditions affect expression (Wang et al., 2009). There are multiple applications for RNAseq, including looking at global changes to expression levels, or just particular pathways, alternative splicing and post-transcriptional modifications. Despite mRNA being the primary transcripts analysed, RNAseq can be tailored to study other RNA populations such as miRNA and tRNAs. However within this chapter, the analysis will be focused on mRNA analysis.

RNAseq has now superseded the prior technology for gene expression studies, microarrays (Zhao et al., 2014). Unlike microarrays, which require a good reference genome, RNAseq can be more readily applied to look at expression in a non-targeted way and is more sensitive for SNP detection. This makes RNAseq ideal for identifying rare mutations, which would go unnoticed using a microarray. Although genomic sequencing is typically used for variant detection, RNAseq can also be used in validation (Wilkerson et al., 2014). RNAseq can also be used to improve on current gene annotations, identify unannotated genes, splice variants and improve on the identification of exon boundaries (Morin et al., 2008; Wilhelm et al., 2008).

4.1.7 Library preparation considerations

Although direct RNA sequencing is available, RNAseq libraries are primarily generated using a RNA to cDNA conversion step, despite this conversion introducing bias (Ozsolak et al., 2009). Total RNA is the starting point for all RNAseq libraries, however for the majority, the high percentage of ribosomal RNA (rRNA), which is typically over 90% of the total transcripts, makes sequencing of the total RNA an unviable option. mRNA is often the RNA species of interest and so methods to select for this in the sample are used (Huang et al., 2011). The two main methods used are enrichment for the poly-adenylated transcripts (poly(A)+) and rRNA depletion (Tariq et al., 2011; Zhao et al., 2014).

4.1.8 RNAseq analysis pipeline

Although the pipeline for processing RNAseq data will vary from application to application, there is a general consensus for the stages RNAseq data needs to be processed through. The data need first to be aligned to a reference genome, and then reads per gene counted, normalised and checked for quality. Then differential gene expression analysis is done to look at the genes that are differentially expressed. This then leads onto a variety of analyses, which can look at the pathways that are involved, and the more functional effects of differences in transcription. A brief overview of the traditional RNAseq pipeline is shown in Figure 4.1.

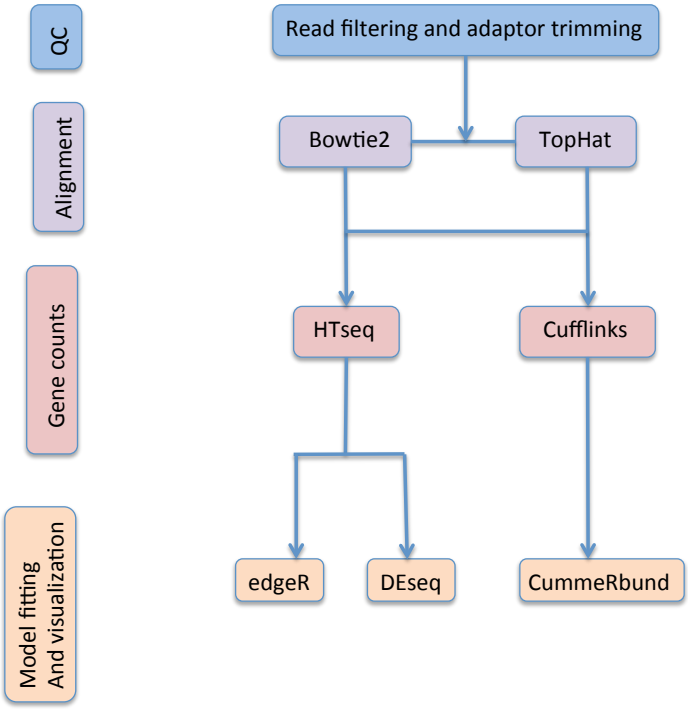

Figure 4.1: Traditional RNAseq workflow is shown. The data is aligned to a reference in either Bowtie2 or TopHat, this can either then be used with cufflinks, which generates reads per kilobase per million mapped reads counts (RPKM), or HTseq which generates counts per individual gene ID. There are other read counting pieces of software available, which will generate data in one of the two formats. Using this count data, this can be modelled using a variety of R packages, for RPKM values, CummeRbund is generally used, and for whole gene counts, edgeR or DEseq are used. These packages model the data and generate the differentially expressed gene lists.

4.1.9 RNAseq aligners

RNAseq data can be assembled into a transcriptome in a *de-novo* or genome-guided manner. As the name suggests, *de-novo* assembly doesn't require a reference and software such as Velvet exists to generate the de novo assembly (Zerbino, 2010). However this is very computationally demanding, particularly for short reads. In contrast, genome-guided assembly is easier and far less computationally demanding. The accuracy of a genome-guided assembly also far exceeds that of de novo assemblies, which are often ambiguous and sub-optimal (Trapnell et al., 2009). If a genome sequence is available, RNAseq data can be mapped onto the genomic sequence using a short read aligner. Although there undoubtedly are differences between the genome and transcriptome, RNAseq reads can be mapped to either, and various papers have aligned to the genome in *T. brucei* and found this doesn't impair the quality of the analysis (Siegel et al., 2011).

There are multiple mappers that deal with RNAseq data, however the primary ones used for RNAseq for analysis are TopHat and Bowtie2. Bowtie2 is different to Bowtie1 because Bowtie1 was intended for very short reads (up to 25-50 nucleotide reads). Bowtie2 is more suited to the length of reads generated by the HiSeq (>100bp), (Langmead and Salzberg, 2012). Bowtie2 is also different to Bowtie1 in that it supports both local and end-to-end alignment, whereas Bowtie1 only supports reads that align end-to-end. Bowtie2 also supports gapped alignments, which means that the reads can be trimmed at the ends to optimise the alignment. By default, although Bowtie2 only reports one alignment for each read, it generates multiple alignments, but only reports the best overall alignment. Unlike with BWA, Bowtie2 bases unique mapping on read quality, with a quality score of more than 20 putting the chance of the alignment being non uniquely mapping at ~1% (Langmead and Salzberg, 2012).

TopHat recognises splice junctions and splits reads across these, it takes unaligned reads, and then splits these at further and further intervals until they align (Trapnell et al., 2009). Bowtie2 was used over TopHat because trans-splicing is ubiquitous in *T. brucei*, with only one known gene known to undergo cis-splicing, another hypothesized, and short intergenic regions, so intron recognition is not an issue (Liang et al., 2003; Kolev et al., 2010). TopHat is predominantly used where splicing variants and the discovery of novel mutations around splice junctions is important. TopHat uses Bowtie as part of its alignment process (Trapnell et al., 2009).

4.1.10 Using gene expression counts to find differentially expressed genes (DEGs)

Analysis of differentially expressed genes (DEGs) between samples can be done using a variety of different software tools. In order to observe differential expression, count data related to a transcript or gene ID is required. These counts are used to normalise the data and identify differences in abundance between samples that are not due to differences in coverage depth (Anders and Huber, 2010; Anders et al., 2013). Counts have typically been derived in one of two ways, using RPKM and FPKM. RPKM which is reads per kilobase of mapped reads, is calculated as the number of reads mapped to a region divided by the transcript length split up into kilobase fragments, then divided by the total read count, once its been divided by a million. FPKM is derived similarly to RPKM and stands for the fragments per kilobase of mapped reads. Instead it measures

the number of fragments of DNA by counting either the number of reads directly in a single end library, or dividing the number of reads by two in a paired library by counting mapped pairs as one fragment. This count type is harder to normalise in situations such as paired end sequencing, whereas counts per gene are simply the number of reads per gene ID.

Due to the different counting and normalisation methods for RNAseq, different software needed for each application. Generally the workflow for RPKM/FPKM methods is mapping using Tophat, then analysis by cufflinks, which generates the RPKM/FPKM values (Trapnell et al., 2009). Following processing through a variety of cufflinks applications such as Cuffmerge and Cuffdiff , the data is then visualised using the R package cummeRbound. The main advantage of using cufflinks is that it is sensitive to detecting iso-forms in the data. However it also has a more conservative method, and compared to HTseq count, finds less genes (Anders et al., 2015).

However analysis can also be done using per gene ID counts, the most popular tool being HTseq count. As with the other analysis methods, this can then be used in conjunction with an R packages for modelling the data and identifying the DEGs. The two most commonly used are DEseq and edgeR. In this chapter I will be focusing on the use of gene count data, and analysis with edgeR.

4.1.11 Using the HTseq package to calculate gene counts

As previously mentioned, HTseq generates counts per gene rather than transcript or by RPKM/FPKM. It uses an annotation file and a BAM file, and counts the number of reads that overlap with exon co-ordinates. Unlike other tools it can be used to discriminate against reads that overlap two genes and would contribute to the frequency count of more than one gene (Anders et al., 2013). It also discards genes mapping to multiple locations, which would also affect differential expression analysis. Data generated from this package can then be used with R packages DEseq, DEseq2 and edgeR (Anders and Huber, 2010; Robinson et al., 2010; Love et al., 2014).

4.1.12 Identification of differentially expressed genes (DEGs) and visualisation

Differentially expressed genes (DEGs) are identified following the fit of data to a model. In both DEseq and edgeR, a negative binomial model is used to best fit the degree of variation seen in RNAseq data. Once identified, these packages can then be used to visualise the DEGs. As mentioned previously this visualization can be done either in cummeRbund for RPKM/FPKM counts or using edgeR and DEseq for gene counts. These edgeR R packages are explained beneath.

4.1.12.1 edgeR

edgeR which stands for empirical analysis of differential gene expression in R, was originally used for serial analysis of gene expression data (SAGE), but has now been widely adopted for use in RNAseq analysis (Robinson et al., 2010). edgeR models data using an overdispensed poisson model and controls the degree of overdispersion between genes using an empirical Bayes method. This accounts for the total size of the library and the abundance of a gene.

4.1.12.2 DEseq

Similarly to edgeR, DEseq models the data based on a negative binomial model. However DEseq is considered to have a very conservative method of DEG identification and so the true positive rate (TPR) observed is much lower than seen in edgeR. This worsens as more outliers, such as very highly and lowly expressed gene counts, are introduced to the analysis (Soneson and Delorenzi, 2013). The main difference between edgeR and DEseq is how dispersion within the model is dealt with. In edgeR this is done using an individual dispersion parameter to fit the dataset, whereas DEseq uses a more flexible approach to deal with the variability seen in low count values (Anders and Huber, 2010).

4.1.12.3 CummeRbund

CummeRbund is also an R package, and this uses the RPKM/FPKM values produced by cuffdiff which is part of the cufflinks package. It uses these cuffdiff values and calculates the relationship between different genes/transcripts/genome regions by creating a SQLlite database (Trapnell et al., 2012).

4.2 Methods

4.2.1 Infection procedure and sample collection

Two female A/J mice were infected intraperitoneally with 10^4 parasites from a blood stabilate of B17 and two female A/J mice were infected with Z310. This mouse strain is particularly susceptible to trypanosome infection, which is important in order to obtain a high parasitaemia blood sample with which to infect subsequent mice. They were subsequently humanely sacrificed following a positive parasitaemia, as identified from daily microscopy screening.

Blood was then collected from these mice and 10^4 parasites were passaged into C57BL/6 mice, five for each isolate. Mice were bled prior to infection and 25µl of blood was then taken twice weekly for metabolomic analysis, 10µl of blood was also taken every other day for qPCR analysis, and daily spots were used for recording the presence of the trypanosomes, and used to create thin films for staining purposes. The mice used for metabolomic and QPCR analysis (Chapter3) were culled prior to schedule due to ill health; subsequent mice used for the collection of infected blood for RNAseq were culled following the first peak of parasitaemia. The blood collected for RNAseq analysis was obtained from another set of infections. The infections for collecting blood for RNAseq were done as above, however mice were only bled for checking parasitaemia and culled at the first peak of parasitaemia.

4.2.2 Procedures for metabolomic sample collection and analysis

4.2.2.1 Collection and processing of samples for metabolomic analysis

Samples were collected for metabolomic analysis by the University of Glasgow and prepared as follows. Microtubes were prepared containing 2µl of 22.5mmol EDTA, to prevent coagulation of collected blood. Up to 30µl of mouse blood was collected by tail vein bleed into the microtubes containing EDTA. However average tail bleed volumes were often lower than this, particularly later in infection. Samples were held on ice and then centrifuged at 6000g for 30seconds to pellet cells and separate plasma. 10µl of the plasma was then added to 40µl of cold acetonitrile (ACN), vortexed for 3 seconds and

then centrifuged to precipitate protein/salts at top speed 13000g for 3 minutes. From this approximately 45µl of particle free supernatant were collected and sent to the University of Glasgow for analysis by LC-MS (Creek et al., 2012b). Acetonitrile (ACN) was prepared with non-infected mouse blood as a control. Due to ill health, the mice were culled earlier than scheduled and so the mice were only sampled prior to infection and at days 3 and 8 post infection. This is described in more detail in Chapter 3's sample collection methods. Days three and eight correlated to approximately early and late stage of infection.

4.2.2.2 Metabolite identification and analysis

This data was produced by LC-MS and analyzed using IDEOM metabolite software (Creek et al., 2012b). From these samples, 579 metabolites were represented in the confidently assigned peaks. In order to draw any meaningful conclusions from this data, a small subset of metabolites were selected based on a number of parameters. Analysis was done in two ways, firstly by looking at the data on an individual metabolite basis, and secondly by focusing on the differential regulation of pathways between the samples. Due to the nature of MS, the metabolites are identified by their weight, and so distinguishing isomeric forms or metabolites with near identical masses is difficult. To reduce the potential effect of this, only metabolites that were either known or had a greater than 7 certainty were selected, with 10 representing a known compound. For individual metabolite analysis, metabolites were then sorted by intensity, selecting only those with an intensity in at least one time point significantly different to the pre-infection intensity. The most abundant were primarily targeted. In subsequent pathway analysis, the data was filtered as above, and then sorted by pathway.

In mice infected with the B17 strain, many of the metabolites were highly significant (P value <0.05) compared to pre-infection and differed greatly in intensity between days post infection. For Z310, this was not the case, many of the metabolites were not significantly different from the control, and very few differed greatly in intensity. As a result of the different profiles for the B17 and Z310 infections, different metabolites of "interest" were selected for each strain, but the relative intensity levels in the other strain are given for comparison. Metabolites with no known role in a pathway or function were also excluded, as it is hard to understand the significance of these metabolites without functional information.

One of the main barriers to analyzing this data is that many of the changes to the metabolites can be attributed to deterioration in the health of the mouse. For instance, metabolites indicating damage to the liver are a result of the damage caused by the infection, they are not indicative of a metabolite that causes the parasite to be virulent even though a high intensity of metabolites associated with damage to the liver is evidence that the parasite is more virulent. Issues arise in trying to distinguish between mouse and trypanosome metabolites since the mouse metabolome is better characterized than that of trypanosomes.

4.2.2.3 Using metaboanalyst for post IDEOM metabolomic analysis

All the metabolites identified were compared using their peak intensities using metaboanalyst software (Xia et al., 2015). Individual rather than m ean values were used to construct heatmaps showing differences between individuals from the same stage of infection, and between the three sampled periods. Heatmaps were constructed for each zymodeme group, to observe differences between the stages in infection for one strain, and then compared across both zymodeme groups.

Metabolites with an intensity outside of the inter-quartile range were filtered out. PCA plots were used to see the degree of variability between strains and infection stages. Dendrograms were also constructed to see whether the metabolite profiles clustered by day post infection as expected. This was done using a Euclidean distance measure and by clustering using Ward's linkage method. Correlations between the treatments were also plotted using a Pearson r measure. Heatmaps were also calculated by clustering by ward and using a Euclidean distance measure.

4.2.3 Procedures for sample collection, processing and analysis for RNAseq

RNAseq was done to tie together differences seen in the genomic data and the phenotypic data (microscopy/QPCR and metabolomics) to allow for a greater understanding of the mechanisms driving these differences in phenotype between these two sets of strains. RNAseq allows us to observe differences at the transcriptional level, whereas metabolomics looks at differences in metabolic pathway products, which can reflect differences at a translation level. The number of genes

outnumbers the number of transcripts, just as the number of transcripts outnumbers the metabolites, and so an effect seen at a genomic or transcriptomic level can often be more pronounced in the metabolomic data.

4.2.3.1 Sample processing for RNAseq post collection

Samples were collected from mice infected as described in section 4.2.1. Blood from these mice was collected via cardiac puncture into syringes containing 0.1ml 20mM EDTA to prevent coagulation. The whole blood was then was then spun at 6000g for 2 minutes, the plasma removed and resuspended in an equal volume of phosphate buffered saline (PBS). Nucleic acid purification lysis buffer (ABI 4305895) that had been already diluted 1:1 with PBS was then added 1:1 to the resuspended blood, mixed then left on ice. Samples were assigned IDs, which relate to the strain they were infected with and these IDs will be used in subsequent analysis, and are shown in Table 4.1.

Table 4.1: Mice were assigned sample IDs based on whether they were infected with strain B, from the Z310 zymodeme group, or strain E, from the B17 zymodeme group, with subsequent letters in the ID referring to cage and batch references used. These IDs are used in subsequent analysis.

Zymodeme group	Sample ID
Z310	
	B1_1_3
	B1_3_2_7
	B4_1_2
	B5_1_1
	B1_4_2_6
B17	
	E1_1_2_5
	E4_1_4
	E3_1_5
	E1_1_7
	E2_1_6

4.2.3.2 RNA extraction

RNA was extracted from blood collected as previously described and processed prior to extraction as above. RNAseq requires higher quality RNA than other downstream processes, and so the Purelink™ RNA mini kit (Life technologies 12183018A) was used in place of the micro kit, because this could be adapted to isolation from blood more

easily and RNA extracted using this kit was less degraded based on RNA Integrity Number values (RIN).

Five different mice were used for each strain, B and E, which represent zymodeme groups Z310 and B17, and samples were taken pre-infection as controls. RNA was extracted using the purifying RNA from whole blood protocol. For optimal yield/quality, 100µl of processed blood was used in the first step and resuspended in 50µl nuclease free H_2O in the last step. Sample volumes greater than this led to insufficient cell lysis and clogging of the column. Lower volumes did not improve quality but greatly reduced yield. On column Purelink™ DNase treatment was performed. Crude extracts were assessed for quality on the nanodrop using their 260/280 and 260/230 scores, and quantity using the qubit RNA assay (Fisher Scientific, UK; Invitrogen, UK). Overall quality, any degradation and fragment size was assessed using the bioanalyzer total RNA pico chip (Agilent, 5067-1513).

4.2.3.3 Purification of RNA

Crude extracts were purified using RNAclean XP beads (Beckman coulter, A63987) which contained Riboguard™ to inhibit RNases. RNA was purified using beads instead of column purification to remove highly degraded fragments less than 200 nucletotides long. The crude extract was added to 1.8 x the volume in beads, mixed by pipetting 10 times, then left to bind to the beads for 5 minutes. The samples were then inserted into a magnetic stand until the solution cleared (approximately 2 minutes). The clear supernatant was then removed from the samples, and 500µl of 70% ethanol was added to clean the beads with the RNA bound. After 30 seconds the ethanol was carefully removed to prevent disruption to the beads, and the ethanol wash repeated. After the second ethanol wash, the ethanol was removed, and the beads left to air dry. To concentrate the sample for library preparation, the sample was then resuspended in 10µl of RNase free water. This process not only removes small fragments but removes salts that can contaminate the sample left over from the column extraction method and can later interfere with downstream processing.

Following purification, the samples were then QC'd using the bioanalyzer to determine RIN values and look at overall quality, nanodrop for 260/280 and 260/230 values and the qubit RNA assay for determining the volumes needed for rRNA depletion (Fisher Scientific, 2008; Invitrogen, UK; Agilent, 5067-1513).

4.2.3.4 rRNA depletion

Due to the samples consisting of both host and parasite, it was essential to deplete the sample of rRNA, which typically comprises of ~90% of total RNA, in order to prevent a further loss of usable reads. This was done using the Scriptseq™ complete gold kit for eukaryotes- low input (Cambio SCL24G), which comprises of the Ribo-Zero™ gold magnetic kit and Scriptseq v2 Library preparation kit. The gold Ribo-zero™ kit (low input) was used as it removes not only eukaryotic cytoplasmic rRNA, but also mitochondrial rRNA from total RNA samples of 100ng- 1µg. The protocol was followed as per manufacturer's instructions however 0.5µl of RiboGuard™ RNase inhibitor was also added to the resuspended magnetic beads (Epicentre, RG90925).

All of the samples had a total RNA input of greater than 250ng apart from sample B2.3, and so all other sample volumes were adjusted to 14µl with RNase free water and had 4µl of Ribo-Zero rRNA removal solution added during stage 3.B of the protocol. Sample B2.3 had its volume adjusted to 16µl and 2µl Ribo-zero rRNA removal solution added. The rRNA depleted sample was subsequently purified using RNAclean™ XP beads, as described in the protocol. Successful rRNA depletion was observed by bioanalyzer Eukaryote Total RNA Pico chip (Agilent, 5067-1513).

4.2.3.5 Library preparation

rRNA depleted samples were then used to prepare libraries for Illumina sequencing using the Scriptseq™ v2 RNA library preparation protocol. The cDNA was purified using Agencourt AMPure™ beads and Illumina barcodes were added in place of the reverse PCR primer in stage 5.E of the protocol (Beckman coulter, A63880). A final purification step was done on the amplified library using AMPure™ XP beads, and the libraries were quantified using both the qubit HS assay and bioanalyzer HS chip for pooling (Fisher, Q32851; Agilent, 5067-4626). The pool was then submitted to the University of Liverpool's Centre for Genomic Research for sequencing. The samples were sequenced on the Hiseq, producing 2 x 100bp reads (Illumina, inc).

4.2.3.6 RNAseq bioinformatic analysis

4.2.3.6.1 Alignment of data using bowtie2

Prior to mapping, the adaptors were trimmed using Cutadapt version 1.2.1 using option –O 3, which trims the 3' ends, which match the index sequence (Martin, 2011). Sickle was subsequently used to trim reads with a quality score of less than 20 and reads shorter than 10bp after trimming were removed. If only one read in a read pair passed the filter, these were rejected and not used in the alignment. Bowtie version 2.1.0 was used using default settings and Tb927 version 8.1 reference was obtained from www.tritrypdb.org and used to build the Bowtie index.

4.2.3.6.2 HTseq-count was used to generate gene counts

Gene expression was calculated using HTseq-count from the HTseq package (Anders et al., 2015). For this, parameters –m union, -f bam, -r name were used. Union mode allows reads to be assigned to a feature even if the full length of the read doesn't map to that feature. Reads mapping to more than one feature are labeled ambiguous and not included within the count. –f was used to define BAM as the input file type, and –r is especially important for paired end data in order to tell htseq-count where to expect both read pair alignments within the file (Anders et al., 2015).

4.2.3.6.3 edgeR was used to fit the data to a negative binomial model

Non-expressed genes were removed from analysis and determined as genes with less than 3 counts (reads aligned) across all 10 samples. These samples were then plotted on a multidimensional scaling (MDS) plot to look at the overall variability and whether the two strains could be separated according to their patterns in gene expression. They were fitted using the glmFit function and normalized for differences in read depth by calculating the dispersion parameter, which estimates the biological variation (BCV) seen between samples. This BCV value was used to generate the BCV plot seen in the results section. Genes upregulated in either B17 or Z310 were then calculated, with a false discovery rate of 5% as cut-off.

4.2.3.6.4 KEGG analysis was used to look differentially regulated pathways

KEGG groups were assigned using the metabolic pathway enrichment tools currently available at www.tritrypdb.org. DEGs were assigned to KEGG groups using the Tb927 database from KEGG with a P-value cutoff of 0.05. Subsequent Bonferroni adjustment was applied to those with assigned groups, and only those with a significant Bonferroni adjusted P-value were used to generate nine KEGG groups.

4.2.3.6.5 REVIGO was used to assign functional groups to the DEGs identified

REVIGO was used with the parameters described in Chapter 3. For each strain, all of the DEGs were subject to analysis, not just the highly differentially expressed genes. This was to see whether there was a significant relationship between differential expression and particular functional groups of genes. Significant was assigned based on a Bonferroni corrected P-value.

4.3 Results and discussion

4.3.1 Metabolomic analysis

4.3.1.1 Pathway analysis shows more global upregulation of pathways in B17 infections

Of the 249 metabolites left after filtering as previously mentioned, 184 (74%) of these showed more than a five fold increase in concentration in the B17 infected mice by day three post infection, and 153 (61%) by day eight post infection. In contrast, only 28 (11%) of these 249 metabolites were increased more than five fold by day three, and 22 (9%) by day eight in the Z310 isolate. The pathways that these correspond to are shown in this analysis.

Figure 4.2 shows the averaged metabolite intensity for each pathway at day three and eight post infection, normalized to a pre-infection control. As you would expect with Z310, it begins its first peak of parasitaemia by day three, and you can see a much lower concentration of the metabolites, compared to that seen in the B17 isolate (as shown by purple bands compared to blue). This higher abundance of metabolites is seen over all of the pathways the metabolites were assigned to. The only exceptions to this are in the

secondary metabolite and polyketide and non ribosomal peptide biosynthesis pathways.

Currently there is little information on polyketide and non-ribosomal peptide biosynthesis in trypanosomes, however the function of these in other eukaryotes is often related to the production of toxic compounds or modulation of the immune system (Taylor, 2008).

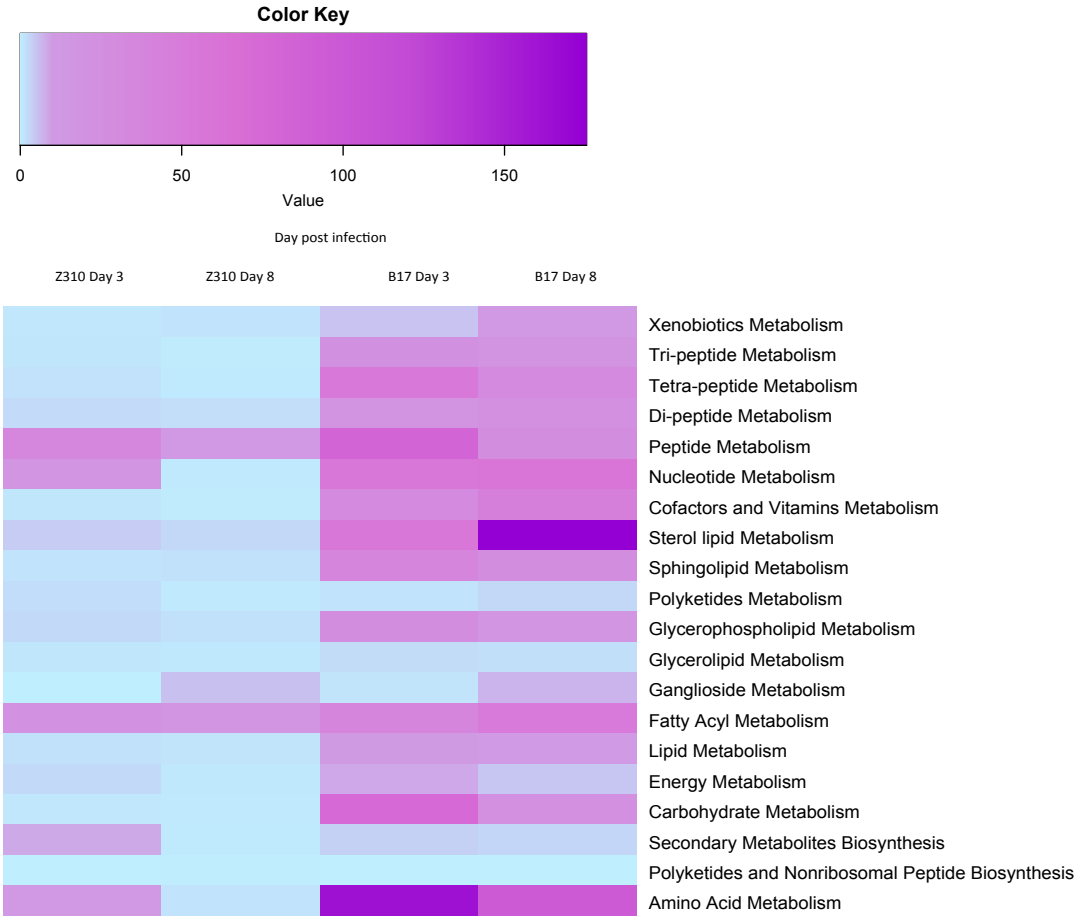

Figure 4.2: Heatmap showing the metabolite intensity for each pathway. Values are averaged across pathways and rows represent the averaged intensity values at day three and eight post infection normalised to a pre-infection control. The key shows the fold increase in concentration relative to the pre-infection control, with blue representing a low increase, and purple a higher increase, with the highest increase in sterol lipid metabolism, with an increase 175 times relative to the control. The pathways shown are as follows left to right: Xenobiotics biodegradation and Metabolism, tri,tetra and di peptide and peptide Metabolism,, Nucleotide Metabolism, Metabolism of Cofactors and Vitamins, Sterol lipid Metabolism, Sphingolipid Metabolism, Polyketide Metabolism, Glycerophospholipid Metabolism, Glycerolipid Metabolism, Ganglioside Metabolism, Fatty Acyl Metabolism, Lipid Metabolism, Energy Metabolism, Carbohydrate Metabolism, Biosynthesis of Secondary Metabolites, Biosynthesis of Polyketides and Nonribosomal Peptides and Amino Acid Metabolism

Spermidine abundance is much higher in the B17 isolate compared to that in the Z310 isolate (shown in amino acid metabolism). By day three, it is 99 times higher than the pre-infection concentration in the B17 strain, which sharply declines to nearly pre-infection levels by day eight, whereas in the Z310 isolates, this doesn't increase from the pre-infection concentration. This metabolite is required for the formation of trypanothione; which is required to reduce oxidative stress on the parasites during infection (Konwar et al., 2013). *T. brucei* is capable of scavenging for spermidine but the bloodstream concentrations of spermidine are very low, insufficient to sustain blood stream forms (BSF), even at low parasitaemias. Due to this, they synthesize their own spermidine, which is then converted to trypanothione for growth (Taylor, 2008). Spermidine's essential role in growth has led to research into designing therapies to prevent the action of spermidine synthase, which is required for the production of spermidine and is downstream of the action of ornithine decarboxylase (ODC) (Taylor, 2008). Difluromethylornithine (DFMO) is a chemotherapy agent, which inhibits the action of ODC and is currently used for the late stage treatment of *T. brucei* (Taylor, 2008). Due to the disparity in growth between B17 and Z310 infections, differential spermidine regulation in these strains may by affecting parasite growth and resulting in these phenotypes.

Z310 infections contain predominantly highly proliferative parasites, and so this difference may also be caused by a higher consumption rate of spermidine. B17 infections will have a lower spermidine consumption rate because the majority of parasites in this infection are cell cycle arrested, and so this could lead to the comparatively high spermidine concentrations. B17 infections also enter the first peak of parasitaemia later than Z310, and so the subsequent low levels of spermidine at day eight will be a reflection of the increase in parasitaemia and consequently spermidine consumption.

4.3.1.2 Establishing a metabolic signature of infection

One of the frequent uses of metabolomics is to identify the global metabolite changes associated with a disease state, and to establish the "signature" changes. Metabolite samples were taken from five mice for the Z310 strain and five for the B17 strain, prior to infection, at days three and eight post infection, using the procedures described in section 4.2.3.1. Post infection measurements correspond to the early and late stages of

infection in days three and eight respectively. The metabolomic analysis for the Z310 strains is shown in Figures 4.3-4, and for the B17 strains in Figures 4.5-4.6. Infections from both strains are shown together on Figures 4.7-4.8. This is done in order to observe more clearly the marked charges in the metabolome in each strain in the early and late stages of the disease.

For Figures 4.3-4.8 the metabolite intensity scores are shown in all of the metabolites identified from these samples. The intensity scores for each individual are shown, not the mean, in order to see the reproducibility of these metabolite changes across individuals. The individuals were clustered according to their associated metabolite intensities, and as shown in Figure 4.3-4.8 the three states investigated, pre-infection, early infection/first peak of parasitaemia, and later infection cluster, suggesting key metabolite changes seen at each of these stages, and consistently between individuals.

Overviews of this "metabolic signature" are shown in three ways, firstly through a heatmap showing the differences in intensities for each stage, which can be used to observe whether pathways are generally being up or down regulated, secondly by correlating the differences between individual and stage, to see the degree of difference, and thirdly by simplifying this variance and seeing how these stages relate using a PCA plot. This is shown for each zymodeme group and then with both for comparison. Figures 4.3-4.4 relate to Z310 infections, Figures 4.5-4.6 to B17 infections, and both are shown in Figures 4.7-4.8.

4.3.1.3 Metabolite profile in infections in Z310

The metabolite profile for the five mice infected with the Z310 strain is shown in Figure 4.3, in which samples taken pre-infection, at day three post infection and day eight post infection, and are shown in green, red and blue respectively. Metabolite intensities are shown in the heatmap in Fig 4.3, and blue correlates with a low intensity of the metabolite, red with a high intensity. Fig 4.3 shows that in the samples taken prior to infection, the majority of the metabolites identified were found at low levels, with the exception of a few metabolites, which were more abundant prior to infection. These are most likely to be the host metabolites that are diminished due to the burden of the parasite. In Z310, the infection predominantly consists of highly proliferative forms, and so depletion of host reserves of resources necessary to facilitate growth such as iron and other resources such as pyrimidines, is expected.

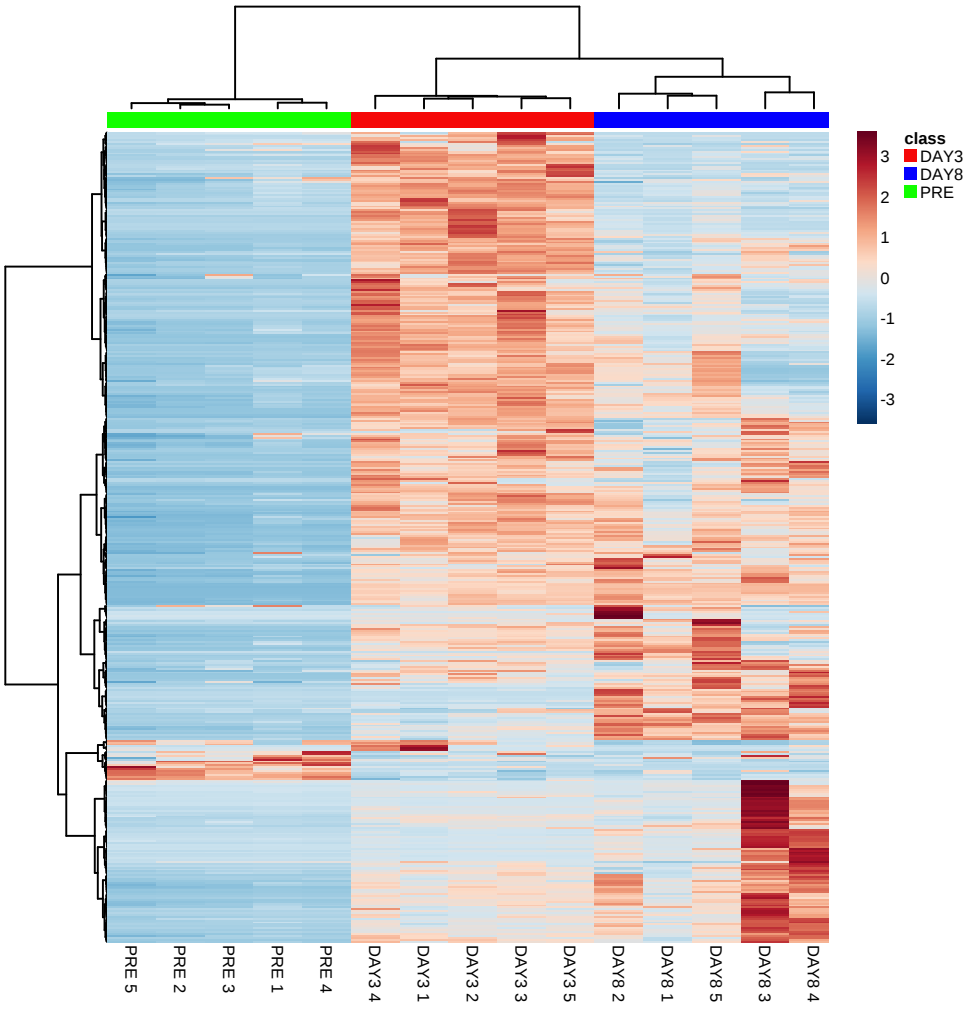

Figure 4.3: Heatmap showing the intensities of metabolites prior to infection, at the first stage of infection at day three, and later in infection, at day eight. Prefixes day three, day eight and pre refer to the stage the metabolite was collected at, and the subsequent number refers to the individual. These cluster by stage and are represented in green, red and blue for pre-infection, day three and day eight respectively. The clustering per stage indicates both that there is consistency between individuals at each stage, and that the stages have distinct metabolic profiles. Blue represents a low metabolite intensity, red high metabolite intensity.

In the Z310 strain, as mentioned in Chapter 3, the parasitaemia peaked earlier than seen in the B17 strain, and so the high abundance of the majority of metabolites taken at day three is not surprising. This general increase in abundance at the peak of parasitaemia is seen across all individuals. Following the first peak of parasitaemia, approximately half of the metabolites elevated during the peak returned to much lower levels by day eight. Metabolites with lower levels compared to pre-infection stayed at low levels, supporting the idea that these are depleted host metabolites.

However the metabolites that were at low abundance by day three, appear to be in much higher abundance by day eight. This metabolite accumulation could be as a result of parasite byproducts generated by the predominantly highly proliferative parasite population observed in Z310 infections. It could also be due to the release of metabolites following short stumpy cell death due to the high parasitaemia, or it could represent metabolites forming the host response. However as observed in Chapter 3, this strain has a smaller proportion of parasites differentiating into short stumpy stages, and so if these metabolites correlated with differentiation in the short stumpy stages, we would expect this effect to be more pronounced in strain B17.

The pattern of metabolite abundance is distinct between the early and late stages of disease, however the later stage of the disease gives rise to a non-uniform metabolic response in the individuals compared to the earlier stage. Although the overall metabolic response appears similar in the later stage, there is a greater degree of variation between the individuals. After sample collection at day eight, further samples could not be collected due to the deterioration in health and moribund condition of some of the infected individuals. At this stage, trypanosomiasis can cause multiple organ failure and systemic effects and due to the complexity of host response, it is unsurprising that the metabolic profile of each host deviates at this point.

This variance in metabolite abundance for each individual at each of these three stages was used to cluster each sample in Figure 4.4. This again supports the effects seen in Figure 4.3, in which the metabolomic profile is distinct pre and post infection. The area shaded surrounding each point shows the 95% confidence intervals (CI) for each group. Despite distinct clusters for each stage, the samples taken at day three do overlap with the samples at day eight. The higher degree of difference between individuals at day eight is evident from the PCA plot, as they do not cluster as tightly as the other two conditions.

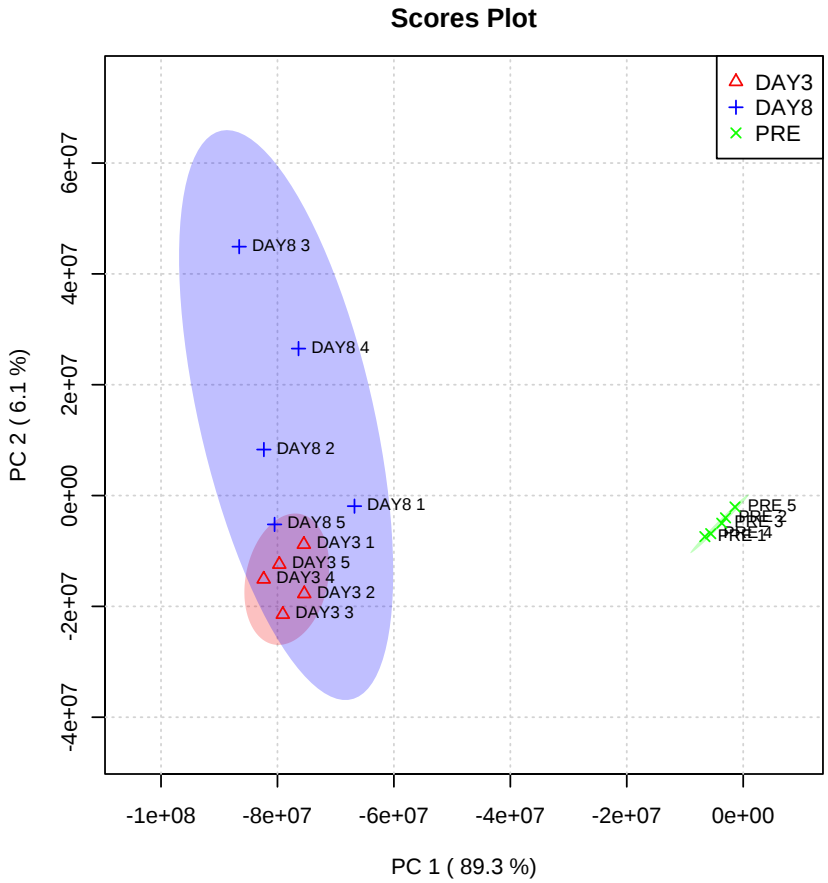

Figure 4.4: PCA plot showing the variance between samples, with pre-infection samples clustering separately to the infected samples. Day three and day eight can be separated, however are closely clustered. Due to ill health of the host, the metabolomic profile in the late stage samples is more variable and clusters less than is seen either prior to infection or in day three. Circles surrounding these values represent the 95% confidence intervals (CI).

4.3.1.4 Metabolite profile in infections in B17

The metabolomic profiles for mice infected with the B17 strain are shown in Fig 4.5. As with the Z310 infection data, the samples were clustered using metabolite abundances, and this again showed metabolic patterns unique to each stage and replicated between individuals. As with the Z310 infections, at day three the overall metabolite abundances were higher than seen pre infection or at day eight. However the difference in metabolite abundance at day three is significantly higher than observed in the Z310 infections. However there is also a high abundance of metabolites at day eight, which cause day three infection samples to cluster with day eight samples.

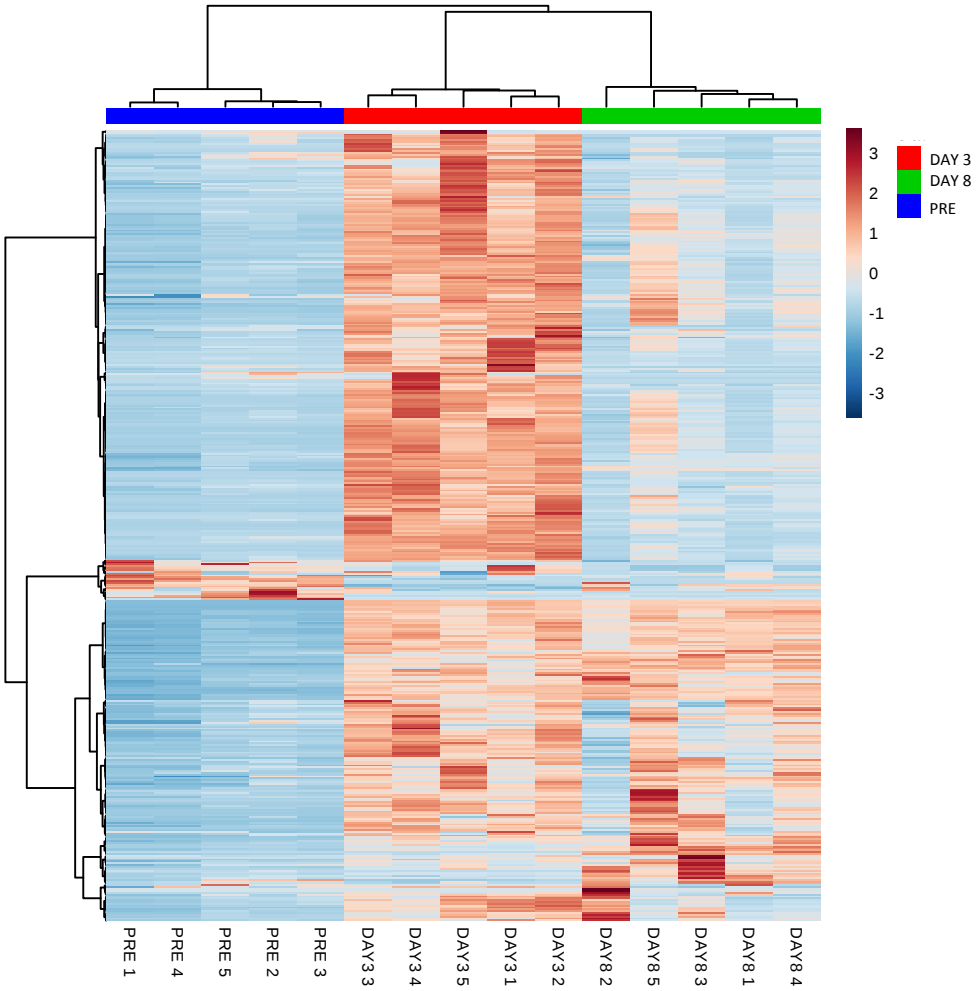

Figure 4.5: Heatmap showing the intensities of metabolites at the three stages previously described in Figure 4.3. Pre-infection is represented by blue, day three by red and day eight by green. As was observed in the Z310 infections, consistency was seen both between individuals at a particular stage, and the stages were distinct from each other. Due to the high abundance of metabolites at days three and eight, these cluster separately from pre-infection metabolites.

As before, there are a small number of metabolites that are more abundant prior to infection, and are depleted by day three of infection. In day three almost all of the identified metabolites have a high abundance, and over half of these have an abundance comparable to pre-infection by day eight. There are differences between individuals as expected, however the variation in the abundance of individual metabolites is greater at day eight post infection than day three. As previously mentioned, this could in part be due to differences in the host response. However the significant decrease in abundance of a high proportion of the metabolites at day eight is likely to be associated with the high abundance of short stumpy forms in these infections. Although more immunogenic, these stages of the life cycle have entered into cell cycle arrest, and so multiple metabolic processes are suspended. This would explain the general low

abundance of metabolites at this later stage of infection. Due to the immunogenicity of the short stumpy parasites, this suggests that the metabolites that are highly abundant are either components of the host response, or are metabolites indicative of damage caused by the parasite.

In Fig 4.6, the highest degree of variability in the abundance of individual metabolites is seen in samples taken eight days post infection. As before, the 95% CI are shown. Pre-infection and both post-infection stages cluster separately, with both post-infection stages clustering closely. Both pre-infection and day three stages exhibited similar degrees of variation between individuals.

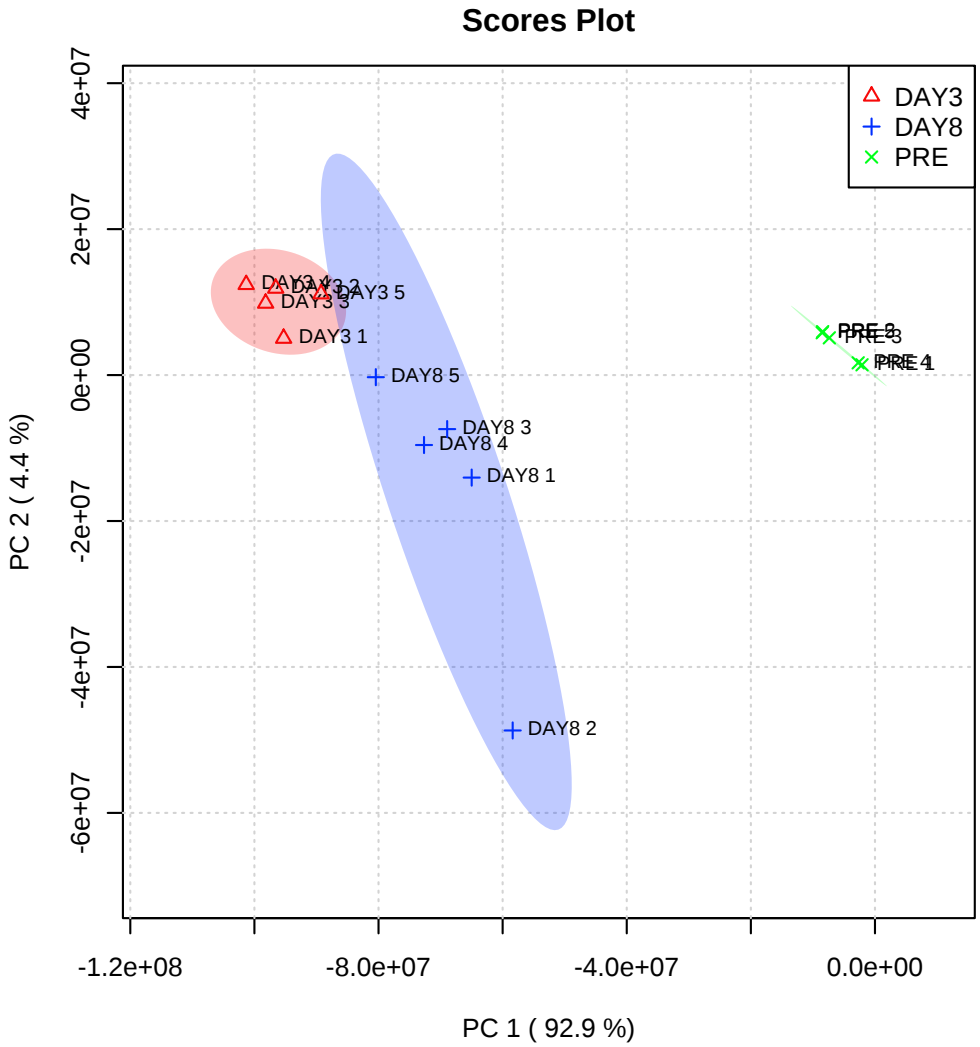

Figure 4.6: PCA plot showing that pre-infection does cluster separately from the infected samples, and as previously seen, the day eight metabolomic profile is more variable than seen earlier in infection

4.3.1.5 Comparison of metabolites between zymodeme groups

Figure 4.6 shows both the metabolic profile for infections with Z310 and B17 strains alongside each other for comparison. As before, the individuals cluster per date relative to infection, but also by strain. One of the pre-infection B17 mice clusters amongst the Z310 mice, however because they are all uninfected at this stage and from the same breeding background, this is of no biological consequence.

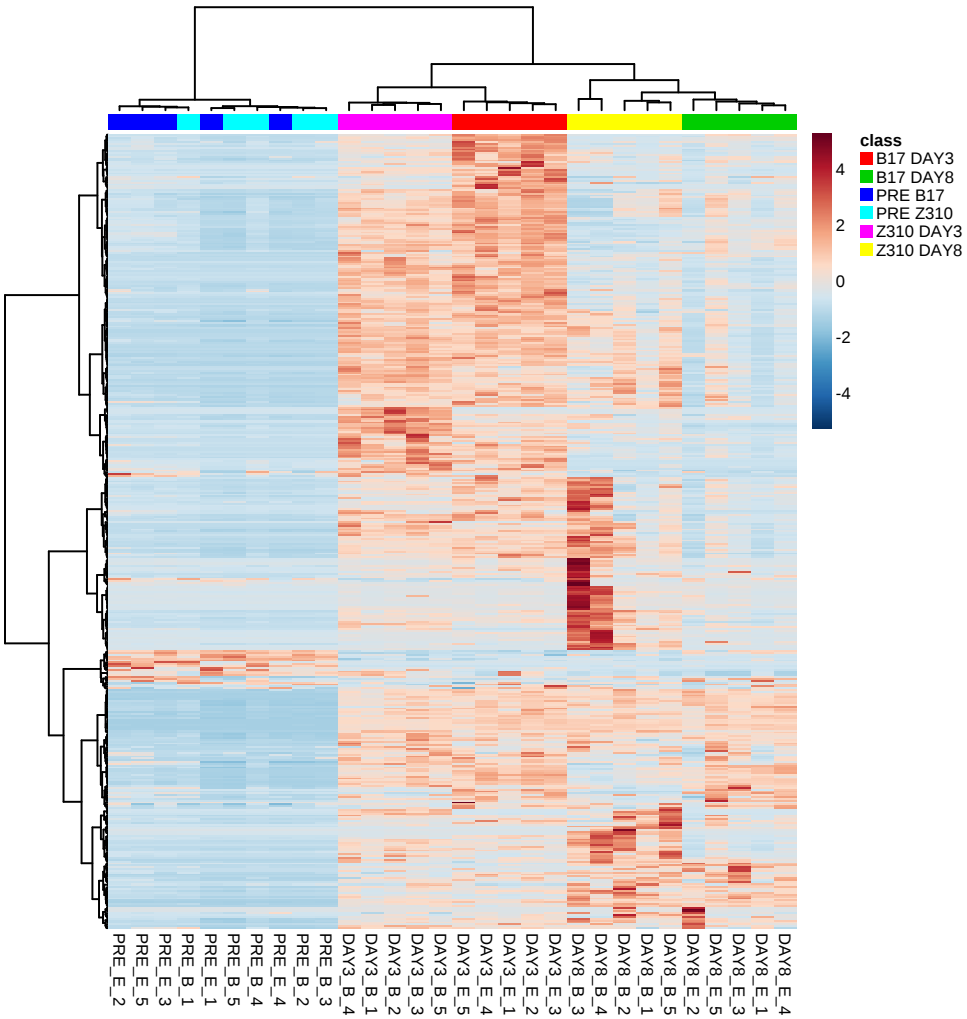

Figure 4.7: This heatmap demonstrates that for both strains, the samples cluster by both strain and by stage of infection. In both strains, there are a few presumably host metabolites that are a higher abundance prior to infection, and are depleted once infected. This also shows a more global abundance of metabolites in B17 infections by day three, compared to Z310, however this is diminished by day eight, with much lower abundances seen in B17 than in Z310.

As previously mentioned, both strains have a small number of metabolites of a higher abundance prior to infection, which are most likely to be host metabolites which are diminished by the increasing burden of infection. In both infections, the highest abundance of metabolites across the majority of the metabolites identified is observed at day three. As described in Chapter 3, strain Z310 goes into the first peak of parasitaemia first, however at its peak at day three, the abundance across multiple metabolites is much lower than observed in the B17 strain, which did not reach the first peak of parasitaemia until day five. Despite not being at its initial peak by day three, there was a general high abundance of metabolites in B17, however this dissipated by day eight, with approximately two-thirds of metabolites back to pre-infection levels. However in both strains, about a third of the metabolites maintained elevated levels into day eight post infection, and these are shown towards the bottom of the heatmap in Figure 4.7. It is hard to determine whether these are related to the host response, because the host response will be different to both of these strains due to the high degree of difference in disease manifestation. They could also be trypanosome-derived metabolites, however the lack of contrast between these in both strains indicates that they are not controlling factors of virulence.

The metabolites in the top third of the heatmap are more abundant at day three in both strains, and decrease in abundance significantly by day eight. Regions with the highest differences in abundance across strains in Figure 4.7 are the most likely candidates for metabolites correlating with differences in phenotype.

Figure 4.8 shows the degree of variability between strains at each stage. As anticipated and shown in Figure 4.7, there is very little variability in the mice prior to infection. For both strains, the metabolomic profile of both strains changes considerably from the pre-infection profile by day three. There is also the highest degree of variability at day eight in both strains, however the degree of variability in mice infected with the Z310 strain far exceeds the variability observed in B17 infected mice, however their 95% CIs do overlap.

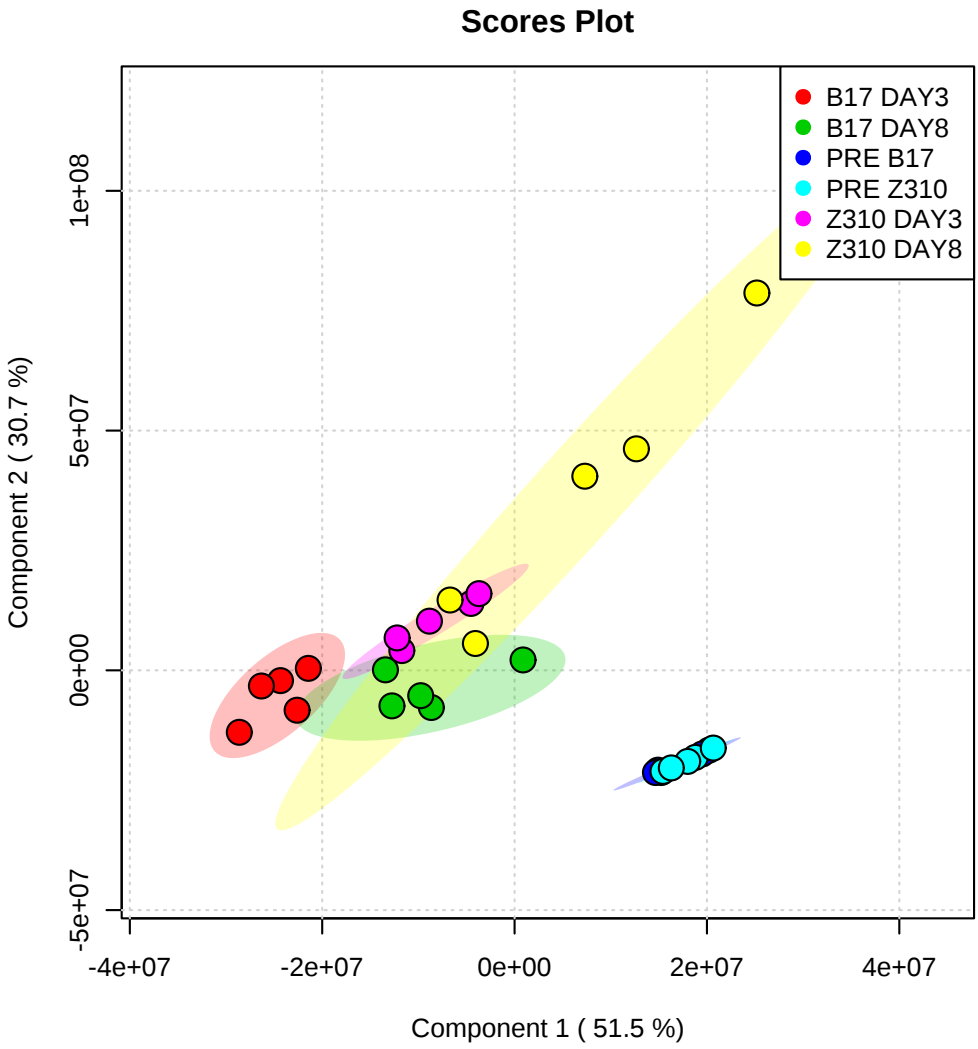

Figure 4.8: This demonstrates that the post and pre-infection metabolomic profile is different for both strains. It also shows that Z310 has a much more variable late stage, and although there is overlap between the two strains and these infected stages, they can be clustered to individual strain and stage

For both strains the day three and day eight samples do cluster separately, but for both strains the CIs of these stages do overlap. Interestingly the variability in B17 infected mice is higher by day three, but this does not increase significantly by day eight. In contrast, the variability in metabolite abundance is very limited by day three of the Z310 infections, but increases drastically by day eight. At day three in the Z310 infections, the parasite is entering its first peak of parasitaemia, however by day eight, the parasite population is still largely comprised of long slender forms. The burden of this and the accumulation of trypanosome specific metabolites as a result of a highly

proliferative parasite population could account for the variability seen in the later stage of infection. In the B17 infections, by day three the parasites had not yet reached the first peak of parasitaemia, and short stumpy forms were predominant in this infection. Due to the cell cycle arrest, and the far fewer highly proliferative forms, so fewer trypanosome specific metabolites produced, it is understandable to see less variation between these two stages.

4.3.1.6 Individual metabolite analysis

The above analysis was used in conjunction with the selection process described in the methods in order to select metabolites of potential importance/interest in either host response to infection or trypanosome specific metabolites important in eliciting these host responses.

4.3.1.6.1 Metabolites of interest in Z310 infections

In mice infected with the Z310 strain, fewer metabolites were statistically different from the pre-infected state, and the intensity of statistically different metabolites, even the most intense metabolites, were much lower than those seen in the B17 infected mice. Table 4.2 presents a list of metabolites, selected as previously described, that were of interest. As previously mentioned, metabolites of interest were identified as those with a high confidence, assigned function and a high abundance/ large difference in the abundance observed between stages. As indicated in Table 4.2, some of these metabolites had a higher concentration in the plasma of the B17 infected mice. This was due to the overall lower abundance of metabolites.

Metabolomics was done primarily in order to ascertain if a difference in host response could be observed in these two infections, or a metabolomic profile for pre and post-infections could be established. As previously stated, it is hard to differentiate between a host response and changes to the metabolome due to damage caused by the parasite. The presence of [FA (18:1)] 9Z-octadecenamide is potential evidence of a difference in host response, (Table 4.2), it is involved in amide production, which is responsible for killing bloodstream trypanosome forms. It was found to be in the highest abundance in the B17 late stage of infections, which is interesting as this strain had a lower parasitaemia but a greater proportion of non-dividing to dividing forms, which are considered to be more immunogenic (Troeberg et al., 1999).

Other metabolites of interest include [ST methoxy,hydroxy(3:0)] 2-methoxy-3-hydroxy-estra-1,3,5(10)-trien-17-one 3-sulfate, which is a regulator of sterol lipid production. Sterol lipids are not only necessary for parasite growth, but there is also a known difference in the lipid content of slender and stumpy forms (Venkatesan and Ormerod, 1976). There is also a difference in cholesterol metabolism between mouse strains after infection (Noyes et al., 2010).

Table 4.2- Shown are the metabolites of interest from mice infected with Z310 strain. These either had a high abundance in the Z310 infection or had an interesting differential pattern of abundance between the stages sampled. Due to the high abundance of metabolites in the B17 infection, although they have a high abundance in the Z310 infection, relative to the other metabolites identified, the abundance observed in B17 infections for the same metabolite may be higher. Metabolites with significant t test values are shown in bold.

	Z310		B17		
Z310 metab	Day 3	Day 8	Day 3	Day 8	**Function**
[FA (18:1)] 9Z-octadecenamide	**560.89**	**475.66**	**578.15**	**665.33**	Fatty amide production
[ST methoxy,hydroxy(3:0)] 2-methoxy-3-hydroxy-estra-1,3,5(10)-trien-17-one 3-sulfate	**34.28**	**4.82**	39.36	74.79	Sterol lipid production
3-Oxohexobarbital	**9.36**	**11.57**	**31.71**	125.97	Modulator of the neuronal GABA-A receptor
Carnosine	**14.13**	0.20	**23.74**	**5.45**	Dipeptide of beta-alanine and histidine, it has antioxidant, antiglycator and metal chelator properties
N-Acetyl-D-tryptophan	**18.45**	**3.51**	**53.64**	**11.68**	Amino acid derivative of D-tryptophan
Serotonin	**20.67**	**1.40**	**4.00**	0.42	Amino acid/taurine metabolism
1H-Imidazole-4-ethanamine	**106.20**	**0.91**	**11.17**	0.52	Histidine/amino acid metabolism
5-Acetylamino-6-formylamino-3-methyluracil	**36.89**	**0.00**	**1.94**	**0.00**	Purine synthesis and caffeine production
Hypotaurine	**15.57**	**0.33**	**110.15**	**10.22**	Intermediate of taurine, endogenous neurotransmitter. Antioxident
Inosine	**274.67**	**0.62**	3.57	**0.07**	Purine and nucleotide metabolism
NG,NG-Dimethyl-L-arginine	**4.45**	**0.69**	**640.76**	**349.88**	Inhibitor of nitrous oxide
Xanthine	**75.78**	**2.15**	**32.54**	**2.76**	Intermediate in peroxide production and control of parasites in cape buffalo (Wang *et al.*, 2002)

Serotonin has previously been studied in *T. brucei*, and is commonly secreted by many unicellular organisms, the most well-known is *E. histolytica*. Previous work has demonstrated a significant decrease in the levels of serotonin following *T.b. gambiense* infection. Patients infected with Z310 strains have presented with symptoms more characteristic of a *T.b. gambiense* infection than a *T.b. rhodesiense* infection, and with these infections, although a decrease of serotonin was not observed, B17 infected mice had higher serotonin levels than those in seen in the Z310 infection, especially late in infection. In comparison, low levels of serotonin were maintained throughout the course of infection in Z310. Serotonin specific neurons are a suspected target of *T. brucei*, and lesions from this may be in part responsible for some of the neurophysiological changes seen upon infection. As a more acute infection in humans, B17 patients present with severe late stage and are thought to naturally progress to the encephalitic stage quicker; this may explain the higher intensities of metabolites with neural related function, such as serotonin and 3-Oxohexobarbital in the B17 infection (Stibbs, 1984).

Interestingly, the majority of metabolites that were found in a higher concentration in B17 infected mice had functions in amino acid metabolism. Due to the higher proportion of stumpy forms, the B17 strain is far less proliferative than the Z310 strain. As a result, you would expect Z310 to synthesize purines at a greater rate, however because *T. brucei* lack the ability to synthesize purines de novo, they utilize scavenging pathways instead. In higher parasitaemia infections, such as Z310, the demand for purines is far greater, and so there are less purines available in the more proliferative population. Due to a higher percentage of non-proliferative forms in the acute isolate, there is a higher concentration of purines left to scavenge because comparatively, the demand is less.

As seen in Table 4.2, the most abundant metabolite of interest is 9Z-octadecenamide. As previously mentioned, 9Z-ocatadeceanamide indicates a host response, and so a high concentration early in both infections is expected. Increases in N-Acetyl-D-tryptophan in the B17 strain are also interesting as they indicate differences in tryptophan metabolism, which has been studied in trypanosomes since in was noted that trypanosomes can metabolise tryptophan to indole-3-ethanol (tryptophol) *in vitro* (Stibbs & Seed, 1975). Due to the abundance of non-proliferative forms in the B17

infections, the metabolism of many of the parasites is reduced, and so tryptophan metabolism is down-regulated compared to in the Z310 infections.

Another metabolite of interest, NG,NG-Dimethyl-L-arginine, has already been shown to be an important inhibitor of nitrous oxide, and nitrous oxide is important in host response and controlling parasites *in vivo* (Vincendeau and Bouteille, 2006). It is also interesting to note the higher concentration of carnosine observed in the B17 infections compared to Z310 as carnosine is considered to have a protective effect against oxidative damage in mammals, and has previously been shown to reduce parasite burden in *S.mansoni* infections (Soliman et al., 2001). It is hard to decipher whether the higher concentration in carnosine early in the B17 infection is indicative of the host responding to the greater immunogenicity of the predominantly stumpy population, or the result of a later first peak of parasitaemia compared to the Z310 strain and subsequent delay in observing low levels of Carnosine (Soliman et al., 2001).

4.3.1.6.2 Metabolites of interest in B17 infection

All of the metabolites of interest in the B17 infection had a higher intensity/ concentration than in the Z310 infection. Table 4.3 details a list of metabolites sorted using the same criteria as the Z310 infected mice metabolites.

Table 4.3: . This shows the relative abundance of the metabolites of interest in the B17 strain. Metabolites selected using the criteria of confidence >7, intensity significantly different from pre-infection metabolome, and whether they had a known function. The values of the Z310 infected mice compared to their pre-infected values are given for comparison. Intensity values are given relative to the control (pre-infection) values. Unlike with the Z310 strain, all of these metabolites were higher throughout infection than in the mice infected with the B17 strain. Metabolites with significant t test values are shown in bold.

	B17		Z310		
Metabolite	Day 3	Day 8	Day 3	Day 8	**Function**
Guanidinoacetate	**26.33**	**13.81**	**0.52**	**0.25**	Glycine, serine and threonine metabolism
Gamma-Glutamylglutamine	**81.59**	**26.57**	**36.85**	**12.91**	Peptide found in hyperammonaemic patients
[PC (16:0/22:6)] 1-hexadecanoyl-2-(4Z,7Z,10Z,13Z,16Z,19Z-docosahexaenoyl)-sn-glycero-3-phosphocholine	**9.70**	**17.15**	**0.89**	**1.46**	Glycine, serine and threonine metabolism
[PC (18:0)] 1-octadecanoyl-sn-glycero-3-phosphocholine	**11.12**	**5.82**	**2.37**	**2.19**	Main phospholipid of nerve cell membranes
N6-Methyl-L-lysine	**110.39**	**836.19**	0	0	Amino acid metabolism
Indolelactate	**34.62**	**395.49**	0	0	Amino acid metabolism, tryptophan metabolism
[PC (14:0/18:1)] 1-tetradecanoyl-2-(11Z-octadecenoyl)-sn-glycero-3-phosphocholine	**9.17**	**5.79**	**0.92**	**0.94**	Amino acid metabolism
Methyloxaloacetate	**356.39**	**12.17**	**2.19**	**0.10**	C5 branched dibasic acid metabolism/ carbohydrate metabolism

As was observed with the Z310 infection, amino acid metabolism is important for maintaining the parasite population; particularly as the Z310 strain is highly proliferative and will exhaust its resources quicker.

Several of the metabolites are an indicator of the greater extent of damage typically caused by the late stage of the disease in the B17 strain. Guanidinoacetate is an intermediate in the synthesis of creatine, which occurs primarily in the liver and kidneys (Konwar et al., 2013). *T. brucei* is known to cause extensive damage and enlargement of several organs including the liver during infection, and an increase in this metabolite may be a result of this strain causing more extensive damage to the liver (Wang et al., 2008).

Gamma-Glutamylglutamine is another indicator that the infection is damaging the liver. This peptide often indicates hyperammonaemia, an excess of ammonia and can result from liver damage, and/or disorders resulting in an accumulation of ammonia in the brain, which can then lead to malfunctions in the regulation of circadian rhythms, causing symptoms such as daytime sleeping and confusion (Maclean et al., 2012). Free amino acids, and metabolites that are intermediate of amino acid metabolism have been shown to alter during *T. brucei* infection, with alterations seen particularly to hepatic and bloodstream amino acid concentrations.

Alongside amino acid metabolism, carbohydrate metabolism is necessary for controlling parasites as it is required for access to glucose. As such, factors controlling carbohydrate metabolism in the parasite have long been considered a good drug targets (Maclean et al., 2012). Differences in abundance of metabolites involved in carbohydrate metabolism between these strains is expected due to the differing energy requirements resulting from predominantly proliferative and non-proliferative parasite populations. Table 4.3 shows the most abundant metabolites of interest in B17 infections are indolelacetate and methyloxaloacetate, which are involved in tryptophan metabolism and carbohydrate metabolism respectively. The upregulation of which is also seen in the RNAseq data and is discussed later in this chapter.

4.3.2 RNAseq analysis

4.3.2.1 Alignment of RNAseq data shows a high percentage of reads can be aligned from directly sequenced infected host samples

Figure 4.14 shows the percentage of reads mapped to *T. brucei* 927 v8.1 using Bowtie2. Due to sequencing directly from an infected host, there is a high percentage of reads, which did map to *T. brucei* reference, however due to sampling from the first peak of parasitaemia, the number of *T. brucei* transcripts is higher than the host in the majority of these samples. The number of reads mapped is given in bold and represents millions of reads. Sampling at the first peak of parasitaemia, ensured transcript abundance is to be at its highest. The two libraries which had the fewest number of reads mapping to *T. brucei* were Z310 strains, and had both the lowest number of actual reads and percentage of total reads mapped. However these still represented over 40% of the total reads. The remaining libraries had at least 70% of the reads mapping to the Tb927 reference. The zymodeme group each of sample is given on the left hand side.

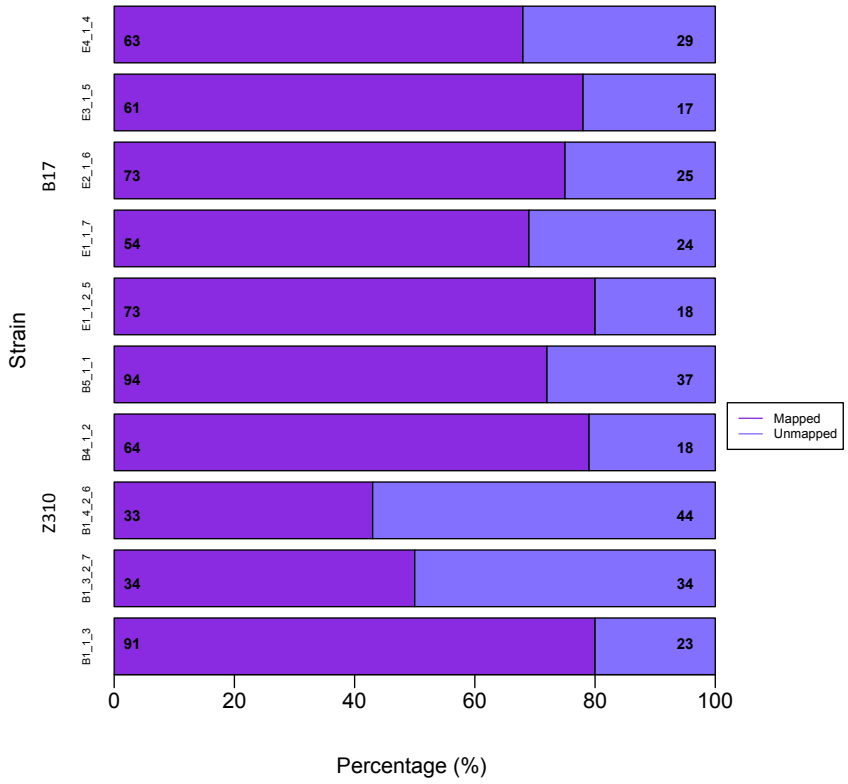

Figure 4.14: This shows the percentage of reads which were aligned to the Tb927 v8.1 genome using Bowtie2 with its default settings. The actual reads mapped is given in millions in bold within the bars. Despite samples being derived from infected hosts, a high percentage of *T. brucei* reads are found within this data.

4.3.2.2 The biological variation seen within the RNAseq data is best explained by a negative binomial model

Figure 4.15A is a plot of the biological coefficient of variation (BCV), which shows the degree of dispersion/variation of a gene in its true abundance between biological replicates (McCarthy et al., 2012). It is expected that there will be a higher degree of biological variation between lower abundance transcripts, and this is observed. However this plateaus for the lower logged counts per million, which reflects that the majority of the transcripts seen had a logged counts per million (logCPM) of at least 4. Actual values are plotted in black and the predicted trend when fitted using edgeR is shown in blue, which shows this model fits the data well.

edgeR was used to fit the data using a negative binomial model. Figure 4.15B shows the logged variation when all the samples are pooled in relation to mean gene expression. Unlike Figure 4.15A, which looks at variation between libraries, this plot looks at the pooled variation against gene expression. Their variation is derived from count variation and adjusted for library size. An increase in the pooled variation is expected with an increase in gene expression and this is observed in Figure 4.15B. The black line represents the poisson mean-variance, which would mean that the mean expression is equal to the mean variance. However this plot shows the data varies significantly from this distribution. Poisson distribution underestimates the variation seen in RNAseq data, especially in highly expressed data, as previously described (Anders and Huber, 2010). This is why a negative binomial model, as used by edgeR and DEseq is a better fit for the data. Binned variances are shown in brown, individual data points in blue, and the blue line represents the predicted gene expression variance using a negative binomial model.

4.3.2.3 Strains from both zymodeme groups show a low degree of inter-sample variation

In order to derive meaningful biological conclusions from the RNAseq data, there should be only small variations in the transcript abundance of each gene between biological replicates. Ideally there should be a high concordance between all replicates for each of the two strains. This is seen in Figure 4.16A and B, with a strong correlation in transcript abundance for each gene seen in both the five B17 and Z310 biological replicates.

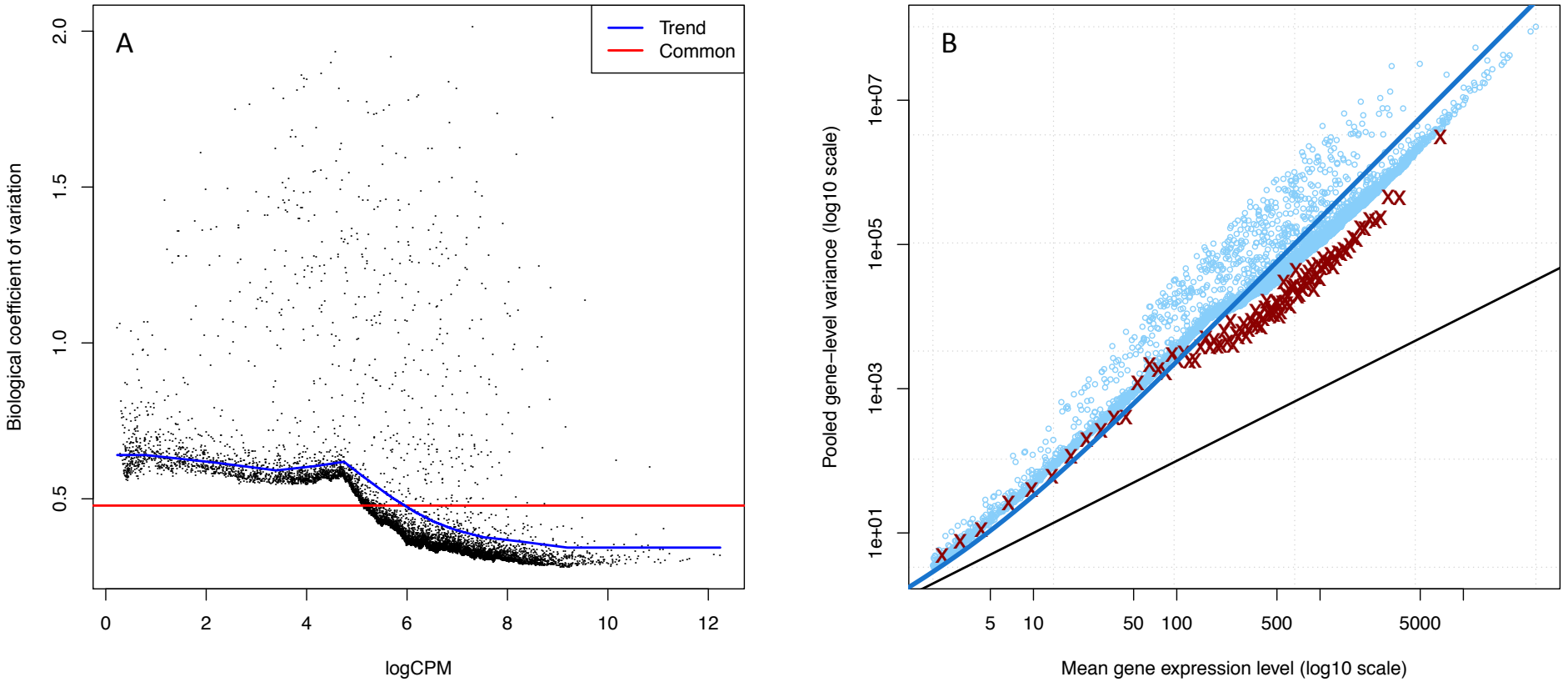

Figure 4.15: A shows the biological coefficient of variation (BCV) in relationship to the logged count per million (logCPM). A higher degree of variation is expected for lower abundance transcripts and this is observed. The plateau seen reflects that the majority of the transcripts had a logCPM value of greater than 4. B shows the variation between genes in relation to their expression level. The black line represents the predicted poisson distribution, the blue line represents the edgeR model. Blue points represent the actual data, and the brown crosses represent binned values.

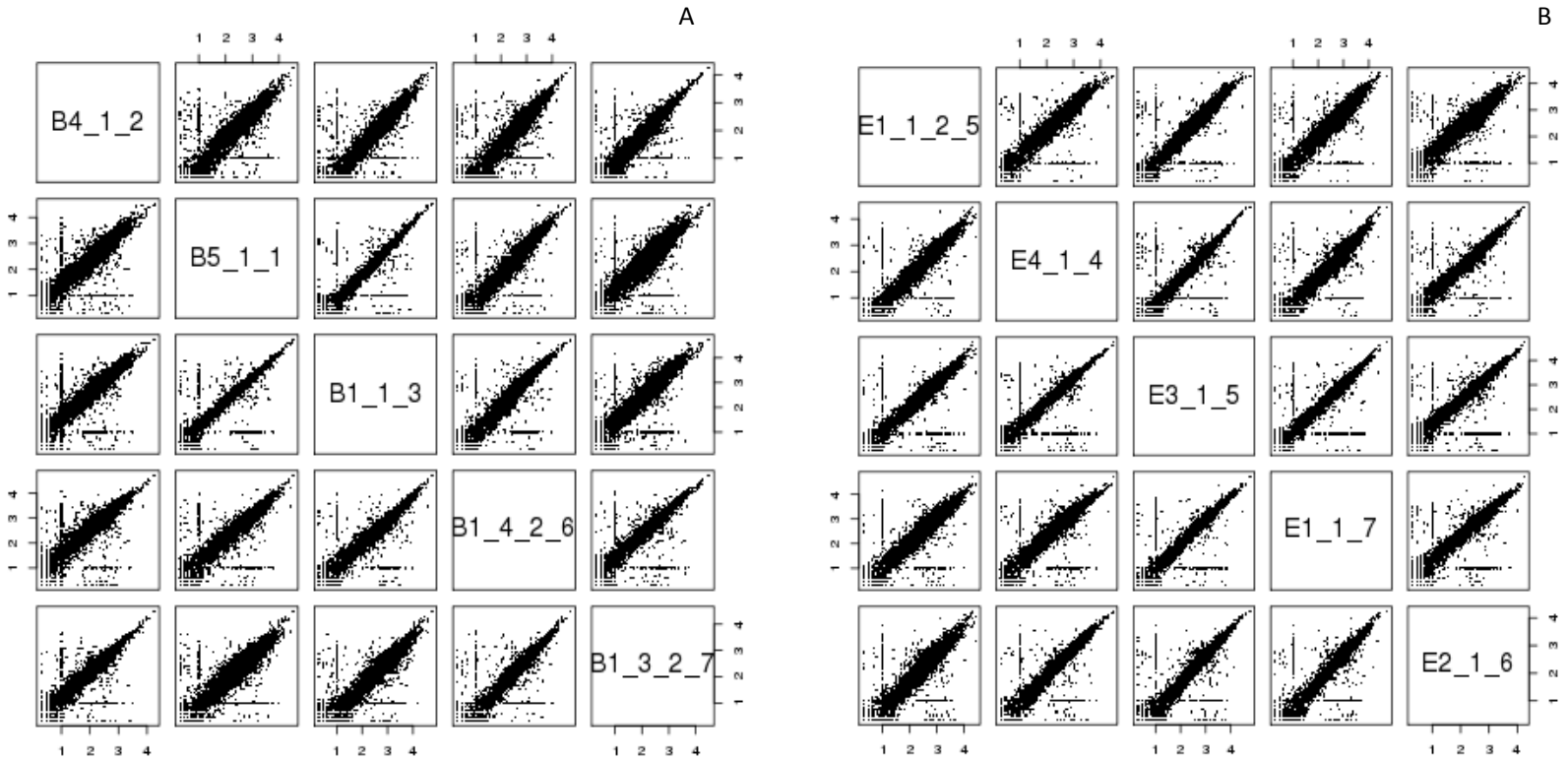

Figure 4.16: This shows the correlation between samples in the expression observed per gene ID. A high correlation as observed here between samples suggests that there isn't a significant difference in the abundances recorded for each sample. All samples from Z310 infections are shown in A, and all samples from B17 infections are shown in Figure B.

4.3.2.4 Both B17 and Z310 strains can be clustered based on BCV values between samples

Although infections with B17 and Z310 strains lead to very different disease manifestations, they are both genetically very similar, with less than 1% difference between *T. brucei* subspecies, and less in instances such as these, where the strains belong to the same subspecies. Due to this, we wouldn't anticipate a large number of genes to be differentially expressed. The MDS plot in Figure 4.17 shows that despite the high degree of similarity between these strains, both the B17 and Z310 strains can be distinguished based on the fold changes observed. This plot clusters the samples based on BCV values between replicates and per treatment group. Although a division between both groups can be seen, they do not cluster tightly because the variation between B17 and Z310 groups is not large, as mentioned above. Due to small between strain differences, the number of differentially expressed genes (DEG) should also not be large.

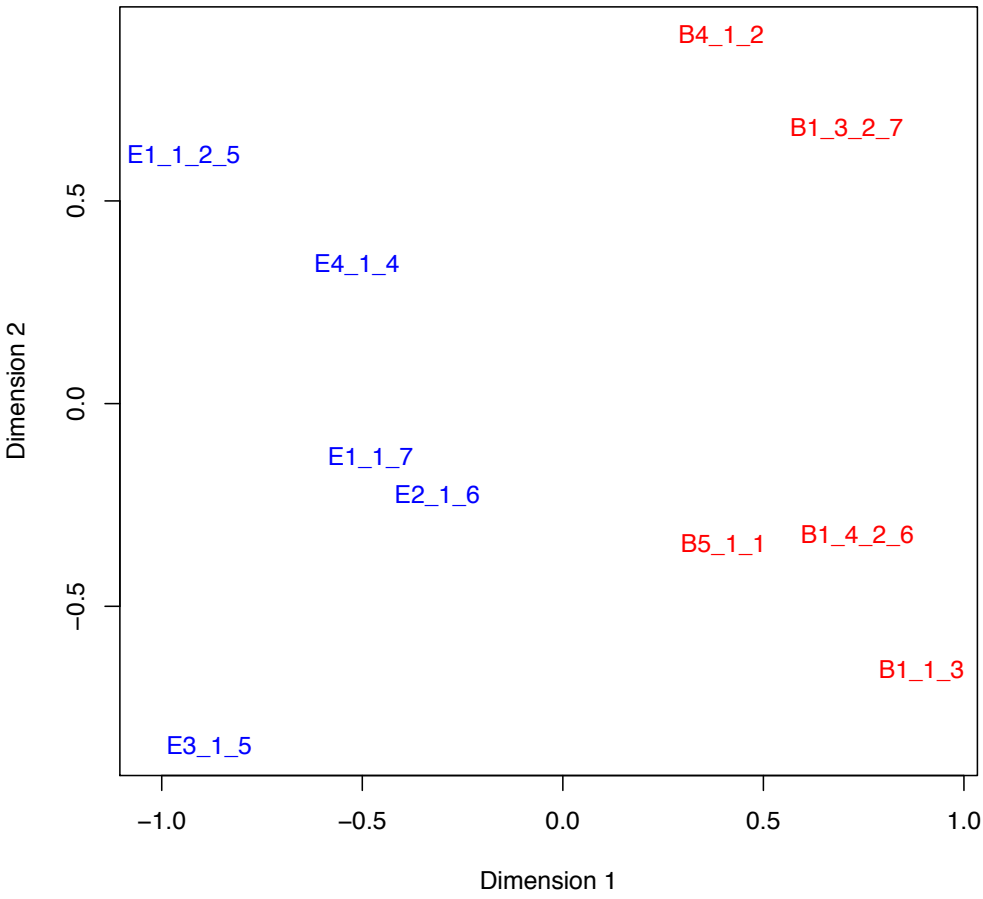

Figure 4.17: MDS plot showing the degree of variability between B17 and Z310 infected samples. B17 samples are shown in blue, Z310 in red. This plot shows that although the two strains can be separated based on fold change variance, the difference isn't large, and so a relatively small number of differentially expressed genes is expected.

4.3.2.5 Only a small subset of the transcriptome is differentially regulated between B17 and Z310 strains

As shown in Table 4.4, 8356 genes were found to be transcriptionally active in both strains. Genes were considered transcriptionally active if they had greater than three reads per million across all samples. A high percentage of these (95%) were not differentially regulated, as expected. Differential expression was defined as significantly higher expression compared to the other strain, using the number of reads mapped to a gene as a measure of expression, and accounting for biological variation between biological replicates. No data is available to provide a background level of transcription, and so fold change represents higher expression relative to the other strain, and so genes identified as upregulated may actually represent a normal level of expression for that strain, and a down regulation of expression in the other strain.

The percentage of differentially expressed genes is almost equal between the two sets of strains, with 189 genes upregulated across all B17 strains and 188 genes across all Z310 strains. This is also seen in Figure 4.18, which shows the log fold change against the counts per million against each gene. Figure 4.18 is a ratio average plot (RA), which uses count frequencies to plot logged fold changes and average abundances using a Bland-Altman plot method. However unlike a MA plot, an epsilon factor is added prior to logging values, which allows for easier identification of genes with similar log fold changes, which may overlap in a MA plot. This also allows for the inclusion of points unique to one condition, and these are shown in yellow.

In Figure 4.18 you can see that the majority of genes aren't differentially regulated. Differentially regulated genes (DEG) are shown in red and there is a clustering of genes that have only a slight difference in fold change, and that have a high fold change instead of a whole range of differential expression. This effect is also exaggerated by the inclusion of the epsilon factor as previously mentioned. Fold change values are assigned so that genes with a higher expression level in the B17 strains are given a minus value and those with a higher expression level in the Z310 strains are given a positive value. This method of assigning fold change values is used throughout subsequent plots.

Table 4.4: 8,356 genes were determined as transcriptionally active in both of these strains. Transcriptionally inactive genes were determined as having less than 3 gene counts across these 10 samples. Only a small number of the active genes were differentially expressed, and an almost equal number of differentially expressed genes were found in B17 and Z310.

	Gene count
Transcriptionally active genes	8,356
Non differentially expressed genes	7,979
Upregulated in B17	189
Upregulated in Z310	188

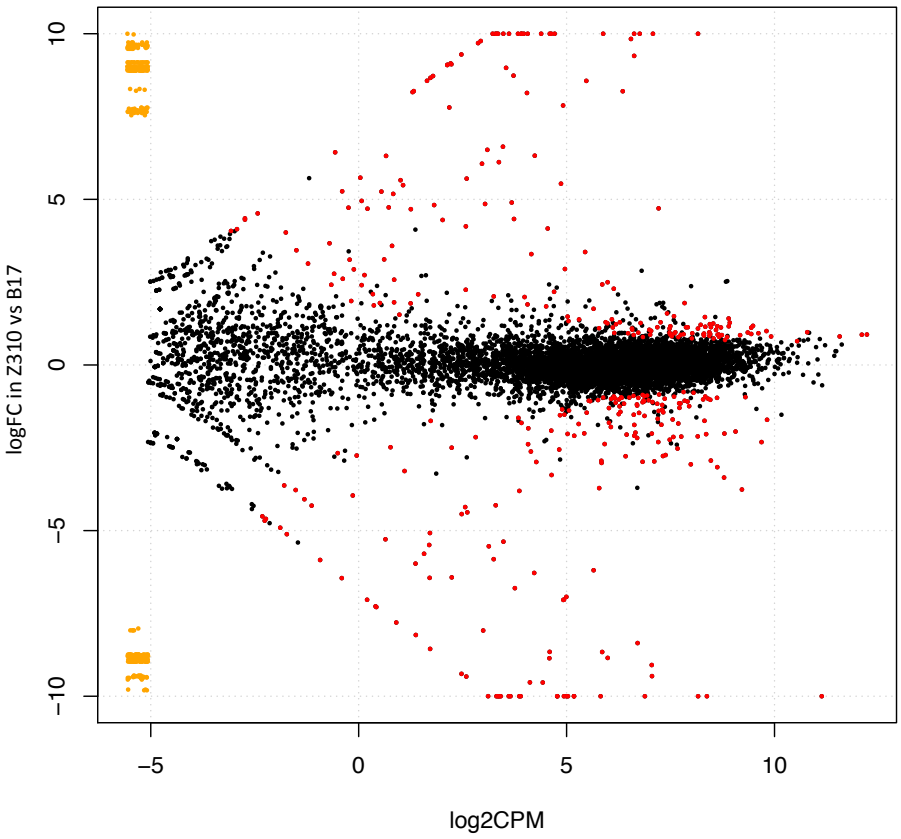

Figure 4.18: This RA plot shows the logged ratio of normalized expression levels between two strains. Those in yellow are those that are unique expression to one strain. The fanned appearance is due to the use of an epsilon factor prior to logging values, which prevents data from being lost due to overlapping, and exaggerates the effects of differential expression so it can be observed more easily. This plot shows as did Table 4.4, that the majority of genes are not differentially expressed.

4.3.2.6 Fold changes in differentially expressed genes are approximately equal in B17 and Z310

As previously mentioned, fold change values are assigned as negative if the gene is upregulated in the B17 strain, and positive if upregulated in the Z310 strain and are shown in Figure 4.19 in blue and red respectively. This figure shows that not only do both strains have near identical numbers of differentially expressed genes, they also have approximately equal fold change distributions, with similar numbers genes with a fold change difference of greater than ten, and approximately half of the DEGs with a fold change of less than 2.

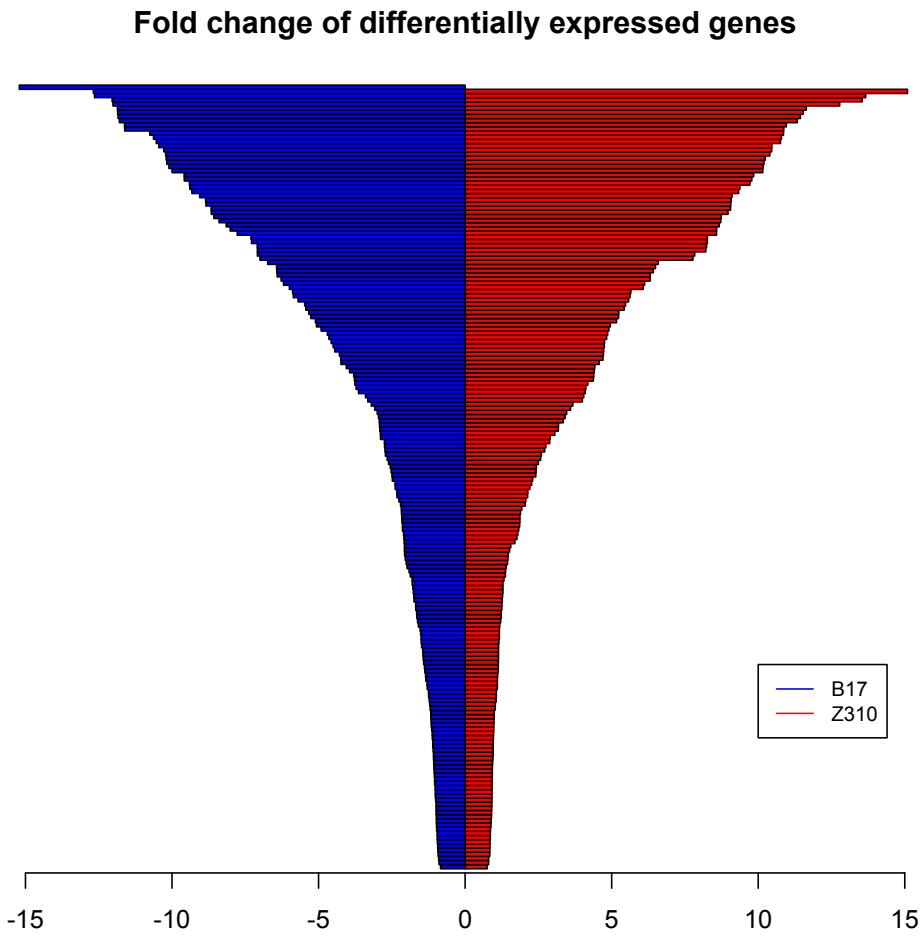

Figure 4.19: This shows the ranked abundance of fold change in the differentially expressed genes identified above. DEGs in blue represent those from the B17 strain, and those in red, Z310. This shows that not only is the absolute number of DEGs relatively equal, but there is an even distribution in the number of DEGs with high and low fold changes.

4.3.2.7 Regions of the genome associated with differential gene expression

Figures 4.20 and 4.21 show the genomic locations of the differentially expressed genes. These plots account for both gene length, with wider bars for longer gene lengths, and for the fold change of expression. Increases in B17 expression are shown in blue, increases in Z310 in red. In both Figures 4.20 and 4.21, small fold changes to both strains are shown across each chromosome at sporadic intervals and are probably of very little functional importance.

Several of these chromosomes appear to have uninteresting patterns of differential gene expression, in particular chromosomes 2, 4, 6 and 7. These chromosomes do have peaks of differential expression, however very few of these have a fold change of greater than five fold. Those that are greater than five fold are located within the telomeric regions of the chromosomes where we would expect either greater sequence bias due to the repetitive nature of the telomeric regions and VSGs, which are likely to be highly expressed and vary between strains Due to the large proportion of the genome dedicated to antigenic variation, and the regular switching and generation of novel antigens, differences in the expression of different VSGs are unlikely to relate to phenotypic differences.

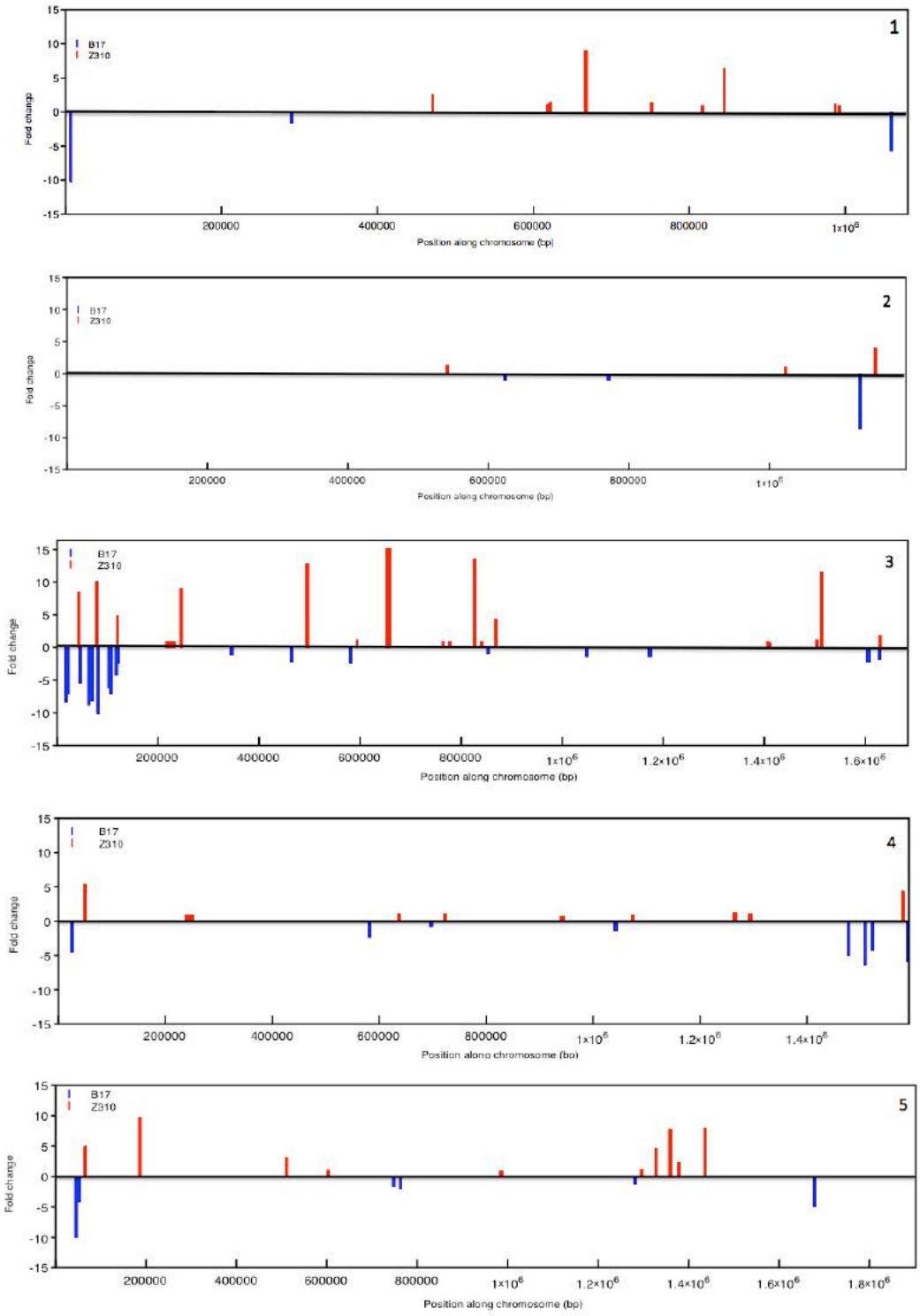

Figure 4.20: This shows the chromosomal locations of differential gene expression in chromosomes 1-5. Blue peaks represent genes that are upregulated in B17 strains, red peaks represent those that are upregulated in Z310. Within these five chromosomes, chromosome 3 shows the most interesting distribution of DEGs.

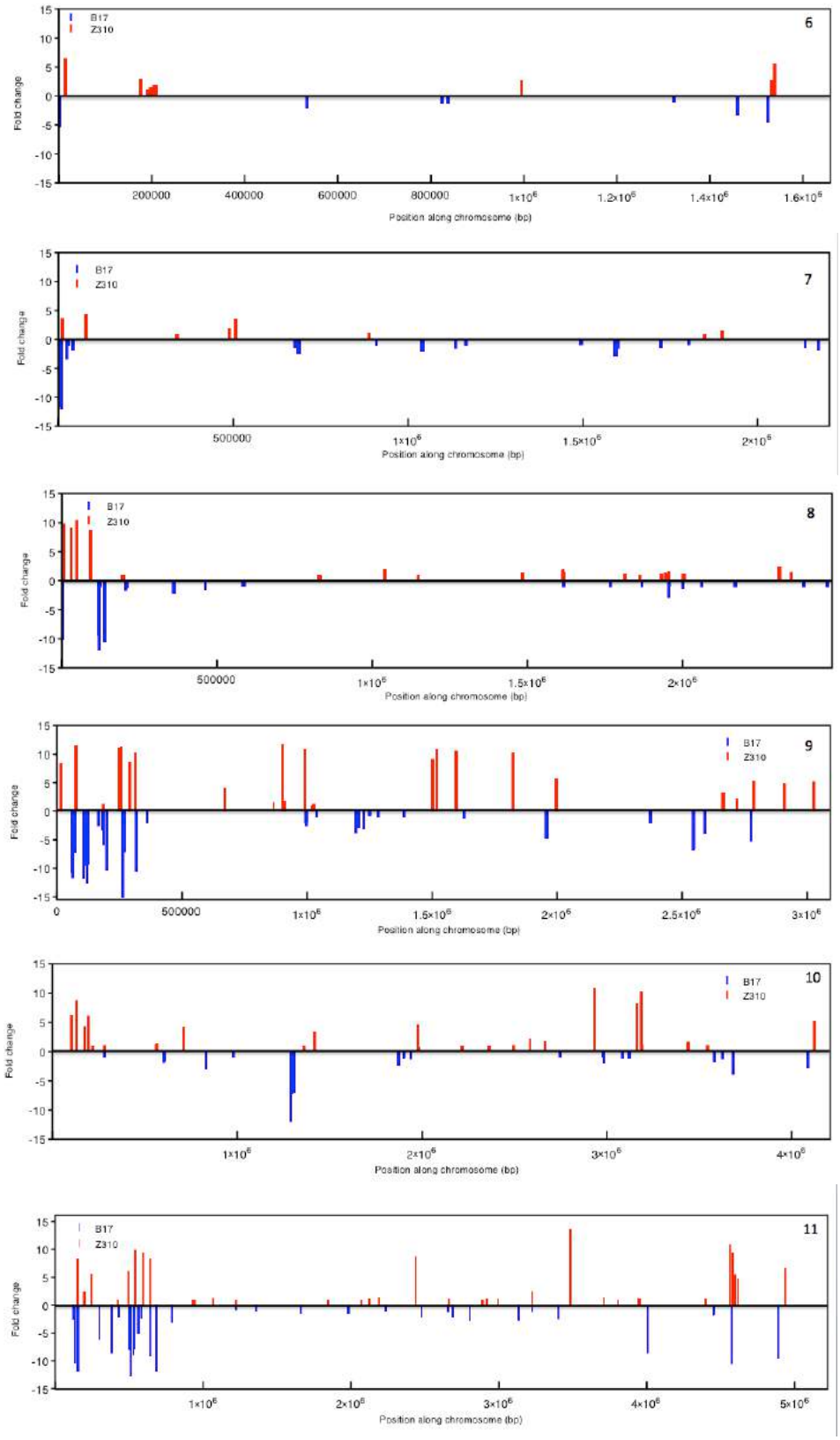

Figure 4.21: This shows the chromosomal location for the DEGs in chromosomes 6-11. As in Figure 4.20, blue peaks represent upregulation in B17, red represent upregulation in Z310. Chromsomes 8,9 and 11 show clusters of differential gene expression in the telomeric regions of the chromosome.

Chromosomes 3,8,9 and 11 all show clusters of high fold change at the 5' end of the chromosome. The majority of these genes are VSGs. Short stumpy forms have greatly reduced antigenic variation because they are incapable of generating new variants on account of their cell cycle arrest (Turner and Barry, 1989; MacGregor and Matthews, 2012). However it appears that there is an increased number of differentially expressed telomeric genes in the B17 strain, which has a much higher abundance of short stumpy forms. This suggests that perhaps prior to cell senescence, the parasites in the B17 infection underwent a greater number of VSG switches, or that the remaining long slender stages have an increased rate of VSG switching compared to the Z310 parasites. Conversely, the Z310 parasites cause the manifestation of a particularly chronic infection, and the long slender stages here could have a reduced number of switches because they are not mounting a high host response.

However it is more likely that this increase in VSGs is an artefact of mapping. The VSG complement of these strains is likely to be different to those seen in Tb927, and for these strains small sequence differences may be leading to reads mapping to different VSGs, leading to these genes wrongly being identified as differentially expressed. The DEGs of most interest are those with the highest logged fold change, those that are clustered with other DEGs, and those that are from more centric regions of the chromosome.

4.3.2.8 Genes generating antigenic variation account for the majority of the high fold change DEGs

As suspected from their chromosomal position, the majority of the genes that had high fold changes were VSGs. Beneath are discussed only genes with a high logged fold change, in this case genes with a fold change of greater than ten. There were twenty genes with a greater than ten fold difference in the B17 infections, and nineteen in the Z310 infections.

4.3.2.9 Genes with a high logged fold change in B17 infections and their chromosomal position

Figure 4.22 shows the function of each of the differentially regulated genes with the highest fold changes in B17. Fourteen of the twenty shown are VSG genes. Four of the remaining DEGs have no assigned function and are hypothetical. As seen in the

previous figure, the majority of these genes have telomeric positions but clusters of high differential expression are not seen except for in the 5' telomeric region on chromosome 9. Of the remaining genes, one is an expression site associated gene, ESAG3, Tb927.11.20150, and has a logged fold change of ~12.64, and the other, Tb927.9.280, is a transferrin receptor. ESAG3 is contained in all but one BES site, however little is known of its function, apart from that it codes for a membrane associated protein (Pays et al., 2001; Hertz-Fowler et al., 2008).

4.3.2.10 Transferrin binding is upregulated in chronic infections

The transferrin receptor is required by the parasite in order to bind and internalize host transferrin, which is necessary for parasite growth. Previous studies have shown that this is facilitated by expression site associated genes ESAG6 and ESAG7 forming a heterodimer complex (Graham and Barry, 1991; Chaudhri et al., 1994; Steverding, 2000). Unlike some ESAGs, ESAG6 and ESAG7 are found only within bloodstream form expression sites (BES). The structure of BESs remains relatively constant between BESs, with multiple copies of each ESAG within the genome, however these still remain poorly characterized (Hertz-Fowler et al., 2008). Only one BES is active at any time, and the resulting transferrin receptor complex has been known to have an altered binding affinity for transferrin dependent on the expression site (ES) it is transcribed from (van Luenen et al., 2005).

Interestingly, ESAG6 is not differentially regulated in these strains, and ESAG7 does not appear to be transcriptionally active. ESAG6 and ESAG7 are known to form the transferrin receptor in order to bind to host transferrin, and previous work has shown transferrin is not bound without complex formation (Steverding et al., 1994). Tb927.9.280 is also known to produce a transferrin binding protein and is putatively regarded as a transferrin receptor (van Luenen et al., 2005). As shown in Figure 4.22, Tb927.9.280 was upregulated greater than ten fold in the B17 strains. Both strains did not have inhibited growth, despite ESAG7 not appearing to be transcriptionally active from the RNAseq data, and so the parasite is evidently still capable of binding transferrin. This suggests that Tb927.9.280 is likely to be an ESAG7 gene from another BES.

In B17 strains, there is a logged 10.76 fold increase in expression of this putative transferrin receptor gene. Due to parasites within the B17 infection being

predominantly cell cycle arrested, you would anticipate their iron requirement to be lower. However haemolysis does occur in trypanosome infections, and this will effect the serum concentrations of iron. Older stumpy forms, which are no longer capable of being transmitted to a tsetse are highly immunogenic, in part due to the products released by them as a result of apoptosis. Stumpy stages can only survive 48 hours after differentiation before cell death, and the products released following apoptosis can cause haemolysis and oxidative damage (Duszenko et al., 2006).

One of the host's responses to *T. brucei* infections is to starve the parasite of iron, limiting their growth. However starving *T. brucei* of iron has been shown to increase transferrin receptor expression levels 3-10 fold, and leads to the relocation of the trypanosomal transferrin receptor complex to the parasite surface instead of the flagellar pocket (Mussmann et al., 2004; Mehlert et al., 2012). In chronic infections, a lower level of host transferrin receptor is seen, and in these infections *T. brucei* has been shown to adapt to low iron conditions by increasing their transferrin receptor levels to allow for greater transferrin uptake (Mussmann et al., 2004; van Luenen et al., 2005).

Gene	Function	logCPM	logFC	P value	FDR
Tb927.9.1010	VSG	8.374067571	15.21067731	7.63E-135	7.27E-131
Tb927.9.490	VSG	5.818099462	12.68392269	1.07E-06	0.000100829
Tb927.11.20150	ESAG3	8.161828957	12.64405913	3.04E-11	6.04E-09
Tb927.10.5200	Hypothetical	5.179299006	12.0426041	2.83E-05	0.001703986
Tb927.7.140	VSG	5.177117541	12.01481549	3.78E-09	5.55E-07
Tb927.11.20700	Atypical VSG	6.882203019	11.86385392	3.55E-05	0.002089804
Tb927.11.19060	VSG	4.997174894	11.85521352	2.27E-06	0.000189169
Tb927.8.440	VSG	5.032819737	11.84138605	1.80E-06	0.000154133
Tb927.9.430	VSG	4.935823353	11.79760949	1.31E-08	1.76E-06
Tb927.7.110	VSG	4.774057948	11.62197075	4.15E-05	0.002373487
Tb927.9.300	VSG	4.77843045	11.61477232	5.11E-34	6.09E-31
Tb927.9.280	Transferrin binding protein	3.89700739	10.76116616	8.94E-05	0.004437533
Tb927.9.1240	VSG	3.862070807	10.63143343	0.000111612	0.005137912
Tb927.11.17070	VSG	3.651250279	10.53704469	0.000144668	0.0062661
Tb927.8.490	Hypothetical	3.609079756	10.45223744	0.000177073	0.007210826
Tb927.11.19020	Hypothetical	3.415290882	10.28001567	0.000146713	0.006325913
Tb927.9.830	VSG	3.348454732	10.2162014	4.27E-07	4.15E-05
Tb927.8.100	VSG	3.377893574	10.1986047	1.07E-05	0.000737958
Tb927.1.10	Hypothetical	3.312929165	10.17780536	0.000157138	0.006596322
Tb927.3.390	VSG	11.13567318	10.10412808	1.15E-66	2.74E-63

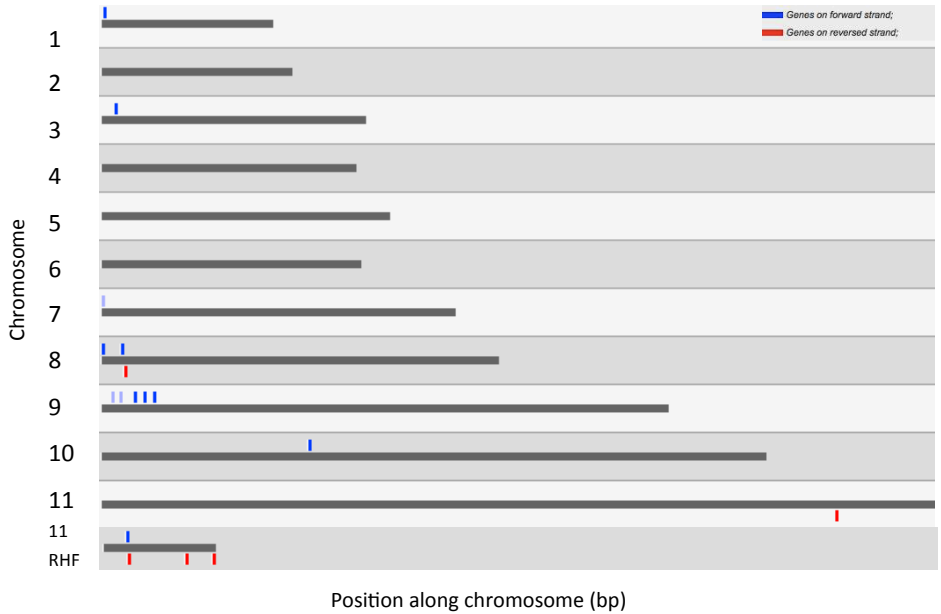

Figure 4.22: Table on the left shows the genes, which have a greater than 10 logged fold change and a P value and FDR cut off of less than 0.05 in B17 infections. The majority of these highly differentially expressed genes are VSGs. On the right are the chromosomal positions of these genes. 11 RHF stands for the chromosome 11's right hand fork and contains several of these DEGs and so is included with the mega chromosomes.

4.3.2.11 Genes with a high logged fold change in Z310 infections and their chromosomal position

As above Figure 4.23 shows the function of each of the genes with the highest differential expression, but in Z310 strains. Although many of the DEGs in Z310 are VSGs, the number is substantially lower than seen in B17 infections. Only six out of nineteen DEGs were VSGs, however a large proportion of the other genes have no assigned function. The remaining three genes encode for an ATP dependent DEAD/H helicase, a phospholipid ATPase and calmodulin. Unlike in the B17 strains, the DEGs have a less sporadic spread across the genome, with the majority on chromosomes 3 and 9, and a couple on chromosomes 8, 10 and 11. Whereas the majority of the DEGs in B17 are positioned at the telomeres, very few are in the Z310 strains. Calmodulin and the phospholipid ATPase both located on chromosome 11.

4.3.2.12 Calmodulin regulates calcium-signaling pathways, which are essential for parasite transversal across the BBB

Calmodulin is a calcium binding protein, which regulates calcium signaling pathways. It is present in all eukaryotes, but trypanosomal calmodulin is distinct from host calmodulin (Ruben and Patton, 1985). Calcium signaling is required for growth and for the migration of parasites across the blood brain barrier (Nikolskaia et al., 2006). Calcium regulation has also been shown to be key to virulence, cell invasion, differentiation and replication and stores of calcium are sequestered in *T. brucei* in acidocalcisomes, which are present in multiple trypanosomatids and apicomplexans, but not in mammals (Docampo and Moreno, 2001).

Calmodulin has also been found in all protozoan parasites that have been investigated (Moreno and Docampo, 2003). There are multiple copies of the calmodulin gene in *T. brucei's* genome, another of the genomes encoding calmodulin, Tb927.9.6130, is also upregulated in Z310, albeit weakly, with a fold change increase of ~1. In Z310 infections, calmodulin expression is increased 10.4 fold. This may be due to higher demand because a high proportion of the parasites in Z310 strains are highly replicative. Calcium binding proteins called calflagins, have also been studied and shown inhibition of calcium signaling can lead to prolonged host survival and a reduction of parasitaemia (Emmer et al., 2010). Due to the plethora of pathways calcium signaling is involved in, which include replication and differentiation, it could

also be potentially responsible for the difference in bloodstream form abundance we see in these different strains (Emmer et al., 2010).

4.3.2.13 ATP dependent DEAD/H helicases are very highly upregulated in Z310 infections

Helicase activity was described in Chapter 3, because all the strains used in first design sequence capture contained a conserved deleterious SNP within a helicase. As mentioned in Chapter 3, helicases are central to many core processes including transcription, replication and degradation, because they control the unwinding and restricting of RNA using ATP (Linder and Jankowsky, 2011; Mehta and Tuteja, 2011; Gargantini et al., 2012). Due to the highly proliferative nature of these Z310 infections, helicase upregulation may be required to enable the parasites to replicate at a rate that will maintain its current population size. This gene had the highest upregulation, with a fold change of greater than 15.

4.3.2.14 Phospholipid ATPase upregulation in a highly replicative parasite population

Phospholipid ATPases have not been extensively studied, and like the upregulated helicase gene, their action is dependent on ATP, however they are known to be cell surface associated (Richmond et al., 2010). *T. brucei* has a dense phospholipid coat, and this allows the parasite to respond to environmental cues and exert partial control over metabolites entering and leaving the cell (Richmond et al., 2010). This was the gene with the second highest fold change with a 13.66 increase in Z310 strains. Similarly to the H helicase, this gene may be upregulated in an attempt to maintain a highly proliferative parasite population because phospholipids constitute a large proportion of the parasite's content. Phospholipid demand will be high in a highly proliferative infection because they will be needed to construct new cell membranes.

Gene	Function	logCPM	logFC	P value	FDR
Tb927.3.2600	ATP dependent DEAD/H helicase	8.160253118	15.09263794	9.35E-11	1.71E-08
Tb927.11.13000	Phospholipid ATPase	6.758900435	13.66454494	4.64E-57	8.84E-54
Tb927.3.3200	Hypothetical	6.623608273	13.54769223	1.43E-07	1.51E-05
Tb927.3.1900	Hypothetical	5.879615637	12.7804061	1.02E-69	3.23E-66
Tb927.9.5000	Hypothetical	4.713187787	11.64150165	8.84E-07	8.51E-05
Tb927.3.5400	Hypothetical	4.631026826	11.53589164	1.25E-06	0.000114686
Tb927.9.340	VSG	4.595613865	11.43813999	2.53E-05	0.001574693
Tb927.9.970	VSG	4.388648948	11.33249247	4.02E-05	0.002320893
Tb927.9.960	VSG	4.064870278	10.95965137	2.43E-06	0.000195844
Tb927.9.5800	Hypothetical	3.971747599	10.871956	1.18E-08	1.60E-06
Tb927.11.17050	VSG	3.918345461	10.86203878	2.71E-05	0.001644533
Tb927.10.12050	Hypothetical	3.910662915	10.79273793	1.60E-06	0.000139734
Tb927.9.9500	Hypothetical	3.832520111	10.75941754	2.21E-06	0.000187617
Tb927.9.10100	Hypothetical	7.076387543	10.46637481	9.56E-101	4.55E-97
Tb927.8.260	VSG	3.615013912	10.45871001	7.05E-05	0.003712087
Tb927.11.13040	Calmodulin	3.478579555	10.39622336	1.01E-05	0.000708032
Tb927.9.11700	Hypothetical	3.319560602	10.24017238	8.19E-05	0.004173646
Tb927.9.1230	VSG	3.352236419	10.19641947	8.64E-05	0.004372921
Tb927.10.13080	Hypothetical	3.28986042	10.17198672	0.000112193	0.005139858

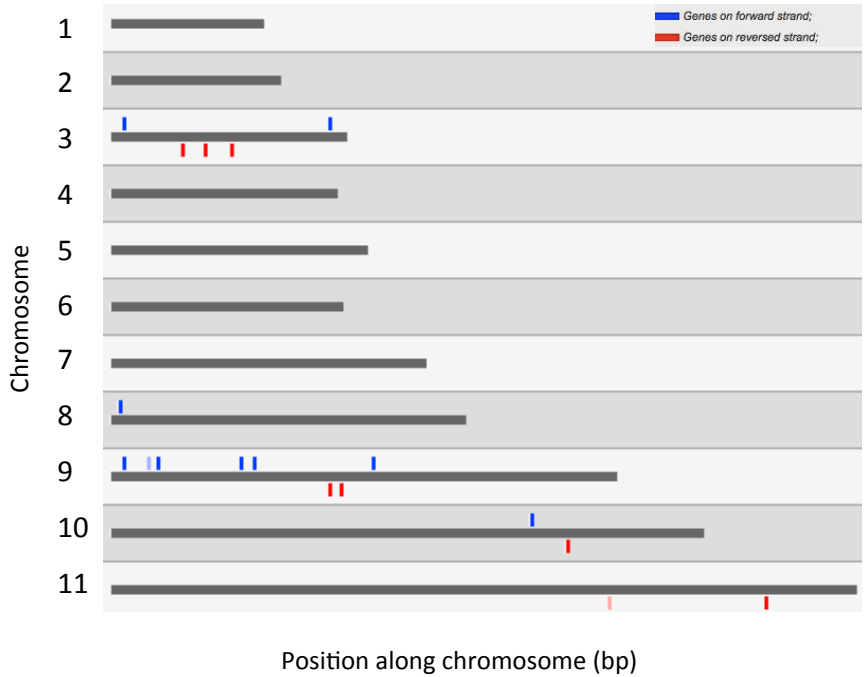

Figure 4.23: As above with the B17 strains, the table on the left shows the genes which have a greater than 10 logged fold change and a P value and FDR cut off of less than 0.05 in the Z310 infections. A high percentage of these highly differentially expressed genes are VSGs, however this is significantly less than was observed in B17 infections. On the right are the chromosomal positions of these genes. There were no DEGs on the 11 RHF and so this isn't included in the figure.

4.3.2.15 Fold change in the DEGs identified is not correlated with the frequency of SNPs in WGS data

The genomic and transcriptomic data can be tied together by examining the frequency of SNPs in the WGS data in the genes that are differentially expressed. The WGS data did not identify any sole causal variants that could explain the different phenotypes observed in these strains. Similarly, it is hard to identify a key regulatory change in these strains causing these phenotypic differences from transcriptomics data alone. In view of this, Figure 4.24 shows the frequency of SNPs found in the WGS data from the genes identified as differentially expressed in the RNAseq data. These are SNPs that are unique to either one of these strains. DEGs upregulated in the B17 strain is shown in blue, DEGs upregulated in the Z310 strain in red. Figure 4.23A shows the total number of SNPs per DEG and Figure 4.23B shows the number of heterozygous SNPs per DEG only.

There is no strong correlation between and increase in fold change and the frequency of SNPs as seen in Figures 4.24A and 4.24B. Many genes have a low fold change and a SNP frequency of up to 10.

Another link between the transcriptomic data and the genomic data is iron transport. The gene putatively encoding a transferrin receptor is differentially expressed between these strains in the RNAseq data, as discussed in section 4.3.2.10, but this has only 5 homozygous SNPs in B17 and 4 in Z310. Another gene, Tb927.11.4430, which encodes the transmembrane component of ferric reductase, is not differentially expressed in the RNAseq data but does contain 50 SNPs in the enrichment data, which were conserved between all seven strains used in both capture experiments. These variations in iron regulation at both a genomic and transcript level, albeit on different genes, suggests it could be important in generating the phenotypes observed in these strains.

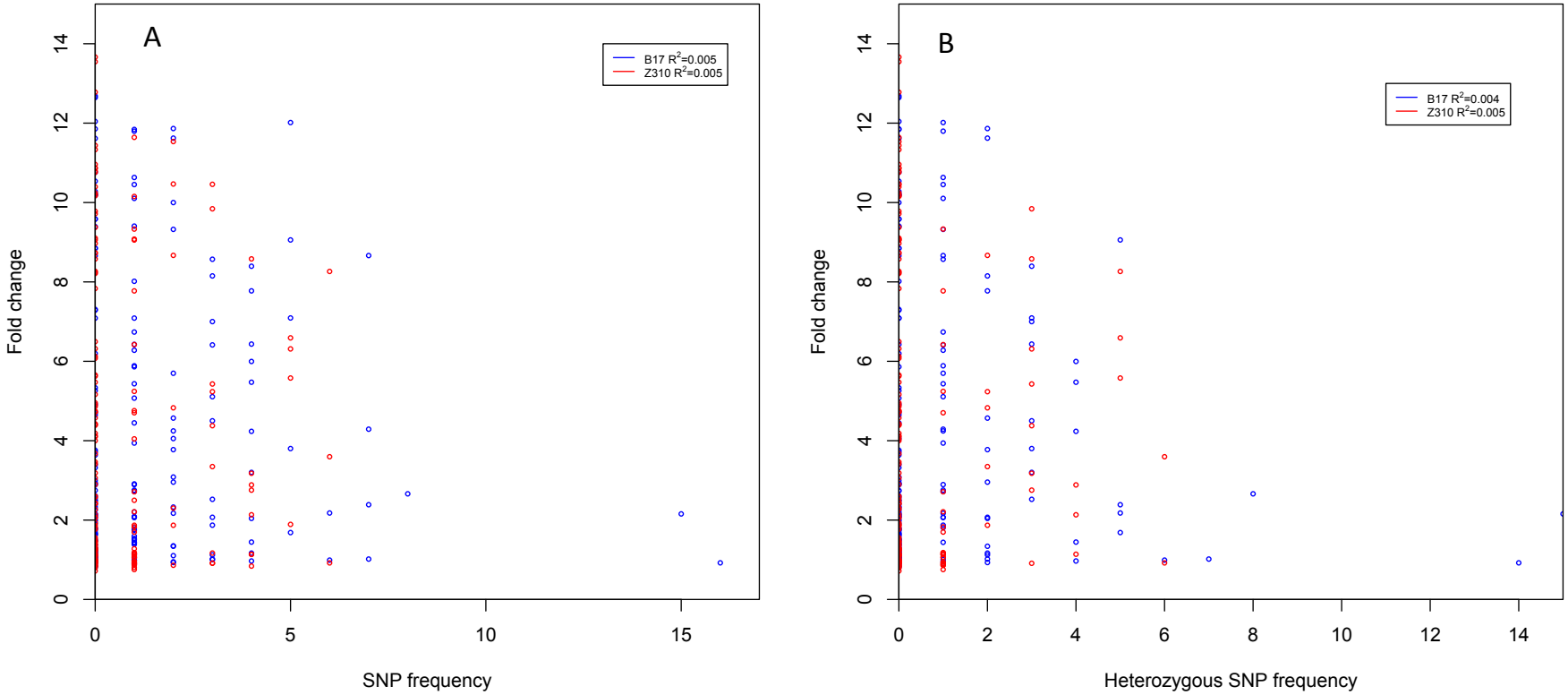

Figure 4.24: Plots A and B show the SNP frequency in the DEGs from the WGS data available and unique to each strain. Strain B17 is shown in blue, Z310 in red. A shows the total frequency of SNPs unique to either B17 or Z310 compared to their fold change in the RNAseq data. B shows ony the heterozygous SNPs compared to the fold change in RNAseq data. Although there are a couple of genes with high SNP frequencies and a high fold change, there is no strong correlation between SNP frequency and fold change.

4.3.2.16 KEGG analysis demonstrates differences in the metabolic pathways enriched in transcriptomic data between B17 and Z310 strains

All of the 377 DEGs from both strains were assigned KEGG pathway IDs and these are shown per strain in Figure 4.25, with metabolic pathways upregulated in B17 strains in blue, and upregulated in Z310 in red. Hypothetical genes could not be assigned a KEGG group, but the remaining genes fit into nine pathways. DEGs from B17 were enriched for eight of these pathways; Z310's DEGs only enriched two. Only one of these pathways was enriched in both of the strains, this was glycolysis and gluconeogenesis. Due to vast differences in the relative abundance of slender to stumpy stages in these strains, differential expression of glycolysis is to be expected due to the differing energy needs of slender and stumpy stages. The fold enrichment is approximately equal in this pathway, but slightly increased in B17 strains. Due to upregulation in both of these strains, this suggests that different parts of the glycolytic pathways are being differentiatially regulated in each strain and Figure 4.26 shows this is correct. Five differentially regulated genes involved in glycolysis and gluconeogenesis were found in B17, and are shown in blue, three in Z310 and are shown in red.

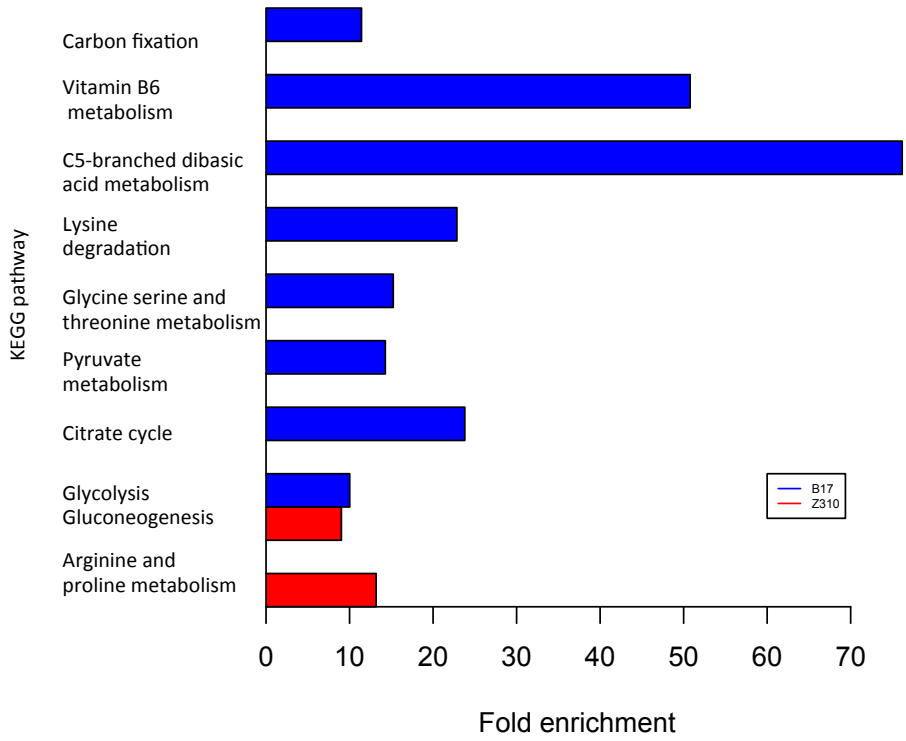

Figure 4.25: DEGs from each strain were assigned to KEGG pathway, and only 9 pathways were significantly enriched . DEGs from B17 are shown in blue, DEGs from Z310 are shown in red. There is little overlap between the two strains, with DEGs from both only seen in the glycolysis and gluconeogenesis pathway. B17 also has DEGs in 8 of these pathways, whereas Z310 is only enriched in 2.

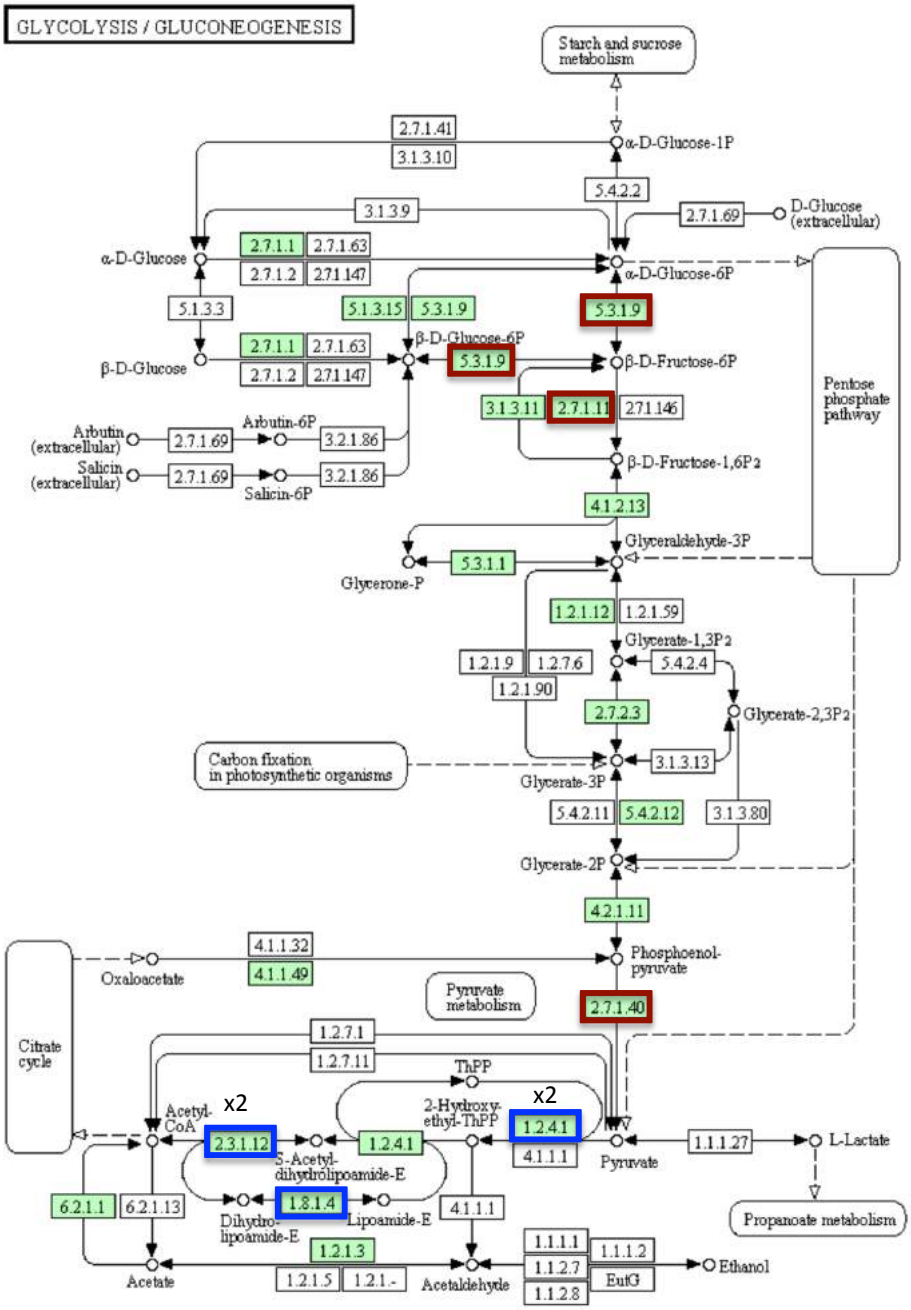

Figure 4.26: Annotated KEGG map for glycolysis and gluconeogenesis. KEGG IDs are given for each gene involved in the pathway for example 1.8.1.4 in the white boxes. Green boxes show genes that have been validated to be part of the glycolytic pathways in *T. brucei*. DEGs involved in this pathways are shown for B17 in blue, and for Z310 in red. There is no otherlap in the DEGs in each strain, showing different parts are being differentially regulated in each strain. Three genes are shown for Z310, five for B17. KEGG ID 5.3.1.9 is used by two separate pathways, and is highlighted twice. Two genes in B17 related to KEGG IDs 2.3.1.12 and 1.2.4.1, and this is indicated on the graph. This KEGG map is adapted from www.genome.jp/kegg/

4.3.2.17 Pyruvate metabolism is upregulated within a predominantly stumpy stage population

Both stumpy and slender bloodstream stages solely depend on glycolysis to produce ATP. Pyruvate is the primary end product of glycolysis, however bloodstream forms lack a functional mitochondria or Krebs cycle and so pyruvate can not be further metabolized, it can only be excreted back into the bloodstream to prevent an accumulation within the parasite, which would be toxic (van Grinsven et al., 2009b). The upregulation of genes belonging to the pyruvate metabolism KEGG group suggests and upregulation of glycolysis, which is seen in this figure, and the excretion of pyruvate. This agrees with previous observations, in which acetate has been found to be produced by stumpy forms, but not slender, and with the predominantly stumpy population present in B17 infections this is unsurprising (van Grinsven et al., 2009b).

C5-branched dibasic acid metabolism is by far the most enriched pathway within the DEGs, with an enrichment in excess of 70 fold in B17 parasites. This pathway was also shown to be upregulated in the metabolomic data. Part of this metabolic pathway involves the metabolism of pyruvate and the production of acetate (van Grinsven et al., 2009b). This pathway is also fed into by the citrate cycle, which is also upregulated in B17 parasites. The assignment of DEGs to this metabolic pathway gives an underestimation of the actual upregulation of pyruvate metabolism.

4.3.2.18 An upregulation in amino acid metabolism could increase the parasite's exposure to nitric oxide

Glycine and serine are non-essential amino acids and trypanosomes scavenge these nutrients from their host. Threonine is an essential amino acid, and is used both for the production of glycine and for the production of pyruvate (Linstead et al., 1977). Given the upregulation of pyruvate metabolism within the B17 parasites, it is unsurprising that glycine, serine and threonine metabolism is also upregulated within these parasites.

Interestingly, proline metabolism is increased in Z310 parasites, despite containing a predominantly slender population. Glucose is the primary energy source for bloodstream form parasites and proline is used by procyclic forms due to the high abundance of proline in tsetse haemolymph and scarce quantities of glucose (Bursell,

1981; van Grinsven et al., 2009a). Arginine metabolism is also upregulated, and this has previously been related to parasite killing by the host. Trypanosomes require arginine for trypanothione, polyamine and DNA synthesis (Gobert et al., 2000). Arginine availability enables the host to increase its production of nitric oxide (NO), which *T. brucei* is susceptible to (Gobert et al., 2000). Upregulation of this KEGG pathway may be necessary to maintain the demands of a highly proliferative population.

4.3.2.19 Vitamin B6 metabolism

Vitamin B6 metabolism is the second most upregulated pathway in B17 parasites, however little research has been done on the metabolism of vitamin B6 in *T. brucei*. Vitamin B6 is synthesized by all complex organisms, and some protozoa, however *T. brucei* are most likely to scavenge their requirements from the host. One protozoan has been shown to require vitamin B6 for normal growth, *Tetrahymena geleii*, however whether it is absolutely required in *T. brucei* is unknown.

4.3.2.20 GO term analysis agrees with KEGG analysis and shows that only a few metabolic pathways are significantly upregulated/ differentially expressed between strains

Similarly to the KEGG analysis, all of the DEGs for each strain were used for assigning GO terms to. As explained in Chapter 3, which also makes use of GO term enrichment analysis, each GO category is coloured to represent the degree of significance for each term, and significance is scored based on a Bonferroni corrected P value. Terms range from red to blue in relation to insignificant to significant. As can be observed from Figure 4.27, DEGs are assigned to multiple functional groups, however the majority of these are not significantly enriched. However a few of these terms are significant and correlate with the observations from the KEGG enrichment.

4.3.2.21 Significant GO terms in the Z310 infections reflect the high abundance of slender stages in these infections

The assigned GO terms from Z310 infections are shown on in Figure 4.26A. Here the most significant terms are glycolysis and carbohydrate metabolism related, with glycolytic process, gluconeogenesis and carbohydrate catabolism GO terms significant.

This agrees with what was seen in the KEGG data, that glycolysis and gluconeogenesis is upregulated. The most significant GO term is the generation of precursor metabolites and energy, which contains all the central pathways for metabolism including glycolysis. This upregulation will be required due to high demand by the highly abundant and proliferative slender forms.

Although the effect is more exaggerated in the B17 strains, Z310 also had several VSGs with high differential expression, this is seen by an increase in antigenic variation and defense response terms also shown in Figure 4.27A. Genes related to locomotion and microtubule based processes were also enriched, which is a reflection of the slender dominant population. Long slender stages are both more motile due to their elongated flagella compared to the stumpy stages, and will require structural reassortment by microtubules due to their constant replication.

Similarly to the GO terms enriched in the Z310 infections, there are a variety of GO terms in the DEGs of the B17 infection, however only a select few are significantly enriched. More DEGs could be assigned GO terms from the B17 DEGs, and these form more distinct clusters than seen in Z310, see Figure 4.27B. A high percentage of DEGs in B17 were assigned to antigenic variation or host region, and these cluster at the bottom of Figure 4.27B. This relates to the high number of VSGs found to be highly differentially expressed in B17. Other significant GO terms of interest include GPI anchor metabolism, acetyl coA metabolism, carboxylic acid catabolism and small molecule catabolism. GPI anchor metabolism involves the mechanism for attaching proteins to the surface of the glycolipid bilayer covering the surface of the trypanosome. If VSG switching or expression is higher in these parasites, then the glycolipid membrane is more likely to be being actively remodeled by the parasite.

As previously shown in the KEGG analysis, pyruvate metabolism is upregulated in B17 parasites, and acetate is produced from stumpy stages. Acetate and pyruvate metabolism are the precursors for acetyl-coA, and so acetyl coA metabolism enrichment reaffirms the upregulation of glycolysis and pyruvate metabolism. Similarly, enrichment in the DEGs related to carboxylic acid catabolism reflects the C5-branched dibasic acid metabolism upregulation seen in the KEGG data. Small molecule catabolism is the breakdown of small molecules such as low molecular weight saccharides, which you agrees with the upregulation of gluconeogenesis seen in the KEGG analysis.

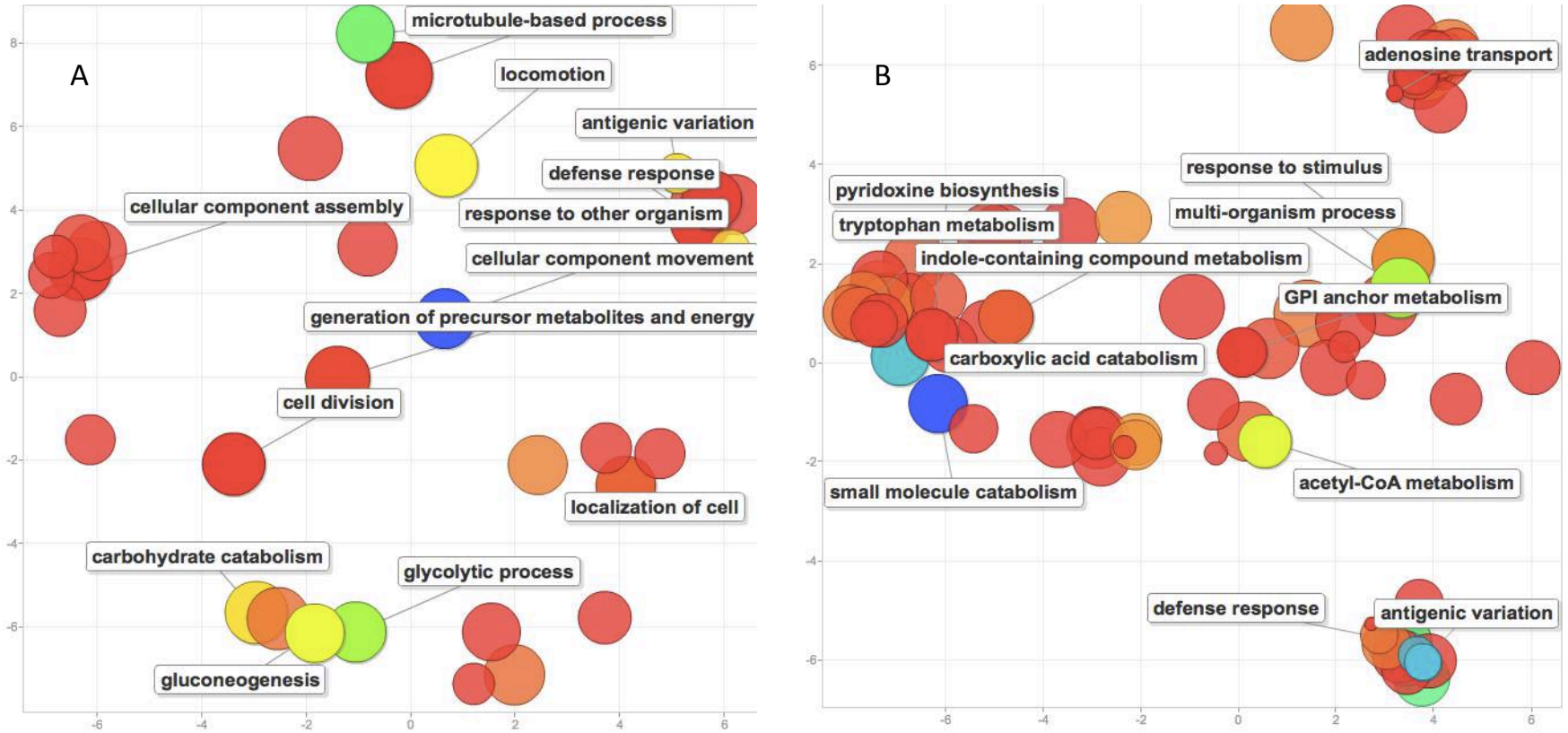

Figure 4.27: GO term analysis generated from REVIGO software (Supek et al., 2011) shows the GO terms assigned to all of the DEGs in Z310 on the left in A, and B17 on the right in B17. Insignificance to significance is shown from red to blue, and although many GO terms are assigned in both, only a few GO terms are highly enriched in either strain. In Z310, A shows that gluconeogenesis, antigenic variation and the generation of precursor metabolites and energy are among the few that are significantly enriched in the DEGs. Similarly in B17 in B, small molecular catabolism, acetyl-CoA metabolism and antigenic variation are the key significantly enriched GO terms. Although VSGs were constituted a high percentage of the DEGs in both strains, this was considerably higher in B17, and this is shown by the cluster in the bottom right hand side of the B.

4.3.2.22 Other genes of interest are also differentially expressed but have smaller fold change differences

Some of the genes with potential functional interest were differentially expressed between B17 and Z310 strains, but had a lower fold change, and so weren't included in the above discussion of the genes with the highest fold changes. Beneath is a brief discussion of some of the genes with a fold change typically of less than five fold which are of interest because they have either been linked to a particularly important pathway, have been implicated in drug resistance and virulence or could be potentially important in explaining the disparity in bloodstream form abundances. All of the genes discussed below where located within chromosomes 6-11 and their chromosomal position is shown in Figure 4.28 below. Interestingly, all of the genes from this subset on chromosomes 9-11 localize to the more centric regions of the chromosome.

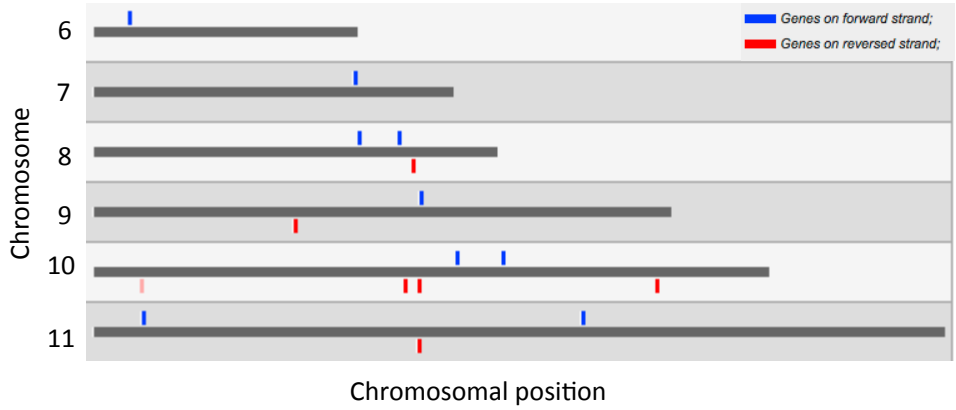

Figure 4.28: This shows the chromosomal position of the genes discussed in this section. Only chromosomes 6-11 are shown because all of these genes were located within this region. Interestingly, the genes cluster to the centric regions of the chromosome on chromosomes 9, 10 and 11.

As discussed in Chapter 3, the haptoglobin haemoglobin receptor (HpHbR) is important because it binds the TLF1 complex and is used by *T.b.gambiense* to resist lysis. As was shown in Chapter 3, the HpHbR in the Z310 strain is more similar to DAL972's copy of HpHbR than Tb927's. The expression of HpHbR was also two-fold higher in Z310 than in B17 in the RNAseq data.

4.3.2.23 Multiple genes with implicated roles in drug resistance and virulence were differentially expressed

Seven genes, which had either been implicated in drug resistance or increased virulence, were differentially expressed; six of these were upregulated in B17. Two of these were peptidases, which are known virulence factors in *T. brucei*. One, CBP1, is a serine peptidase and this has previously been linked to suramin resistance, the other, Tb927.11.1100, was a cysteine peptidase. Amino acid permease 24 (AAT6) was also upregulated in B17, and has linked to eflornithine efficacy and resistance (Alsford et al., 2012). A NT10 purine nucleoside transporter linked to nufurtimox efficacy and resistance was also three fold higher in B17 (Alsford et al., 2012). A pyrophosphate encoded by Tb927.11.7060 which has previously shown to be essential for virulence in mice, was also upregulated (Lemercier et al., 2004; Kotsikorou et al., 2005). The only gene with upregulated expression implicated in drug resistance in the Z310 strain was aquaglycerporin 3 (AQP3), and this has been linked to potential drug resistance and the efficacy of suramin (Alsford et al., 2012).

4.3.2.24 Genes with glycolytic roles are upregulated in Z310 strains

Phospholipase A1, phosphoglycerate mutase, pyruvate kinase 1 and pyruvate dehydrogenase are important in glycolysis and differentially expressed. Unsurprisingly, due to the highly proliferative population seen in Z310 infections, the glycolytic demand is greater and so three of these genes are upregulated in the Z310 strains. One of these is phospholipase A1, which is upregulated nearly six fold in the Z310 strains and is localized to the glycosome (Smith and Buetikofer, 2010; Guether et al., 2014). The other two that are upregulated in Z310 to a lesser extent are pyruvate kinase 1 and phosphoglycerate mutase. Both are important in glycolysis and phosphoglycerate mutase has been implicated in virulence in Leishmania (Chevalier et al., 2000; Mercaldi et al., 2012; Kramer et al., 2014). The only glycolytic gene upregulated in B17 is pyruvate dehydrogenase (Flynn and Bowman, 1970; 1973; Sykes and Hajduk, 2013).

4.3.2.25 Several genes are indicative of the differences in bloodstream form abundances

As described in Chapter 3, the parasite population in B17 infections is comprised largely of non-proliferative bloodstream forms, whereas Z310 infections contain primarily highly proliferative bloodstream forms. The stumpy stages are less motile

than the long slender stages on account of their shortened flagellum, and evidence of this is seen in the comparatively higher levels of paraflagellar components, PFC17 and PFC19, in Z310 strains (Portman and Gull, 2010). PAD2 is a stumpy stage specific marker, is also increased in B17 infections, and this correlates with the QPCR and microscopy data in Chapter 3. Stumpy stage parasites have approximately 48 hours before cell death, but the parasitaemia in B17 infections does not stay at low levels despite only a small proportion of the parasites being capable of replication (Duszenko et al., 2006). Interestingly, in B17, gene translationally controlled tumor protein (TCTP) is upregulated three fold and related to the promotion of growth and apoptosis prevention (Erben et al., 2014).

4.3.2.26 Correlation between metabolomic and transcriptomic data

Transcriptomic data and metabolomic data can both give a snapshot of a biological system, which is generally transient. Transcripts are continually produced, regulated and degraded, similarly to in metabolomics in which metabolomic pathways are constantly regulated, and the products continually produced and degraded at a non-constant rate. Metabolomic data has a much lower resolution in comparison to transcriptomic data, with less than 1000 metabolites that can be identified, less that can be confirmed. In comparison, transcriptomic data can be used to look at all transcriptionally active genes, which in these experiments was over 8000 genes. However multiple genes, which regulate a pathway may be differentially expressed under a treatment, and metabolomic data using the same treatments, can also be used to look at the pathways upregulated from the end products of that metabolic pathway. However consistency between the observations in RNAseq data and the metabolomic data is key, one for showing reproducibility of the data, and two to improve any analysis derived using both methods in conjunction.

However in this instance observations seen in both the metabolomic and transcriptomic data do correlate. Analysis of metabolomics data from infected hosts is inherently difficult due to the presence of host metabolites. However using pathways upregulated in both transcriptomic and metabolomic data can reduce this issue.

The QPCR and microscopy data within Chapter 3 demonstrated that B17 infections contained predominantly non-proliferative stage bloodstream forms, although proliferative stages were also present. Due to the nature of these stumpy forms, we

would anticipate them to have reduced metabolic activities, although they are known to be more immunogenic to the host. This is supported by both the transcriptomic data and metabolomic data. There are many more highly abundant metabolites seen in B17 infections, however this appears to be the result of the hosts response to these more immunogenic stages, rather than an upregulation in the pathways of the parasite. Several metabolites of host origin show the damaging effect of the host mounting an immune response against the parasite, with metabolites consistent with liver damage present among others. However in the RNAseq data, a relatively equal number of genes were differentially expressed, with neither parasite strain illustrating unusually high levels of differential expression.

There were several overlaps in the metabolomic data and transcriptomic data, which were consistent; the most important of these is glycolysis. Glycolysis was shown to be upregulated in both metabolomic and transcriptomic data. Other consistencies included the upregulation of glycine, serine and threonine metabolism and C5 branched dibasic acid metabolism, which was found to be higher in the B17 strain in both sets of data. Tryptophan metabolism was also found to be higher in B17 infections in both datasets, with increased amounts of indole-3-ethanol (tryptophol) found in metabolomic data and indole- containing metabolism in B17's DEGs. Tryptophan is tryptophol's precursor, and tryptophol is used by the parasite to induce sleep within its mammalian host. Cross overs in the RNAseq and genomic data from Chapter 3 were also seen, with differential expression in the transferrin receptor in B17 in the RNAseq data, and deleterious SNPs within a ferric transporter in the genomic data.

4.4 Discussion and concluding remarks

Both the metabolomic and transcriptomic data were useful in providing more insight into the mechanistic differences in chronic and acute *T. brucei* infection. The metabolomics data was useful in determining several things, one of which was that for these two strains, and with only a pre-infection, early stage and late stage time points, each stage of the infection results in a different metabolomic signature. Each time point was taken from five individuals infected with B17 or Z310, and these signatures were reproducible across these individuals.

Importantly, each strain was distinctive enough that they gave rise to a different metabolomic signature, and B17 and Z310 infected individuals clustered separately.

From the disease manifestation, you would presume that the metabolites of infected individuals would vary significantly from pre-infected individuals, and this is observed. The pre-infection metabolomic profile is very different to the post infection profile. Furthermore the metabolomic profiles of both the early and late stages are also significantly different and cluster separately. Early infections clustered closely in both strains, however due to ill health and host response, the later stage of infection results in a more variable metabolomic profile than seen in the earlier stage of the disease.

Metabolites of interest can be identified by looking at individual metabolites or groups of metabolites belonging to pathways. In this instance, both ways were applied. In individual analysis confidently identified metabolites with the highest abundance and metabolites with the largest differences between stages were used and in pathway analysis the upregulation of pathways was derived from the abundances of several metabolites from the same pathway, to look at the overall state of regulation.

This chapter also demonstrated that RNAseq can be done directly from infected host blood samples and yield a good depth of coverage and a high percentage of reads belonging to the parasite. However these samples were taken from high parasitaemia infections at the first peak of parasitaemia, and so the percentage of usable reads from a low parasitaemia infection may yield unusable data. For each strain there were five biological replicates, and the correlation in gene counts between the libraries were very consistent.

Despite being very genetically similar both B17 and Z310 could be separated based on fold changes between genes. A high number of genes were found to be transcriptionally active, but only a small proportion of these were differentially expressed; the number DEGs in each strain were approximately equal. The distribution in terms of how many DEGs had a certain level of fold change was also relatively equal.

In B17, the DEGs with the highest fold change were found in the telomeric regions and were primarily VSGs. In Z310 strains, this was more varied, with highly DEGs found in more centric regions of the chromosome. Unfortunately, many of these highly expressed centrally located genes had no assigned functions. However the most interesting patterns of DEGs over the entire genome were found on chromosomes 3, 9 and 11.

KEGG analysis showed that the 377 DEGs could be allocated to 9 KEGG pathways, one pathway, glycolysis and gluconeogenesis, was enriched in both strains but at different parts of this pathway. 7 of the 9 KEGG pathways were enriched for in the B17 strain, 2 in the Z310 strain.

GO term analysis on the DEGs gave similar findings, with multiple functional groups assigned to DEGs but only a few were significantly enriched. As with the KEGG analysis it shows that gluconeogenesis and carbohydrate metabolism was enriched in Z310, and it also reflected that the highest fold change DEGs were in VSGs, with antigenic variation enriched for significantly, particularly in B17, where there was a significantly higher number of VSGs upregulated. This higher number of VSGs in a highly non-proliferative population could suggest either that the remaining long slender forms in the B17 infection switch at a higher frequency than seen in Z310 or that the Z310 parasites switch particularly slowly. However as previously discussed, this is more likely to be an artefact of mapping and differences in Tb927's catalogue of VSGs compared to these strains. There were also several genes, which were of interest that did not have a high fold change, but were differentially upregulated and had previously been related to virulence or drug resistance.

The transcriptomic data and metabolomic data were consistent and both showed the upregulation of glycolysis in both strains, amongst other observations discussed in the previous section. However having analysis combined from both data sets was invaluable, with metabolomics providing additional information on the host response, and transcriptomics for increasing the data resolution which enables you to look at individual genes that are being differentially expressed rather than just the pathways that have been upregulated. Using both methods also helps validate the observations seen.

CHAPTER 5

FINAL DISCUSSION

This project aimed to build on previous knowledge of two phenotypically distinct *T.b. rhodesiense* strains, in an attempt to elucidate the genetic mechanisms driving the differences in disease manifestation in both natural and experimental infection. As previously stated, virulence is not easily characterized, and in parasitic infections attempts predominantly rely on using key phenotypic differences rather than understanding the genetic mechanisms. Previous studies phenotyped these strains based on prepatent period differences, chancre presentation, and iso-enzyme differences, however it wasn't until one of the B17 and Z310 strains were sequenced by (Goodhead et al., 2013), that a greater understanding of the underlying molecular cause for this disparity in phenotype could be greater understood. As previously mentioned, due to both the reduction in cost and chemistry improvements, parasitology projects are being guided away from these traditional low throughput methodologies towards OMIC technologies and from studying strains in isolation to population studies.

This project utilized technological advances to decipher these strains at a genomic, transcriptomic and metabolomic level, in an attempt to increase the power of these analyses by using all methods in combination. Individually, these technologies have provided us with great insights into the genetic mechanisms in *T. brucei*, however in combination they have the potential for us to thoroughly interrogate datasets. Discussed beneath are the key insights gained through this project.

5.1 Developing a method for sequencing directly from clinical samples

Chapter two discussed the feasibility of sequencing directly from blood in place of whole genome sequencing from cultured parasites. This was done through the design of an enrichment array covering approximately 12% of the Tb927 genome, and nesting a subsequent enrichment design within the original target region. Minute volumes of blood were applied to Whatman FTA™ cards and successfully enriched and sequenced. As previously mentioned, this not only circumvents multiple logistical issues arising from the CAT3 nature of the parasite, but also allows for the sequencing of previously

catalogued samples, and the preparation of field samples for sequencing, which have primarily been collected using Whatman FTA™ cards.

The first capture region design was limited by the annotation of the Tb927 genome at the time of design, with subsequent re-annotation of several of the genes included within the first design. This was reflected in the probes ordered for the second design and allowed for capture of the UTR regions of genes, which may have not been fully targeted in the first design. The availability of whole genome sequence data for two of these strains enabled the validation of variants and the robustness of enrichment sequencing to be investigated. Seven strains were used in the first design, sixteen in the second, however only the two which whole genome sequence data was available for could be analyzed in this way. All seven strains enriched well within the first design, and the strains used in both designs were also successfully enriched. Unfortunately, despite enrichment, strain B17 did perform poorly in the second design compared to the other six strains used in both. This analysis could've been improved if WGS data was available for the other strains used in these designs.

The enrichment data was benchmarked against WGS data using a number of parameters, which examined whether the enrichment data was as sensitive in terms of the number of variants called, whether zygosity was conserved and correctly assigned, and whether allelic drop out was observed. A greater number of SNPs were identified in the enrichment data compared to the WGS data, indicating a greater sensitivity, which is what we would anticipate from Illumina versus SOLiD chemistry. Despite the parasite originally representing only less than 0.1% of the sample, not only was the enrichment substantial, but the data did not appear to suffer from allelic drop out effects despite the amplification required both from WGA and library preparation. Nesting a second design within the first design also illustrated that design can have a significant impact on the evenness of coverage and reduce off-target effects.

In Chapter 3, samples from low parasitaemia human infections were sequenced using the methodology described in Chapter 2. Due to their lower parasitaemias, these libraries did not enrich sufficiently to make the datasets usable, with the exception of strain G7, which demonstrated the ratio of parasite to host DNA is more important than the quantity of DNA. The un-enriched parasite to host DNA ratio was significantly lower than seen in the experimental infections, and the resulting percentage of parasite DNA in the enriched data still resulted in a significant enrichment. However, this still

represented only a small proportion of the dataset (~1%) and so a greater depth of sequencing would be required to make use of this data.

Overall, enrichment data from Chapters 2 and 3 showed that providing the parasitaemia of the infection was not too low, enrichment data could reliably be used in place of WGS data for resolving parasite DNA from an infected host sample.

5.2 Using multiple strains to understand zymodeme group structure and identify mutations potentially associated with virulence

Chapter 3 discussed the data generated using the methods described in Chapter 2 for seven strains, which were used both in the first and second capture design. Three strains were available for zymodeme groups B17 and Z310, however only one strain was available for the intermediate phenotype. Strain availability is the main limiting factor in this analysis, as more strains would make inferences on variants seen in individuals and variants conserved across zymodeme groups more valid. Due to the lack of a *T.b. rhodesiense* reference and the inferior quality of the DAL reference, determining the biologically relevant variation seen and separating this from the natural accumulation of mutations resulting from genetic drift, is difficult. The analysis is also limited by the data being generated from a small proportion of the genome and the high percentage similarity between strains. If the strains were more dissimilar and WGS data was available in place of enrichment data, or the target region was larger, a greater number of non-conserved SNPs could be used to understand inter/intra zymodeme variation.

The microscopy and QPCR data within this chapter illustrated striking phenotypic differences between these strains, and indicated that differences in differentiation may be responsible for this. Deleterious SNPs found to be unique to each zymodeme group suggested potential variants that could be causing this difference in phenotype, but no clear causal mutation could be identified. However multiple SNPs were found in the transmembrane component of ferric reductase, which suggested iron regulation was potentially important in generating this phenotype. This idea was further supported in Chapter 4, by the greater than ten fold increase in transferrin receptor expression in B17. Data within this chapter indicated that the phenotypic differences observed were likely to arise through differential regulation of transcripts rather than genomic variants.

5.3 Understanding host-parasite interaction through the combined analysis of metabolomic and transcriptomic data

Chapter 4 used both transcriptomic and metabolomic analysis to further phenotype these strains. Variants called from the genomic data and the difference in the bloodstream form abundance discussed in Chapter 3 indicated that few genotypic differences were responsible for the severe difference in phenotype observed and that differential regulation was more likely to be driving the phenotypes generated. As suggested by the genomic data, only a small subset of genes were differentially regulated. Not all genes are transcriptionally active and there are substantially fewer metabolites, which means that smaller genomic changes are often amplified downstream in the metabolomic data.

Extracting biologically relevant observations from the transcriptomic and metabolomic data is hindered by the presence of host metabolites and by the nature of LC-MS, which means that many metabolites can not be identified with complete certainty. However by using transcriptomic data in conjunction, I could identify pathways that were differentially upregulated in the RNAseq data and the metabolomic data, and observe whether SNPs were also identified in these genes. The highest differential expression was observed primarily in VSG genes, however as previously discussed, this is more likely to be an artefact of mapping to the Tb927 reference than biologically relevant. However multiple differentially expressed genes were also identified which had previously been associated with virulence.

The RNAseq data and metabolomic data were largely consistent and showed upregulation of multiple pathways that could be related back to the phenotypic differences observed, with differential regulation in pathways associated with growth observed. The high fold change seen in the transferrin receptor in B17 again suggested iron regulation may be partially responsible for these phenotypic differences. However greater conclusions could have been drawn if transcriptomic and metabolomic data was available for more than the first peak of parasitaemia, as RNAseq libraries generated from samples taken at the late stage of disease could provide useful insights into the regulatory differences resulting in either maintaining a chronic infection or acute.

5.4 Final conclusions

The combined use of multiple OMIC approaches within this body of work has highlighted several potential mechanistic differences, which could be driving the disparity in virulence observed in these strains. No causal variant or regulatory process regulating this phenotype can conclusively be identified from this work, however multiple deleterious variants identified within Chapter 3 and differentially regulated genes identified in Chapter 4, suggest several mechanisms which may be driving this. The microscopy and QPCR data highlighted the extreme differences in bloodstream form abundance observed in these two strains and suggested mechanisms regulating differentiation could be regulating the phenotype. Stumpy forms predominating B17 infections correlates with the chronic nature of the strain and chancre presence observed because stumpy forms illicit a greater immune response, and have reduced motility, which could lead to the chancre formation. This also agrees with the metabolomic data, because despite the reduced metabolic capabilities of stumpy forms, a higher abundance of metabolites are observed compared to Z310 infection, suggesting host response. Iron is essential for parasite growth, and the high density of SNPs found in iron regulatory pathways and the high fold changes observed in the RNAseq data suggest that this is one of the key differentially regulated pathways involved.

References

Abdulla, M.-H., O'Brien, T., Mackey, Z. B., Sajid, M., Grab, D. J., and McKerrow, J. H. (2008). RNA interference of *Trypanosoma brucei* cathepsin B and L affects disease progression in a mouse model. *PLoS Negl Trop Dis* 2, e298. doi:10.1371/journal.pntd.0000298.

Afework, Y., Mäser, P., Etschmann, B., Samson-Himmelstjerna, von, G., Zessin, K.-H., and Clausen, P.-H. (2006). Rapid identification of isometamidium-resistant stocks of *Trypanosoma b. brucei* by PCR-RFLP. *Parasitol Res* 99, 253–261. doi:10.1007/s00436-006-0141-z.

Ahmed, H. A., MacLeod, E. T., Hide, G., Welburn, S. C., and Picozzi, K. (2011). The best practice for preparation of samples from FTA. *Parasites & Vectors* 4, 68. doi:10.1186/1756-3305-4-68.

Aitcheson, N., Talbot, S., Shapiro, J., Hughes, K., Adkin, C., Butt, T., Sheader, K., and Rudenko, G. (2005). VSG switching in *Trypanosoma brucei*: antigenic variation analysed using RNAi in the absence of immune selection. *Molecular Microbiology* 57, 1608–1622. doi:10.1111/j.1365-2958.2005.04795.x.

Al-Khodor, S., Price, C. T., Kalia, A., and Abu Kwaik, Y. (2010). Functional diversity of ankyrin repeats in microbial proteins. *Trends Microbiol.* 18, 132–139. doi:10.1016/j.tim.2009.11.004.

Ali, J. A. M., Creek, D. J., Burgess, K., Allison, H. C., Field, M. C., Mäser, P., and de Koning, H. P. (2013a). Pyrimidine salvage in *Trypanosoma brucei* bloodstream forms and the trypanocidal action of halogenated pyrimidines. *Mol. Pharmacol.* 83, 439–453. doi:10.1124/mol.112.082321.

Ali, J. A. M., Tagoe, D. N. A., Munday, J. C., Donachie, A., Morrison, L. J., and de Koning, H. P. (2013b). Pyrimidine Biosynthesis Is Not an Essential Function for *Trypanosoma brucei* Bloodstream Forms. *PLoS ONE* 8, e58034. doi:10.1371/journal.pone.0058034.s003.

Alizon, S., and Lion, S. (2011). Within-host parasite cooperation and the evolution of virulence. *Proceedings of the Royal Society B: Biological Sciences* 278, 3738–3747. doi:10.1098/rspb.2011.0471.

Alsford, S., Eckert, S., Baker, N., Glover, L., Sanchez-Flores, A., Leung, K. F., Turner, D. J., Field, M. C., Berriman, M., and Horn, D. (2012). High-throughput decoding of antitrypanosomal drug efficacy and resistance. *Nature* 482, 232–U125. doi:10.1038/nature10771.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol* 11, R106. doi:10.1186/gb-2010-11-10-r106.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* 8, 1765–1786. doi:10.1038/nprot.2013.099.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638.

Anderson, J., Fuglsang, H & de Marshall, T,F (1976). Effects of suramin on ocular onchocerciasis. *Tropenmed Parasitol* 27, 279–296.

Anderson, N. E., Mubanga, J., Fevre, E. M., Picozzi, K., Eisler, M. C., Thomas, R., and Welburn, S. C. (2011). Characterisation of the wildlife reservoir community for human and animal trypanosomiasis in the Luangwa Valley, Zambia. *PLoS Negl Trop Dis* 5, e1211. doi:10.1371/journal.pntd.0001211.

Arneson, N., Hughes, S., Houlston, R., and Done, S. (2008). GenomePlex Whole-Genome Amplification. *CSH Protoc* 2008, pdb.prot4920.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29. doi:10.1038/75556.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Gajria, B., Gao, X., et al. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* 38, D457–62. doi:10.1093/nar/gkp851.

Babokhov, P., Sanyaolu, A. O., Oyibo, W. A., Fagbenro-Beyioku, A. F., and Iriemenam, N. C. (2013). A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis. *Pathog Glob Health* 107, 242–252. doi:10.1179/2047773213Y.0000000105.

Bacchi, C. J., Garofalo, J., Mockenhaupt, D., McCann, P. P., Diekema, K. A., Pegg, A. E., Nathan, H. C., Mullaney, E. A., Chunosoff, L., Sjoerdsma, A., et al. (1983). Invivo Effects of Alpha-Dl-Difluoromethylornithine on the Metabolism and Morphology of *Trypanosoma-Brucei-Brucei*. *Mol. Biochem. Parasitol.* 7, 209–225.

Bainbridge, M. N., Wang, M., Burgess, D. L., Kovar, C., Rodesch, M. J., D'Ascenzo, M., Kitzman, J., Wu, Y.-Q., Newsham, I., Richmond, T. A., et al. (2010). Whole exome capture in solution with 3 Gbp of data. *Genome Biol* 11, R62. doi:10.1186/gb-2010-11-6-r62.

Balmer, O., Beadell, J. S., Gibson, W., and Caccone, A. (2011). Phylogeography and Taxonomy of *Trypanosoma brucei*. *PLoS Negl Trop Dis* 5, e961. doi:10.1371/journal.pntd.0000961.

Banuls, A. L., Brisse, S., Sidibé, I., Noël, S., and TIBAYRENC, M. (1999). A phylogenetic analysis by multilocus enzyme electrophoresis and multiprimer random amplified polymorphic DNA fingerprinting of the Leishmania genome project Friedlin reference strain. *Folia Parasitol.* 46, 10–14.

Baptista-Fernandes, T., Marques, C., Roos Rodrigues, O., and Santos-Gomes, G. M. (2007). Intra-specific variability of virulence in *Leishmania infantum* zymodeme MON-1 strains. *Comparative Immunology, Microbiology and Infectious Diseases* 30, 41–53. doi:10.1016/j.cimid.2006.10.001.

Barrett, M. P., Boykin, D. W., Brun, R., and Tidwell, R. R. (2007). Human African trypanosomiasis: pharmacological re-engagement with a neglected disease. *Br. J. Pharmacol.* 152, 1155–1171. doi:10.1038/sj.bjp.0707354.

Barry, J. D., and McCulloch, R. (2001). Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Adv. Parasitol.* 49, 1–70.

Beament, T. (2002). Investigation into differences in pathogenesis of human isolates of African Trypanosomiasis in mice. Thesis. University of Liverpool.

Ben Abderrazak, S., Guerrini, F., Mathieu-Daudé, F., Truc, P., Neubauer, K., Lewicka, K., Barnabé, C., and TIBAYRENC, M. (1993). Isoenzyme electrophoresis for parasite characterization. *Methods Mol. Biol.* 21, 361–382. doi:10.1385/0-89603-239-6:361.

Benaim, G., Lopez-Estraño, C., Docampo, R., and Moreno, S. N. (1993). A calmodulin-stimulated Ca2+ pump in plasma-membrane vesicles from *Trypanosoma brucei*; selective inhibition by pentamidine. *Biochem. J.* 296 (Pt 3), 759–763.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi:10.1038/nature07517.

Berriman, M. (2005). The Genome of the African Trypanosome *Trypanosoma brucei*. *Science* 309, 416–422. doi:10.1126/science.1112642.

Berriman, M., Hall, N., Sheader, K., Bringaud, F., Tiwari, B., Isobe, T., Bowman, S., Corton, C., Clark, L., Cross, G. A. M., et al. (2002). The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 122, 131–140. doi:10.1016/S0166-6851(02)00092-0.

Bertram, M. A., Meyer, E. A., Lile, J. D., and Morse, S. A. (1983). A comparison of isozymes of five axenic *Giardia isolates*. *J. Parasitol.* 69, 793–801.

Beschin, A., Van Den Abbeele, J., de Baetselier, P., and Pays, E. (2014). African trypanosome control in the insect vector and mammalian host. *Trends in Parasitology* 30, 538–547. doi:10.1016/j.pt.2014.08.006.

Black, S. J., Jack, R. M., and Morrison, W. I. (1983). Host-parasite interactions which influence the virulence of *Trypanosoma* (Trypanozoon) *brucei brucei* organisms. *Acta Tropica* 40, 11–18.

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., Kieleczawa, J., Lyons, R. H., Neubert, T. A., Noll, A. C., et al. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech* 24, 73–86. doi:10.7171/jbt.13-2402-002.

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T., et al. (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* 400, 532–538. doi:10.1038/22964.

Brun, R., Blum, J., Chappuis, F., and Burri, C. (2010). Human african trypanosomiasis. *The Lancet*.

Bucheton, B., MacLeod, A., and Jamonneau, V. (2011). Human host determinants influencing the outcome of *Trypanosoma brucei gambiense* infections. *Parasite Immunology* 33, 438–447. doi:10.1111/j.1365-3024.2011.01287.x.

Bursell, E. (1981). Energetics of Hematophagous Arthropods - Influence of Parasites. *Parasitology* 82, 107–108.

Canduri, F., Cardoso Perez, P., Caceres, R. A., and de Azevedo, W. F. J. (2007). Protein kinases as targets for antiparasitic chemotherapy drugs. *Curr Drug Targets* 8, 389–398.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi:10.1101/gr.6743907.

Canuto, G. A. B., Castilho-Martins, E. A., Tavares, M., López-Gonzálvez, A., Rivas, L., and Barbas, C. (2012). CE-ESI-MS metabolic fingerprinting of *Leishmania* resistance to antimony treatment. *Electrophoresis* 33, 1901–1910. doi:10.1002/elps.201200007.

Capewell, P., Clucas, C., DeJesus, E., Kieft, R., Hajduk, S., Veitch, N., Steketee, P. C., Cooper, A., Weir, W., and MacLeod, A. (2013a). The TgsGP Gene Is Essential for Resistance to Human Serum in *Trypanosoma brucei gambiense*. *PLoS Pathog* 9, e1003686. doi:10.1371/journal.ppat.1003686.

Capewell, P., Cooper, A., Duffy, C. W., Tait, A., and Turner, C. (2013b). Human and animal trypanosomes in Côte d'Ivoire form a single breeding population. *PLoS ONE*.

Capewell, P., Veitch, N. J., Turner, C. M. R., Raper, J., Berriman, M., Hajduk, S. L., and MacLeod, A. (2011). Differences between *Trypanosoma brucei gambiense* Groups 1 and 2 in Their Resistance to Killing by Trypanolytic Factor 1. *PLoS Negl Trop Dis* 5, e1287. doi:10.1371/journal.pntd.0001287.s005.

Caruccio, N. (2011). Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods Mol. Biol.* 733, 241–255. doi:10.1007/978-1-61779-089-8_17.

Chappuis, F. (2007). Melarsoprol-free drug combinations for second-stage gambian sleeping sickness: The way to go. *Clin. Infect. Dis.* 45, 1443–1445. doi:10.1086/522983.

Chappuis, F., Loutan, L., Simarro, P., Lejon, V., and Büscher, P. (2005). Options for field diagnosis of human african trypanosomiasis. *Clinical Microbiology Reviews* 18, 133–146. doi:10.1128/CMR.18.1.133-146.2005.

Chappuis, F., Stivanello, E., Adams, K., Kidane, S., Pittet, A., and Bovier, P. A. (2004). Card agglutination test for trypanosomiasis (CATT) end-dilution titer and cerebrospinal fluid cell count as predictors of human African Trypanosomiasis (*Trypanosoma brucei gambiense*) among serologically suspected individuals in southern Sudan. *Am. J. Trop. Med. Hyg.* 71, 313–317.

Chaudhri, M., Steverding, D., Kittelberger, D., Tjia, S., and Overath, P. (1994). Expression of a glycosylphosphatidylinositol-anchored *Trypanosoma brucei* transferrin-binding protein complex in insect cells. *Proc. Natl. Acad. Sci. U.S.A.* 91, 6443–6447.

Cheeseman, I. H., McDew-White, M., Phyo, A. P., Sriprawat, K., Nosten, F., and Anderson, T. J. C. (2014). Pooled Sequencing and Rare Variant Association Tests for Identifying the Determinants of Emerging Drug Resistance in Malaria Parasites. *Molecular Biology and Evolution*. doi:10.1093/molbev/msu397.

Chevalier, N., Rigden, D. J., Van Roy, J., Opperdoes, F. R., and Michels, P. (2000). *Trypanosoma brucei* contains a 2,3-bisphosphoglycerate independent phosphoglycerate mutase. *Eur. J. Biochem.* 267, 1464–1472.

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., and Meza-Zepeda,

L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 15, 449. doi:10.1186/1471-2164-15-449.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., and Lu, X. (2012a). Using *Drosophila melanogaster* as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet* 3, 35. doi:10.3389/fgene.2012.00035.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012b). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi:10.4161/fly.19695.

Coppens, I., and Courtoy, P. J. (2000). The adaptative mechanisms of *Trypanosoma brucei* for sterol homeostasis in its different life-cycle environments. *Annu. Rev. Microbiol.* 54, 129–156. doi:10.1146/annurev.micro.54.1.129.

Cox, A., Tilley, A., McOdimba, F., Fyfe, J., Eisler, M., Hide, G., and Welburn, S. (2005). A PCR based assay for detection and differentiation of African trypanosome species in blood. *Exp. Parasitol.* 111, 24–29. doi:10.1016/j.exppara.2005.03.014.

Creek, D. J., Anderson, J., McConville, M. J., and Barrett, M. P. (2012a). Metabolomic analysis of trypanosomatid protozoa. *Mol. Biochem. Parasitol.* 181, 73–84. doi:10.1016/j.molbiopara.2011.10.003.

Creek, D. J., Jankevics, A., Burgess, K. E. V., Breitling, R., and Barrett, M. P. (2012b). IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data. *Bioinformatics* 28, 1048–1049. doi:10.1093/bioinformatics/bts069.

Cross, G. A., and Manning, J. C. (1973). Cultivation of *Trypanosoma brucei* sspp. in semi-defined and defined media. *Parasitology* 67, 315–331.

Dacks, J. B., and Doolittle, W. F. (2002). Novel syntaxin gene sequences from *Giardia, Trypanosoma* and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *Journal of Cell Science* 115, 1635–1642.

Daily, J. P., Le Roch, K. G., Sarr, O., Ndiaye, D., Lukens, A., Zhou, Y., Ndir, O., Mboup, S., Sultan, A., Winzeler, E. A., et al. (2005). In vivo transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. *J. Infect. Dis.* 191, 1196–1203. doi:10.1086/428289.

Damodaran, S., and Kinsella, J. E. (1983). Dissociation of nucleoprotein complexes by chaotropic salts. *FEBS Lett.* 158, 53–57.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330.

Daniels, J.-P., Gull, K., and Wickstead, B. (2010). Cell biology of the trypanosome genome. *Microbiol. Mol. Biol. Rev.* 74, 552–569. doi:10.1128/MMBR.00024-10.

Das, A., Banday, M., and Bellofatto, V. (2008). RNA Polymerase Transcription Machinery in Trypanosomes. *Eukaryotic Cell* 7, 429–434. doi:10.1128/EC.00297-07.

de Atouguia, J. L. M., and Kennedy, P. G. E. (2000). Neurological aspects of human African trypanosomiasis.

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261–5266. doi:10.1073/pnas.082089499.

Dean, P., Major, P., Nakjang, S., Hirt, R. P., and Embley, T. M. (2014). Transport proteins of parasitic protists and their role in nutrient salvage. *Front Plant Sci* 5. doi:10.3389/fpls.2014.00153.

DeJesus, E., Kieft, R., Albright, B., Stephens, N. A., and Hajduk, S. L. (2013). A single amino acid substitution in the group 1 Trypanosoma brucei gambiense haptoglobin-hemoglobin receptor abolishes TLF-1 binding. *PLoS Pathog* 9, e1003317. doi:10.1371/journal.ppat.1003317.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491–. doi:10.1038/ng.806.

Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrom Rev* 26, 51–78. doi:10.1002/mas.20108.

Dhalia, R., Reis, C. R. S., Freire, E. R., Rocha, P. O., Katz, R., Muniz, J. R. C., Standart, N., and de Melo Neto, O. P. (2005). Translation initiation in *Leishmania major*: characterisation of multiple eIF4F subunit homologues. *Mol. Biochem. Parasitol.* 140, 23–41. doi:10.1016/j.molbiopara.2004.12.001.

Dhaliwal, , B.B.S & Juyal, P.D. (2013). *Parasitic Zooneses.*.New Dehli. Springer.

Dieterle, F., Riefke, B., Schlotterbeck, G., Ross, A., Senn, H., and Amberg, A. (2011). NMR and MS methods for metabonomics. *Methods Mol. Biol.* 691, 385–415. doi:10.1007/978-1-60761-849-2_24.

Docampo, R., and Moreno, S. (2001). The acidocalcisome. *Mol. Biochem. Parasitol.* 114, 151–159.

Docampo, R., and Moreno, S. N. J. (2011). Acidocalcisomes. *Cell Calcium* 50, 113–119. doi:10.1016/j.ceca.2011.05.012.

Donelson, J. E. (2003). Antigenic variation and the African trypanosome genome. *Acta Tropica* 85, 391–404.

Doua, F., Miezan, T. W., Sanon Singaro, J. R., Boa Yapo, F., and Baltz, T. (1996). The efficacy of pentamidine in the treatment of early-late stage *Trypanosoma brucei gambiense* trypanosomiasis. *Am. J. Trop. Med. Hyg.* 55, 586–588.

Downing, T., Imamura, H., Decuypere, S., Clark, T. G., Coombs, G. H., Cotton, J. A., Hilley, J. D., de Doncker, S., Maes, I., Mottram, J. C., et al. (2011). Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance.

Drain, J., Bishop, J. R., and Hajduk, S. L. (2001). Haptoglobin-related protein mediates trypanosome lytic factor binding to trypanosomes. *J. Biol. Chem.* 276, 30254–30260.

Duszenko, M., Figarella, K., MacLeod, E. T., and Welburn, S. C. (2006). Death of a trypanosome: a selfish altruism. *Trends in Parasitology* 22, 536–542. doi:10.1016/j.pt.2006.08.010.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113. doi:10.1186/1471-2105-5-113.

El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A. N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., et al. (2005a). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309, 409–415. doi:10.1126/science.1112631.

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C., et al. (2005b). Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309, 404–409. doi:10.1126/science.1112181.

El-Sayed, N., Ghedin, E., Song, J. M., MacLeod, A., Bringaud, F., Larkin, C., Wanless, D., Peterson, J., Hou, L. H., Taylor, S., et al. (2003). The sequence and analysis of *Trypanosoma brucei* chromosome II. *Nucleic Acids Research* 31, 4856–4863. doi:10.1093/nar/gkg673.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol. (Amst.)* 29, 51–63. doi:10.1016/j.tree.2013.09.008.

Emmer, B. T., Daniels, M. D., Taylor, J. M., Epting, C. L., and Engman, D. M. (2010). Calflagin Inhibition Prolongs Host Survival and Suppresses Parasitemia in *Trypanosoma brucei* Infection. *Eukaryotic Cell* 9, 934–942. doi:10.1128/EC.00086-10.

Erben, E. D., Fadda, A., Lueong, S., Hoheisel, J. D., and Clayton, C. (2014). A Genome-Wide Tethering Screen Reveals Novel Potential Post-Transcriptional Regulators in *Trypanosoma brucei*. *PLoS Pathog* 10. doi:10.1371/journal.ppat.1004178.

Fairlamb, A. H. (1990). Future prospects for the chemotherapy of human trypanosomiasis. 1. Novel approaches to the chemotherapy of trypanosomiasis. *Trans. R. Soc. Trop. Med. Hyg.* 84, 613–617.

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Fenn, K., and Matthews, K. R. (2007). The cell biology of *Trypanosoma brucei* differentiation. *Curr. Opin. Microbiol.*

Ferragina, P., and Manzini, G. (2001). An experimental study of a compressed index. *Information Sciences* 135, 13–28.

Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., Ursos, L. M., Sidhu, A. B., Naudé, B., Deitsch, K. W., et al. (2000). Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* 6, 861–871.

Flynn, I. W., and Bowman, I. B. (1970). Comparative Biochemistry of Monomorphic and Pleomorphic Strains of Trypanosoma-Rhodesiense. *Trans. R. Soc. Trop. Med. Hyg.* 64, 175–&.

Flynn, I. W., and Bowman, I. B. (1973). The metabolism of carbohydrate by pleomorphic African trypanosomes. *Comp. Biochem. Physiol., B* 45, 25–42.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169–3177. doi:10.1093/bioinformatics/bts605.

Forrester, S. J., and Hall, N. (2014). The revolution of whole genome sequencing to study parasites. *Mol. Biochem. Parasitol.*

Fowler, K. E., Reitter, C. P., Walling, G. A., and Griffin, D. K. (2012). Novel approach for deriving genome wide SNP analysis data from archived blood spots. *BMC Res Notes* 5, 503. doi:10.1007/s00253-010-2926-3.

Funk, S., Nishiura, H., Heesterbeek, H., and Edmunds, W. J. (2013). Identifying transmission cycles at the human-animal interface: the role of animal reservoirs in maintaining gambiense human african trypanosomiasis. *PLoS Comput Biol*.

Galyov, E. E., Håkansson, S., Forsberg, A., and Wolf-Watz, H. (1993). A secreted protein kinase of *Yersinia pseudotuberculosis* is an indispensable virulence determinant. *Nature* 361, 730–732. doi:10.1038/361730a0.

Garcia-Salcedo, J. A., Perez-Morga, D., Gijon, P., Dilbeck, V., Pays, E., and Nolan, D. P. (2004). A differential role for actin during the life cycle of *Trypanosoma brucei*. *The EMBO Journal* 23, 780–789. doi:10.1038/sj.emboj.7600094.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511. doi:10.1038/nature01097.

Gargantini, P. R., Lujan, H. D., and Pereira, C. A. (2012). In silico analysis of trypanosomatids' helicases. *FEMS Microbiol. Lett.* 335, 123–129. doi:10.1111/j.1574-6968.2012.02644.x.

Gehrig, S., and Efferth, T. (2008). Development of drug resistance in *Trypanosoma brucei rhodesiense* and Trypanosoma brucei gambiense. Treatment of human African trypanosomiasis with natural products (Review). *Int. J. Mol. Med.* 22, 411–419. doi:10.3892/ijmm_00000037.

GE Healthcare. (2010a). Amplification of human genomic DNA from blood on FTA™ with Genomiphi™. Application Note 51638.

GE Healthcare. (2010b) Comparative analysis of FTA™ and NucleoSave™ cards. Application note 28-9822-24AA.

GE Healthcare. (2010c) FTA™ cards. Data File 51613.

Geysen, D., Delespaux, V., and Geerts, S. (2003). PCR-RFLP using Ssu-rDNA amplification as an easy method for species-specific diagnosis of Trypanosoma species in cattle. *Vet. Parasitol.* 110, 171–180.

Gibson, W. C. (1986). Will the real *Trypanosoma b. gambiense* please stand up. *Parasitol. Today (Regul. Ed.)* 2, 255–257.

Gibson, W., and Bailey, M. (2003). The development of *Trypanosoma brucei* within the tsetse fly midgut observed using green fluorescent trypanosomes. *Kinetoplastid Biol Dis* 2, 1.

Gibson, W., and Stevens, J. (1999). "Genetic Exchange in the Trypanosomatidae," in *Advances in Parasitology* Advances in Parasitology. (Elsevier), 1–46. doi:10.1016/S0065-308X(08)60240-7.

Gibson, W., Nemetschke, L., and Ndung'u, J. (2010). Conserved sequence of the TgsGP gene in Group 1 Trypanosoma brucei gambiense. *Infect. Genet. Evol.* 10, 453–458. doi:10.1016/j.meegid.2010.03.005.

Gibson, W., Peacock, L., Ferris, V., Fischer, K., Livingstone, J., Thomas, J., and Bailey, M. (2015). Genetic recombination between human and animal parasites creates novel strains of human pathogen. *PLoS Negl Trop Dis* 9, e0003665. doi:10.1371/journal.pntd.0003665.

Giroud, C., Ottones, F., Coustou, V., Dacheux, D., Biteau, N., Miezan, B., Van Reet, N., Carrington, M., Doua, F., and Baltz, T. (2009). Murine Models for *Trypanosoma brucei gambiense* Disease Progression-From Silent to Chronic Infections and Early Brain Tropism. *PLoS Negl Trop Dis* 3. doi:10.1371/journal.pntd.0000509.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189. doi:10.1038/nbt.1523.

Gobert, A. P., Daulouede, S., Lepoivre, M., Boucher, J. L., Bouteille, B., Buguet, A., Cespuglio, R., Veyret, B., and Vincendeau, P. (2000). L-arginine availability modulates local nitric oxide production and parasite killing in experimental trypanosomiasis. *Infection and Immunity* 68, 4653–4657.

Goodhead, I., Capewell, P., Bailey, J. W., Beament, T., Chance, M., Kay, S., Forrester, S., MacLeod, A., Taylor, M., Noyes, H., et al. (2013). Whole-genome sequencing of *Trypanosoma brucei* reveals introgression between subspecies that is associated with virulence. *MBio* 4. doi:10.1128/mBio.00197-13.

Graham, S. V., and Barry, J. D. (1991). Expression site-associated genes transcribed independently of variant surface glycoprotein genes in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 47, 31–41.

Guether, M. L. S., Urbaniak, M. D., Tavendale, A., Prescott, A., and Ferguson, M. A. J. (2014). High-Confidence Glycosome Proteome for Procyclic Form *Trypanosoma brucei* by Epitope-Tag Organelle Enrichment and SILAC Proteomics. *J. Proteome Res.* 13, 2796–2806. doi:10.1021/pr401209w.

Hajduk, S. L., Moore, D. R., Vasudevacharya, J., Siqueira, H., Torri, A. F., Tytler, E. M., and Esko, J. D. (1989). Lysis of Trypanosoma brucei by a toxic subspecies of human high density lipoprotein. *J. Biol. Chem.* 264, 5210–5217.

Hall, N., Berriman, M., Lennard, N. J., Harris, B. R., Hertz-Fowler, C., Bart-Delabesse, E. N., Gerrard, C. S., Atkin, R. J., Barron, A. J., Bowman, S., et al. (2003). The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism. *Nucleic Acids Research* 31, 4864–4873. doi:10.1093/nar/gkg674.

Hamilton, P. B., Stevens, J. R., Gaunt, M. W., Gidley, J., and Gibson, W. C. (2004). Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. *Int. J. Parasitol.* 34, 1393–1404. doi:10.1016/j.ijpara.2004.08.011.

Handyside, A. H., Robinson, M. D., Simpson, R. J., Omar, M. B., Shaw, M. A., Grudzinskas, J. G., and Rutherford, A. (2004). Isothermal whole genome amplification from single and small numbers of cells: a new era for preimplantation genetic diagnosis of inherited disease. *Mol. Hum. Reprod.* 10, 767–772. doi:10.1093/molehr/gah101.

Harrington, J. M., Widener, J., Stephens, N., Johnson, T., Francia, M., Capewell, P., MacLeod, A., and Hajduk, S.

L. (2010). The Plasma Membrane of Bloodstream-form African Trypanosomes Confers Susceptibility and Specificity to Killing by Hydrophobic Peptides. *J. Biol. Chem.* 285, 28659–28666. doi:10.1074/jbc.M110.151886.

Hellani, A., Abu-Amero, K., Azouri, J., and El-Akoum, S. (2008). Successful pregnancies after application of array-comparative genomic hybridization in PGS-aneuploidy screening. *Reprod. Biomed. Online* 17, 841–847.

Hertz-Fowler, C., Figueiredo, L. M., Quail, M. A., Becker, M., Jackson, A., Bason, N., Brooks, K., Churcher, C., Fahkro, S., Goodhead, I., et al. (2008). Telomeric expression sites are highly conserved in *Trypanosoma brucei*. *PLoS ONE* 3, e3527. doi:10.1371/journal.pone.0003527.

Hide, G., Tait, A., Maudlin, I., and Welburn, S. C. (1996). The origins, dynamics and generation of *Trypanosoma brucei rhodesiense* epidemics in East Africa. *Parasitol. Today (Regul. Ed.)* 12, 50–55.

Hirano, T. (2012). Condensins: universal organizers of chromosomes with diverse functions. *Genes & Development* 26, 1659–1678. doi:10.1101/gad.194746.112.

Hoare, C. A. (1972). The trypanosomes of mammals. A zoological monograph. *... trypanosomes of mammals: a zoological monograph*.

Hu, L., Hu, H., and Li, Z. (2012). A kinetoplastid-specific kinesin is required for cytokinesis and for maintenance of cell morphology in *Trypanosoma brucei*. *Molecular Microbiology* 83, 565–578. doi:10.1111/j.1365-2958.2011.07951.x.

Huang, Q., Lin, B., Liu, H., Ma, X., Mo, F., Yu, W., Li, L., Li, H., Tian, T., Wu, D., et al. (2011). RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS ONE* 6, e26168. doi:10.1371/journal.pone.0026168.

Hutchison, C. A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research* 35, 6227–6237. doi:10.1093/nar/gkm688.

Inoue, R., Tsukahara, T., Sunaba, C., Itoh, M., and Ushida, K. (2007). Simple and rapid detection of the porcine reproductive and respiratory syndrome virus from pig whole blood using filter paper. *Journal of Virological Methods* 141, 102–106. doi:10.1016/j.jviromet.2006.11.030.

International Glossina Genome Initiative, Attardo, G. M., Abila, P. P., Auma, J. E., Baumann, A. A., Benoit, J. B., Brelsfoard, C. L., Ribeiro, J. M. C., Cotton, J. A., Pham, D. Q. D., et al. (2014). Genome Sequence of the Tsetse Fly (*Glossina morsitans*): Vector of African Trypanosomiasis. *Science* 344, 380–386. doi:10.1126/science.1249656.

Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M. A., Adlem, E., Aert, R., et al. (2005). The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309, 436–442. doi:10.1126/science.1112680.

Jackson, A. P., Sanders, M., Berry, A., McQuillan, J., Aslett, M. A., Quail, M. A., Chukualim, B., Capewell, P., MacLeod, A., Melville, S. E., et al. (2010). The Genome Sequence of *Trypanosoma brucei gambiense*, Causative Agent of Chronic Human African Trypanosomiasis. *PLoS Negl Trop Dis* 4. doi:10.1371/journal.pntd.0000658.

Jamonneau, V., Ilboudo, H., Kabore, J., Kaba, D., Koffi, M., Solano, P., Garcia, A., Courtin, D., Laveissiere, C., Lingue, K., et al. (2012). Untreated Human Infections by *Trypanosoma brucei gambiense* Are Not 100% Fatal. *PLoS Negl Trop Dis* 6. doi:10.1371/journal.pntd.0001691.

Kaddurah-Daouk, R., Weinshilboum, R. M., Pharmacometabolomics Research Network (2014). Pharmacometabolomics: implications for clinical pharmacology and systems pharmacology. *Clin. Pharmacol. Ther.* 95, 154–167. doi:10.1038/clpt.2013.217.

Kangethe, R. T., Boulangé, A. F. V., Coustou, V., Baltz, T., and Coetzer, T. H. T. (2012). *Trypanosoma brucei brucei* oligopeptidase B null mutants display increased prolyl oligopeptidase-like activity. *Mol. Biochem. Parasitol.* 182, 7–16. doi:10.1016/j.molbiopara.2011.11.007.

Kanmogne, G. D., Stevens, J. R., Asonganyi, T., and Gibson, W. C. (1996). Characterization of *Trypanosoma brucei gambiense* isolates using restriction fragment length polymorphisms in 5 variant surface

glycoprotein genes. *Acta Tropica* 61, 239–254.

Kayang, B. B., Bosompem, K. M., Assoku, R. K., and Awumbila, B. (1997). Detection of *Trypanosoma brucei*, *T. congolense* and *T. vivax* infections in cattle, sheep and goats using latex agglutination. *Int. J. Parasitol.* 27, 83–87.

Keating, J., Yukich, J. O., Sutherland, C. S., Woods, G., and Tediosi, F. (2015). Human African trypanosomiasis prevention, treatment and control costs: A systematic review. *Acta Tropica* 150, 4–13. doi:10.1016/j.actatropica.2015.06.003.

Kennedy, P. (2004). Human African trypanosomiasis of the CNS: current issues and challenges. *Journal of Clinical Investigation*.

Kennedy, P. G. (2013). Clinical features, diagnosis, and treatment of human African trypanosomiasis (sleeping sickness). *The Lancet Neurology* 12, 186–194. doi:10.1016/S1474-4422(12)70296-X.

Kibona, S. N., Picozzi, K., Matemba, L., and Lubega, G. W. (2007). Characterisation of the *Trypanosoma brucei rhodesiense* isolates from Tanzania using serum resistance associated gene as molecular marker. *Tanzan Health Res Bull* 9, 25–31.

Kieft, R., Capewell, P., Turner, C. M. R., Veitch, N. J., MacLeod, A., and Hajduk, S. (2010). Mechanism of *Trypanosoma brucei gambiense* (group 1) resistance to human trypanosome lytic factor. *Proceedings of the National Academy of Sciences* 107, 16137–16141. doi:10.1073/pnas.1007074107.

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765–777.

Kline, M. C., Duewer, D. L., Redman, J. W., Butler, J. M., and Boyer, D. A. (2002). Polymerase chain reaction amplification of DNA from aged blood stains: quantitative evaluation of the "suitability for purpose" of four filter papers as archival media. *Anal. Chem.* 74, 1863–1869.

Koffi, M., Solano, P., Barnabe, C., de Meeûs, T., Bucheton, B., Cuny, G., and Jamonneau, V. (2007). Genetic characterisation of *Trypanosoma brucei* s.l. using microsatellite typing: New perspectives for the molecular epidemiology of human African trypanosomosis. *Infection, Genetics and Evolution* 7, 675–684. doi:10.1016/j.meegid.2007.07.001.

Kolev, N. G., Franklin, J. B., Carmi, S., Shi, H., Michaeli, S., and Tschudi, C. (2010). The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog* 6, e1001090. doi:10.1371/journal.ppat.1001090.

Konwar, K. M., Hanson, N. W., Pagé, A. P., and Hallam, S. J. (2013). MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* 14, 202. doi:10.1186/1471-2105-14-202.

Kotsikorou, E., Song, Y. C., Chan, J., Faelens, S., Tovian, Z., Broderick, E., Bakalara, N., Docampo, R., and Oldfield, E. (2005). Bisphosphonate inhibition of the exopolyphosphatase activity of the *Trypanosoma brucei* soluble vacuolar pyrophosphatase. *J. Med. Chem.* 48, 6128–6139. doi:10.1021/jm058220g.

Koumandou, V. L., Natesan, S., Sergeenko, T., and Field, M. C. (2008). The trypanosome transcriptome is remodelled during differentiation but displays limited responsiveness within life stages. *BMC Genomics* 9, 298. doi:10.1186/1471-2164-9-298.

Kramer, P. A., Ravi, S., Chacko, B., Johnson, M. S., and Darley-Usmar, V. M. (2014). A review of the mitochondrial and glycolytic metabolism in human platelets and leukocytes:, Implications for their use as bioenergetic biomarkers. *Redox Biol* 2, 206–210. doi:10.1016/j.redox.2013.12.026.

Kraus, R. H. S., van Hooft, P., Waldenström, J., Latorre-Margalef, N., Ydenberg, R. C., and Prins, H. H. T. (2011). Avian influenza surveillance with FTA cards: field methods, biosafety, and transportation issues solved. *J Vis Exp*. doi:10.3791/2832.

Kristjanson, P. M., Swallow, B. M., Rowlands, G. J., Kruska, R. L., and de Leeuw, P. N. (1999). Measuring the costs of African animal trypanosomosis, the potential benefits of control and returns to research. *Agricultural Systems* 59, 79–98.

Lander, E. S., Consortium, I. H. G. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Langlois, M. R., and Delanghe, J. R. (1996). Biological and clinical significance of haptoglobin polymorphism in humans.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10. doi:10.1186/gb-2009-10-3-r25.

Lee, Y.-S., Tsai, C.-N., Tsai, C.-L., Chang, S.-D., Hsueh, D.-W., Liu, C.-T., Ma, C.-C., Lin, S.-H., Wang, T.-H., and Wang, H.-S. (2008). Comparison of whole genome amplification methods for further quantitative analysis with microarray-based comparative genomic hybridization. *Taiwan J Obstet Gynecol* 47, 32–41. doi:10.1016/S1028-4559(08)60052-2.

Lemercier, G., Espiau, B., Ruiz, F. A., Vieira, M., Luo, S. H., Baltz, T., Docampo, R., and Bakalara, N. (2004). A pyrophosphatase regulating polyphosphate metabolism in acidocalcisomes is essential for Trypanosoma brucei virulence in mice. *J. Biol. Chem.* 279, 3420–3425. doi:10.1074/jbc.M309974200.

Leonard, G., Soanes, D. M., and Stevens, J. R. (2011). Resolving the question of trypanosome monophyly: A comparative genomics approach using whole genome data sets with low taxon sampling. *Infect. Genet. Evol.* 11, 955–959. doi:10.1016/j.meegid.2011.03.005.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Liang, X.-H., Haritan, A., Uliel, S., and Michaeli, S. (2003). trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryotic Cell* 2, 830–840.

Linder, P., and Jankowsky, E. (2011). From unwinding to clamping - the DEAD box RNA helicase family. *Nat Rev Mol Cell Biol* 12, 505–516. doi:10.1038/nrm3154.

Linstead, D. J., Klein, R. A., and CROSS, G. (1977). Threonine Catabolism in *Trypanosoma-Brucei. J. Gen. Microbiol.* 101, 243–251.

Liu, X., Han, S., Wang, Z., Gelernter, J., and Yang, B.-Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* 8. doi:10.1371/journal.pone.0075619.

Lonsdale-Eccles, J. D., and Grab, D. J. (1987). Purification of African trypanosomes can cause biochemical changes in the parasites. *J. Protozool.* 34, 405–408.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550. doi:10.1186/s13059-014-0550-8.

Lugli, E. B., Pouliot, M., Portela, M., Loomis, M. R., and Raper, J. (2004). Characterization of primate trypanosome lytic factors. *Mol. Biochem. Parasitol.* 138, 9–20. doi:10.1016/j.molbiopara.2004.07.004.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nat. Rev. Genet.* 4, 981–994. doi:10.1038/nrg1226.

Lynn, K.-S., Cheng, M.-L., Chen, Y.-R., Hsu, C., Chen, A., Lih, T. M., Chang, H.-Y., Huang, C.-J., Shiao, M.-S., Pan, W.-H., et al. (2015). Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information. *Anal. Chem.* 87, 2143–2151. doi:10.1021/ac503325c.

MacGregor, P., and Matthews, K. R. (2012). Identification of the regulatory elements controlling the transmission stage-specific gene expression of PAD1 in *Trypanosoma brucei*. *Nucleic Acids Research*

40, 7705–7717. doi:10.1093/nar/gks533.

MacGregor, P., Savill, N. J., Hall, D., and Matthews, K. R. (2011). Transmission Stages Dominate Trypanosome Within-Host Dynamics during Chronic Infections. *Cell Host and Microbe* 9, 310–318. doi:10.1016/j.chom.2011.03.013.

Mackey, Z. B., O'Brien, T. C., Greenbaum, D. C., Blank, R. B., and McKerrow, J. H. (2004). A cathepsin B-like protease is required for host protein degradation in *Trypanosoma brucei*. *J. Biol. Chem.* 279, 48426–48433. doi:10.1074/jbc.M402470200.

MacLean, L. M., Odiit, M., Chisi, J. E., and Kennedy, P. G. (2010). Focus-specific clinical profiles in human African Trypanosomiasis caused by Trypanosoma brucei rhodesiense. *PLoS Negl Trop ….*

Maclean, L., Reiber, H., Kennedy, P. G. E., and Sternberg, J. M. (2012). Stage progression and neurological symptoms in *Trypanosoma brucei rhodesiense* sleeping sickness: role of the CNS inflammatory response. *PLoS Negl Trop Dis* 6, e1857. doi:10.1371/journal.pntd.0001857.

Magnus, E., Vervoort, T., and Van Meirvenne, N. (1978). A card-agglutination test with stained trypanosomes (C.A.T.T.) for the serological diagnosis of *T. B. gambiense* trypanosomiasis. *Ann Soc Belg Med Trop* 58, 169–176.

Makowski, G. S., Nadeau, F. L., and Hopfer, S. M. (2003). Single tube multiplex PCR detection of 27 cystic fibrosis mutations and 4 polymorphisms using neonatal blood samples collected on Guthrie cards. *Ann. Clin. Lab. Sci.* 33, 243–250.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118. doi:10.1038/nmeth.1419.

Marcello, L., and Barry, J. D. (2007). Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.* 17, 1344–1352. doi:10.1101/gr.6421207.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9, 387–402. doi:10.1146/annurev.genom.9.081307.164359.

Margulies, E. H., NISC Comparative Sequencing Program, Maduro, V. V. B., Thomas, P. J., Tomkins, J. P., Amemiya, C. T., Luo, M., and Green, E. D. (2005). Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 3354–3359. doi:10.1073/pnas.0408539102.

Marine, R., Polson, S. W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M., and Wommack, K. E. (2011). Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl. Environ. Microbiol.* 77, 8071–8079. doi:10.1128/AEM.05610-11.

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

Masocha, W., and Kristensson, K. (2012). Passage of parasites across the blood-brain barrier. *Virulence* 3, 202–212. doi:10.4161/viru.19178.

Matthews, K. R. (1999). Developments in the differentiation of *Trypanosoma brucei*. *Parasitol. Today (Regul. Ed.)* 15, 76–80.

Matthews, K. R., Ellis, J. R., and Paterou, A. (2004). Molecular regulation of the life cycle of African trypanosomes. *Trends in Parasitology* 20, 40–47. doi:10.1016/j.pt.2003.10.016.

Maxam, A. M., and Gilbert, W. (1977). New Method for Sequencing Dna. *Proc. Natl. Acad. Sci. U.S.A.* 74, 560–564.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* 40, 4288–4297. doi:10.1093/nar/gks042.

McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J.-B., and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Med* 6, 26. doi:10.1186/gm543.

McCourt, C. M., McArt, D. G., Mills, K., Catherwood, M. A., Maxwell, P., Waugh, D. J., Hamilton, P., O'Sullivan, J. M., and Salto-Tellez, M. (2013). Validation of next generation sequencing technologies in comparison to current diagnostic gold standards for BRAF, EGFR and KRAS mutational analysis. *PLoS ONE* 8, e69604. doi:10.1371/journal.pone.0069604.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19, 1527–1541. doi:10.1101/gr.091868.109.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069–2070. doi:10.1093/bioinformatics/btq330.

Mebrahtu, Y. B., Lawyer, P. G., Pamba, H., Koech, D., Perkins, P. V., Roberts, C. R., Were, J. B., and Hendricks, L. D. (1992). Biochemical characterization and zymodeme classification of *Leishmania* isolates from patients, vectors, and reservoir hosts in Kenya. *Am. J. Trop. Med. Hyg.* 47, 852–892.

Mehlert, A., Wormald, M. R., and Ferguson, M. A. J. (2012). Modeling of the N-glycosylated transferrin receptor suggests how transferrin binding can occur within the surface coat of *Trypanosoma brucei*. *PLoS Pathog* 8, e1002618. doi:10.1371/journal.ppat.1002618.

Mehlitz, D., Zillmann, U., Scott, C. M., and Godfrey, D. G. (1982). Epidemiological studies on the animal reservoir of Gambiense sleeping sickness. Part III. Characterization of trypanozoon stocks by isoenzymes and sensitivity to human serum. *Tropenmed Parasitol* 33, 113–118.

Mehta, J., and Tuteja, R. (2011). Inhibition of unwinding and ATPase activities of *Plasmodium falciparum* Dbp5/DDX19 homolog. *Commun Integr Biol* 4, 299–303. doi:10.4161/cib.4.3.14778.

Melnikov, A., Galinsky, K., Rogov, P., Fennell, T., Van Tyne, D., Russ, C., Daniels, R., Barnes, K. G., Bochicchio, J., Ndiaye, D., et al. (2011). Hybrid selection for sequencing pathogen genomes from clinical samples. *Genome Biol* 12, R73. doi:10.1186/gb-2011-12-8-r73.

Mendonça, M. B. A., Nehme, N. S., Santos, S. S., Cupolillo, E., Vargas, N., Junqueira, A., Naiff, R. D., Barrett, T. V., Coura, J. R., Zingales, B., et al. (2002). Two main clusters within *Trypanosoma cruzi* zymodeme 3 are defined by distinct regions of the ribosomal RNA cistron. *Parasitology* 124, 177–184.

Mercaldi, G. F., Pereira, H. M., Cordeiro, A. T., Michels, P. A. M., and Thiemann, O. H. (2012). Structural role of the active-site metal in the conformation of *Trypanosoma brucei* phosphoglycerate mutase. *FEBS Journal* 279, 2012–2021. doi:10.1111/j.1742-4658.2012.08586.x.

Mirchamsy, H., Nazari, F., Stellman, C., and Esterabady, H. (1968). The use of dried whole blood absorbed on filter-paper for the evaluation of diphtheria and tetanus antitoxins in mass surveys. *Bulletin of the World Health Organization* 38, 665–671.

Montilla, M. M., Guhl, F., Jaramillo, C., Nicholls, S., Barnabe, C., Bosseno, M. F., and Breniere, S. F. (2002). Isoenzyme clustering of Trypanosomatidae Colombian populations. *Am. J. Trop. Med. Hyg.* 66, 394–400.

Moreno, S., and Docampo, R. (2003). Calcium regulation in protozoan parasites. *Curr. Opin. Microbiol.* 6, 359–364. doi:10.1016/S1369-5274(03)00091-2.

Morgan, G. W., Hall, B. S., Denny, P. W., Field, M. C., and Carrington, M. (2002). The endocytic apparatus of the kinetoplastida. Part II: machinery and components of the system. *Trends in Parasitology* 18, 540–546.

Morgulis, A., Gertz, E. M., Schäffer, A. A., and Agarwala, R. (2006). Windowmasker: window-based masker for sequenced genomes. *Bioinformatics* 22, 134–141. doi:10.1093/bioinformatics/bti774.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotech.* 45, 81–94. doi:10.2144/000112900.

Morrison, L. J. (2011). Parasite-driven pathogenesis in *Trypanosoma brucei* infections. *Parasite Immunology* 33, 448–455. doi:10.1111/j.1365-3024.2011.01286.x.

Morrison, L. J., McCormack, G., Sweeney, L., Likeufack, A. C. L., Truc, P., Turner, C. M., Tait, A., and MacLeod, A. (2007). Use of multiple displacement amplification to increase the detection and genotyping of *trypanosoma* species samples immobilized on FTA filters. *Am. J. Trop. Med. Hyg.* 76, 1132–1137.

Morrison, L. J., McLellan, S., Sweeney, L., Chan, C. N., MacLeod, A., Tait, A., and Turner, C. M. R. (2010). Role for parasite genetic diversity in differential host responses to *Trypanosoma brucei* infection. *Infection and Immunity* 78, 1096–1108. doi:10.1128/IAI.00943-09.

Morty, R. E., Lonsdale-Eccles, J. D., Morehead, J., Caler, E. V., Mentele, R., Auerswald, E. A., Coetzer, T. H., Andrews, N. W., and Burleigh, B. A. (1999). Oligopeptidase B from *Trypanosoma brucei*, a new member of an emerging subgroup of serine oligopeptidases. *J. Biol. Chem.* 274, 26149–26156.

Morty, R. E., Pellé, R., Vadász, I., Uzcanga, G. L., Seeger, W., and Bubis, J. (2005). Oligopeptidase B from *Trypanosoma evansi*. A parasite peptidase that inactivates atrial natriuretic factor in the bloodstream of infected hosts. *J. Biol. Chem.* 280, 10925–10937. doi:10.1074/jbc.M410066200.

Moscoso, H., Thayer, S. G., Hofacre, C. L., and Kleven, S. H. (2004). Inactivation, storage, and PCR detection of Mycoplasma on FTA filter paper. *Avian Dis.* 48, 841–850.

Murray, M., Morrison, W. I., and Whitelaw, D. D. (1982). Host susceptibility to African trypanosomiasis: trypanotolerance. *Adv. Parasitol.*

Mussmann, R., Engstler, M., Gerrits, H., Kieft, R., Toaldo, C. B., Onderwater, J., Koerten, H., van Luenen, H. G. A. M., and Borst, P. (2004). Factors affecting the level and localization of the transferrin receptor in Trypanosoma brucei. *J. Biol. Chem.* 279, 40690–40698. doi:10.1074/jbc.M404697200.

Muthukrishnan, M., Singanallur, N. B., Ralla, K., and Villuppanoor, S. A. (2008). Evaluation of FTA cards as a laboratory and field sampling device for the detection of foot-and-mouth disease virus and serotyping by RT-PCR and real-time RT-PCR. *Journal of Virological Methods* 151, 311–316. doi:10.1016/j.jviromet.2008.05.020.

Naessens, J. (2006). Bovine trypanotolerance: A natural ability to prevent severe anaemia and haemophagocytic syndrome? *Int. J. Parasitol.* 36, 521–528. doi:10.1016/j.ijpara.2006.02.012.

Ng'ayo, M. O., Njiru, Z. K., Kenya, E. U., Muluvi, G. M., Osir, E. O., and Masiga, D. K. (2005). Detection of trypanosomes in small ruminants and pigs in western Kenya: important reservoirs in the epidemiology of sleeping sickness? *Kinetoplastid Biol Dis* 4, 5. doi:10.1186/1475-9292-4-5.

Nikolskaia, O. V., de A Lima, A. P. C., Kim, Y. V., Lonsdale-Eccles, J. D., Fukuma, T., Scharfstein, J., and Grab, D. J. (2006). Blood-brain barrier traversal by African trypanosomes requires calcium signaling induced by parasite cysteine protease. *J. Clin. Invest.* 116, 2739–2747. doi:10.1172/JCI27798.

Nilsson, D., Gunasekera, K., Mani, J., Osteras, M., Farinelli, L., Baerlocher, L., Roditi, I., and Ochsenreiter, T. (2010). Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of *Trypanosoma brucei*. *PLoS Pathog* 6, e1001037. doi:10.1371/journal.ppat.1001037.s021.

Njiru, Z. K., Ndung'u, K., Matete, G., Ndungu, J. M., and Gibson, W. C. (2004). Detection of *Trypanosoma brucei rhodesiense* in animals from sleeping sickness foci in East Africa using the serum resistance associated (SRA) gene. *Acta Tropica* 90, 249–254. doi:10.1016/j.actatropica.2004.01.001.

Noireau, F., Paindavoine, P., Lemesre, J. L., Toudic, A., Pays, E., Gouteux, J. P., Steinert, M., and Frezil, J. L. (1989). The epidemiological importance of the animal reservoir of *Trypanosoma brucei gambiense* in the Congo. 2. Characterization of the Trypanosoma brucei complex. *Trop. Med. Parasitol.* 40, 9–11.

Noyes, H. A., Agaba, M., Anderson, S., Archibald, A. L., Brass, A., Gibson, J., Hall, L., Hulme, H., Oh, S. J., and Kemp, S. (2010). Genotype and expression analysis of two inbred mouse strains and two derived congenic strains suggest that most gene expression is trans regulated and sensitive to genetic background. *BMC Genomics* 11, 361. doi:10.1186/1471-2164-11-361.

O'Gorman, G. M., Park, S., Hill, E. W., and Meade, K. G. (2009). Transcriptional profiling of cattle infected with *Trypanosoma congolense* highlights gene expression signatures underlying trypanotolerance and trypanosusceptibility. *BMC* Genomics. 10, 207. doi:10.1186/1471-2164-10-207.

Onyango, R. J., Van Hoeve, K., and De Raadt, P. (1966). The epidemiology of *Trypanosoma rhodesiense* sleeping sickness in Alego location, Central Nyanza, Kenya. I. Evidence that cattle may act as reservoir hosts of trypanosomes infective to man. *Trans. R. Soc. Trop. Med. Hyg.* 60, 175–182.

Orenge, C. O., Munga, L., Kimwele, C. N., Kemp, S., Korol, A., Gibson, J. P., Hanotte, O., and Soller, M. (2012). Trypanotolerance in N'Dama x Boran crosses under natural trypanosome challenge: effect of test-year environment, gender, and breed composition. *BMC Genet.* 13, 87. doi:10.1186/1471-2156-13-87.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature* 461, 814–818. doi:10.1038/nature08390.

Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X., Richter, D. J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., et al. (2004). Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Research* 32, e71. doi:10.1093/nar/gnh069.

Paindavoine, P., Zampetti-Bosseler, F., Coquelet, H., Pays, E., and Steinert, M. (1989). Different allele frequencies in *Trypanosoma brucei brucei* and *Trypanosoma brucei gambiense* populations. *Mol. Biochem. Parasitol.* 32, 61–71.

Pan, Z., Gu, H., Talaty, N., Chen, H., Shanaiah, N., Hainline, B. E., Cooks, R. G., and Raftery, D. (2007). Principal component analysis of urine metabolites detected by NMR and DESI-MS in patients with inborn errors of metabolism. *Anal Bioanal Chem* 387, 539–549. doi:10.1007/s00216-006-0546-7.

Parsons, M., NELSON, R. G., Watkins, K. P., and Agabian, N. (1984). Trypanosome Messenger-Rnas Share a Common 5' Spliced Leader Sequence. *Cell* 38, 309–316.

Parsons, M., Worthey, E. A., Ward, P. N., and Mottram, J. C. (2005). Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. *BMC Genomics* 6. doi:10.1186/1471-2164-6-127.

Paterou, A., Walrad, P., Craddy, P., Fenn, K., and Matthews, K. (2006). Identification and stage-specific association with the translational apparatus of TbZFP3, a CCCH protein that promotes trypanosome life-cycle development. *J. Biol. Chem.* 281, 39002–39013. doi:10.1074/jbc.M604280200.

Pays, E., and Vanhollebeke, B. (2008). Mutual self-defence: the trypanolytic factor story. *Microbes Infect.* 10, 985–989. doi:10.1016/j.micinf.2008.07.020.

Pays, E., Lips, S., Nolan, D., Vanhamme, L., and Pérez-Morga, D. (2001). The VSG expression sites of Trypanosoma brucei: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol. Biochem. Parasitol.* 114, 1–16.

Pays, E., Vanhamme, L., and Perez-Morga, D. (2004). Antigenic variation in *Trypanosoma brucei*: facts, challenges and mysteries. *Curr. Opin. Microbiol.* 7, 369–374. doi:10.1016/j.mib.2004.05.001.

Pays, E., Vanhollebeke, B., Vanhamme, L., Paturiaux-Hanocq, F., Nolan, D. P., and Perez-Morga, D. (2006). The trypanolytic factor of human serum. *Nature Reviews Microbiology* 4, 477–486. doi:10.1038/nrmicro1428.

Peacock, C. S., Seeger, K., Harris, D., Murphy, L., Ruiz, J. C., Quail, M. A., 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. doi:10.1038/ng2053.

Peacock, L., Ferris, V., Bailey, M., and Gibson, W. (2008). Fly transmission and mating of *Trypanosoma brucei brucei* strain 427. *Mol. Biochem. Parasitol.* 160, 100–106.

doi:10.1016/j.molbiopara.2008.04.009.

Pettitt, J., Mueller, B., Stansfield, I., and Connolly, B. (2008). Spliced leader trans-splicing in the nematode *Trichinella spiralis* uses highly polymorphic, noncanonical spliced leaders. *RNA* 14, 760–770. doi:10.1261/rna.948008.

Pépin, J., Milord, F., and Khonde, A. (1994). Gambiense trypanosomiasis: frequency of, and risk factors for, failure of melarsoprol therapy. *Trans R Soc Trop Med Hyg*. 88(4):447-52

Pérez-Moreno, G., Sealey-Cardona, M., Rodrigues-Poveda, C., Gelb, M. H., Ruiz-Pérez, L. M., Castillo-Acosta, V., Urbina, J. A., and González-Pacanowska, D. (2012). Endogenous sterol biosynthesis is important for mitochondrial function and cell morphology in procyclic forms of *Trypanosoma brucei*. *Int. J. Parasitol.* 42, 975–989. doi:10.1016/j.ijpara.2012.07.012.

Picozzi, K., Carrington, M., and Welburn, S. C. (2008). A multiplex PCR that discriminates between *Trypanosoma brucei brucei* and zoonotic *T. b. rhodesiense*. *Exp. Parasitol.* 118, 41–46. doi:10.1016/j.exppara.2007.05.014.

Picozzi, K., Fevre, E. M., Odiit, M., Carrington, M., Eisler, M. C., Maudlin, I., and Welburn, S. C. (2005). Sleeping sickness in Uganda: a thin line between two fatal diseases. *BMJ* 331, 1238–1241. doi:10.1136/bmj.331.7527.1238.

Picozzi, K., Tilley, A., Fèvre, E. M., Coleman, P. G., Magona, J. W., Odiit, M., Eisler, M. C., and Welburn, S. C. (2002). The diagnosis of trypanosome infections: applications of novel technology for reducing disease risk. *African Journal of Biotechnology* 1, 39–45. doi:10.4314/ajb.v1i2.14813.

Portman, N., and Gull, K. (2010). The paraflagellar rod of kinetoplastid parasites: From structure to components and function. *Int. J. Parasitol.* 40, 135–148. doi:10.1016/j.ijpara.2009.10.005.

Priotto, G., Fogg, C., Balasegaram, M., Erphas, O., Louga, A., Checchi, F., Ghabri, S., and Piola, P. (2006). Three drug combinations for late-stage *Trypanosoma brucei gambiense* sleeping sickness: A randomized clinical trial in Uganda. *PLoS Clin Trials* 1. doi:10.1371/journal.pctr.0010039.

Priotto, G., Kasparian, S., Mutombo, W., Ngouama, D., Ghorashian, S., Arnold, U., Ghabri, S., Baudin, E., Buard, V., Kazadi-Kyanza, S., et al. (2009). Multicentre clinical trial of nifurtimox-eflornithine combination therapy for second-stage sleeping sickness. *Trop. Med. Int. Health* 14, 43–43.

Priotto, G., Kasparian, S., Ngouama, D., Ghorashian, S., Arnold, U., Ghabri, S., and Karunakara, U. (2007). Nifurtimox-Eflornithine combination therapy for second-stage *Trypanosoma brucei gambiense* sleeping sickness: A randomized clinical trial in Congo. *Clin. Infect. Dis.* 45, 1435–1442. doi:10.1086/522982.

Radwanska, M., Chamekh, M., Vanhamme, L., Claes, F., Magez, S., Magnus, E., de Baetselier, P., Büscher, P., and Pays, E. (2002). The serum resistance-associated gene as a diagnostic tool for the detection of *Trypanosoma brucei rhodesiense*. *Am. J. Trop. Med. Hyg.* 67, 684–690.

Ramsey, J. M., Peterson, A. T., Carmona-Castro, O., Moo-Llanes, D. A., Nakazawa, Y., Butrick, M., Tun-Ku, E., la Cruz-Félix, K. de, and Ibarra-Cerdeña, C. N. (2015). Atlas of Mexican *Triatominae* (Reduviidae: Hemiptera) and vector transmission of Chagas disease. *Mem. Inst. Oswaldo Cruz* 110, 339–352. doi:10.1590/0074-02760140404.

Raper, J., Fung, R., Ghiso, J., Nussenzweig, V., and Tomlinson, S. (1999). Characterization of a novel trypanosome lytic factor from human serum. *Infection and Immunity* 67, 1910–1916.

Ratan, A., Miller, W., Guillory, J., Stinson, J., and Seshagiri, S. (2013). Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS ONE*.

Reuner, B., Vassella, E., Yutzy, B., and Boshart, M. (1997). Cell density triggers slender to stumpy differentiation of *Trypanosoma brucei* bloodstream forms in culture. *Mol. Biochem. Parasitol.* 90, 269–280.

Richmond, G. S., Gibellini, F., Young, S. A., Major, L., Denton, H., Lilley, A., and Smith, T. K. (2010). Lipidomic analysis of bloodstream and procyclic form *Trypanosoma brucei*. *Parasitology* 137, 1357–1392. doi:10.1017/S0031182010000715.

Rico, E., Rojas, F., Mony, B. M., Szoor, B., MacGregor, P., and Matthews, K. R. (2013). Bloodstream form pre-adaptation to the tsetse fly in *Trypanosoma brucei*. *Front Cell Infect Microbiol* 3, 78. doi:10.3389/fcimb.2013.00078.

Rigaud, T., Perrot-Minnot, M.-J., and Brown, M. J. F. (2010). Parasite and host assemblages: embracing the reality will improve our knowledge of parasite transmission and virulence. *Proceedings of the Royal Society B: Biological Sciences* 277, 3693–3702. doi:10.1098/rspb.2010.1163.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., WGS500 Consortium, Wilkie, A. O. M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* 46, 912–918. doi:10.1038/ng.3036.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Rotureau, B., Subota, I., and Bastin, P. (2011). Molecular bases of cytoskeleton plasticity during the *Trypanosoma brucei* parasite cycle. *Cellular Microbiology* 13, 705–716. doi:10.1111/j.1462-5822.2010.01566.x.

Ruben, L., and Patton, C. L. (1985). Antibodies to calmodulin during experimental *Trypanosoma brucei rhodesiense* infections in rabbits. *Immunology* 56, 227–233.

Rusk, N. (2011). Torrents of sequence. *Nat. Methods* 8, 44–44. doi:10.1038/NMETH.F.330.

Saeij, J. P. J., Boyle, J. P., Coller, S., Taylor, S., Sibley, L. D., Brooke-Powell, E. T., Ajioka, J. W., and Boothroyd, J. C. (2006). Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science* 314, 1780–1783. doi:10.1126/science.1133690.

Safar, Al, H. S., Abidi, F. H., Khazanehdari, K. A., Dadour, I. R., and Tay, G. K. (2010). Evaluation of different sources of DNA for use in genome wide studies and forensic application. *Appl Microbiol Biotechnol* 89, 807–815. doi:10.1007/s00253-010-2926-3.

Sakharkar, K. R., Dhar, P. K., and Chow, V. (2004). Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int. J. Syst. Evol. Microbiol.* 54, 1937–1941. doi:10.1099/ijs.0.63090-0.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467.

Sargeaunt, P. G., and Williams, J. E. (1979). Electrophoretic isoenzyme patterns of the pathogenic and non-pathogenic intestinal amoebae of man. *Trans. R. Soc. Trop. Med. Hyg.* 73, 225–227.

Sawyer, W. H., and Puckridge, J. (1973). The dissociation of proteins by chaotropic salts. *J. Biol. Chem.* 248, 8429–8433.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biol* 12, 125. doi:10.1186/gb-2011-12-8-125.

Scheltema, R. A., Jankevics, A., Jansen, R. C., Swertz, M. A., and Breitling, R. (2011). PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal. Chem.* 83, 2786–2793. doi:10.1021/ac2000994.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 7, 302. doi:10.1186/1471-2105-7-302.

Schmidt, R. S., and Bütikofer, P. (2014). Autophagy in *Trypanosoma brucei:* Amino Acid Requirement and Regulation during Different Growth Phases. *PLoS ONE* 9, e93875. doi:10.1371/journal.pone.0093875.s003.

Scory, S., Caffrey, C. R., Stierhof, Y. D., Ruppel, A., and Steverding, D. (1999). *Trypanosoma brucei*: killing of bloodstream forms in vitro and in vivo by the cysteine proteinase inhibitor Z-phe-ala-CHN2. *Exp. Parasitol.* 91, 327–333. doi:10.1006/expr.1998.4381.

Seed, J. R., and Wenck, M. A. (2003). Role of the long slender to short stumpy transition in the life cycle of the african trypanosomes. *Kinetoplastid Biol Dis* 2, 3. doi:10.1186/1475-9292-2-3.

Seed, J. R., Sechelski, J. B., and Loomis, M. R. (2007). A survey for a trypanocidal factor in primate sera. *J. Protozool.* 37, 393–400. doi:10.1111/j.1550-7408.1990.tb01163.x.

Seth-Smith, H. M. B., Harris, S. R., Scott, P., Parmar, S., Marsh, P., Unemo, M., Clarke, I. N., Parkhill, J., and Thomson, N. R. (2013). Generating whole bacterial genome sequences of low-abundance species from complex samples with IMS-MDA. *Nat Protoc* 8, 2404–2412. doi:10.1038/nprot.2013.147.

Shiflett, A. M., Faulkner, S. D., Cotlin, L. F., Widener, J., Stephens, N., and Hajduk, S. L. (2007). African trypanosomes: intracellular trafficking of host defense molecules. *J. Eukaryot. Microbiol.* 54, 18–21. doi:10.1111/j.1550-7408.2006.00228.x.

Shimamura, M., Hager, K. M., and Hajduk, S. L. (2001). The lysosomal targeting and intracellular metabolism of trypanosome lytic factor by *Trypanosoma brucei brucei*. *Mol. Biochem. Parasitol.* 115, 227–237.

Siegel, T. N., Gunasekera, K., Cross, G. A. M., and Ochsenreiter, T. (2011). Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing. *Trends in Parasitology* 27, 434–441. doi:10.1016/j.pt.2011.05.006.

Simarro, P. (2011). African trypanosomiasis: current burden of disease and geographical distribution. *Trop. Med. Int. Health* 16, 22–22.

Simarro, P. P., Cecchi, G., Paone, M., Franco, J. R., Diarra, A., Ruiz, J. A., Fevre, E. M., Courtin, F., Mattioli, R. C., and Jannin, J. G. (2010). The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. *Int J Health Geogr* 9, 57. doi:10.1186/1476-072X-9-57.

Simarro, P. P., Jannin, J., and Cattand, P. (2008). Eliminating Human African Trypanosomiasis: Where Do We Stand and What Comes Next? *PLOS Med* 5, e55. doi:10.1371/journal.pmed.0050055.

Simo, G., Njiokou, F., Tume, C., Lueong, S., de Meeûs, T., Cuny, G., and Asonganyi, T. (2010). Population genetic structure of Central African *Trypanosoma brucei gambiense* isolates using microsatellite DNA markers. *Infect. Genet. Evol.* 10, 68–76. doi:10.1016/j.meegid.2009.09.019.

Simo, G., Njitchouang, G. R., Njiokou, F., Cuny, G., and Asonganyi, T. (2011). *Trypanosoma brucei* s.l.: Microsatellite markers revealed high level of multiple genotypes in the mid-guts of wild tsetse flies of the Fontem sleeping sickness focus of Cameroon. *Exp. Parasitol.* 128, 272–278. doi:10.1016/j.exppara.2011.02.023.

Sloof, P., Bos, J. L., Konings, A. F., Menke, H. H., Borst, P., Gutteridge, W. E., and Leon, W. (1983). Characterization of satellite DNA in *Trypanosoma brucei* and *Trypanosoma cruzi*. *Journal of Molecular Biology* 167, 1–21.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 78, 779–787. doi:10.1021/ac051437y.

Smith, D. H., and Bailey, J. W. (2000). Human African trypanosomiasis in south-eastern Uganda: clinical diversity and isoenzyme profiles. *Ann Trop Med Parasitol* 91, 851–856.

Smith, L. M., and Burgoyne, L. A. (2004). Collecting, archiving and processing DNA from wildlife samples using FTA databasing paper. *BMC Ecol.* 4, 4. doi:10.1186/1472-6785-4-4.

Smith, T. K., and Buetikofer, P. (2010). Lipid metabolism in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 172, 66–79. doi:10.1016/j.molbiopara.2010.04.001.

Soccol, V. T., Barnabé, C., Castro, E., and Luz, E. (2002). *Trypanosoma cruzi*: isoenzyme analysis suggests the presence of an active Chagas sylvatic cycle of recent origin in Paraná State, Brazil. *Experimental ….*

Soliman, K., El-Ansary, A., and Mohamed, A. M. (2001). Effect of carnosine administration on metabolic parameters in bilharzia-infected hamsters. *Comp. Biochem. Physiol. B, Biochem. Mol. Biol.* 129, 157–164.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14, 91. doi:10.1186/1471-2105-14-91.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. (2006). Whole-genome multiple displacement amplification from single cells. *Nat Protoc* 1, 1965–1970. doi:10.1038/nprot.2006.326.

Stangegaard, M., Borsting, C., Ferrero-Miliani, L., Frank-Hansen, R., Poulsen, L., Hansen, A. J., and Morling, N. (2013). Evaluation of Four Automated Protocols for Extraction of DNA from FTA Cards. *J Lab Autom* 18, 404–410. doi:10.1177/2211068213484472.

Sternberg, J. M., and Maclean, L. (2010). A spectrum of disease in human African trypanosomiasis: the host and parasite genetics of virulence. *Parasitology* 137, 2007–2015. doi:10.1017/S0031182010000946.

Stevens, J. R., and Gibson, W. (1999a). The Molecular Evolution of Trypanosomes. *Parasitology Today* 15, 432–437. doi:10.1016/S0169-4758(99)01532-X.

Stevens, J. R., and Tibayrence, M. (1995). Detection of Linkage Disequilibrium in Trypanosoma-Brucei Isolated From Tsetse-Flies and Characterized by Rapd Analysis and Isoenzymes. *Parasitology* 110, 181–186.

Stevens, J. R., Noyes, H. A., Dover, G. A., and Gibson, W. C. (1999). The ancient and divergent origins of the human pathogenic trypanosomes, *Trypanosoma brucei* and *T. cruzi*. *Parasitology* 118 (Pt 1), 107–116.

Stevens, J., and Gibson, W. (1999b). The Evolution of Salivarian Trypanosomes. *Mem. Inst. Oswaldo Cruz* 94, 225–228. doi:10.1590/S0074-02761999000200019.

Steverding, D. (2010). The development of drugs for treatment of sleeping sickness: a historical review. *Parasites & Vectors* 3, 15. doi:10.1186/1756-3305-3-15.

Steverding, D. (2000). The transferrin receptor of *Trypanosoma brucei*. *Parasitol. Int.* 48, 191–198.

Steverding, D., Stierhof, Y. D., Chaudhri, M., Ligtenberg, M., Schell, D., Beck-Sickinger, A. G., and Overath, P. (1994). ESAG 6 and 7 products of *Trypanosoma brucei* form a transferrin binding protein complex. *Eur. J. Cell Biol.* 64, 78–87.

Stibbs, H. H. (1984). Effects of African trypanosomiasis on brain levels of dopamine, serotonin, 5-hydroxyindoleacetic acid, and homovanillic acid in the rabbit. *J. Neurochem.* 43, 1253–1256.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6, e21800. doi:10.1371/journal.pone.0021800.

Sykes, S. E., and Hajduk, S. L. (2013). Dual Functions of alpha-Ketoglutarate Dehydrogenase E2 in the Krebs Cycle and Mitochondrial DNA Inheritance in *Trypanosoma brucei*. *Eukaryotic Cell* 12, 78–90. doi:10.1128/EC.00269-12.

Symula, R. E., Beadell, J. S., Sistrom, M., Agbebakun, K., Balmer, O., Gibson, W., Aksoy, S., and Caccone, A. (2012). *Trypanosoma brucei gambiense* group 1 is distinguished by a unique amino acid substitution in the HpHb receptor implicated in human serum resistance. *PLoS Negl Trop Dis* 6, e1728. doi:10.1371/journal.pntd.0001728.

Tachibana, S.-I., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., Arisue, N., Palacpac, N. M. Q., Honma, H., Yagi, M., et al. (2012). *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet* 44, 1051–. doi:10.1038/ng.2375.

Tait, A., Masiga, D., Ouma, J., MacLeod, A., Sasse, J., Melville, S., Lindegard, G., McIntosh, A., and Turner, M. (2002). Genetic analysis of phenotype in *Trypanosoma brucei*: a classical approach to potentially complex traits. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 357, 89–99. doi:10.1098/rstb.2001.1050.

Tarailo-Graovac, M., and Chen, N. (2009). Using Repeatmasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, Unit 4.10. doi:10.1002/0471250953.bi0410s25.

Tariq, M. A., Kim, H. J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from

minute amount of total RNA. *Nucleic Acids Research* 39, e120–e120. doi:10.1093/nar/gkr547.

Taylor, R. E. (2008). Tedanolide and the evolution of polyketide inhibitors of eukaryotic protein synthesis. *Nat Prod Rep* 25, 854–861. doi:10.1039/b805700c.

Teixeira, A. R. L., and Soulsby, E. J. L. (1987). *The stercorarian trypanosomes.* CRC Press, Inc.

Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjöld, M., Ponder, B. A., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 13, 718–725.

Tibayrenc, M. (1998). Genetic epidemiology of parasitic protozoa and other infectious agents: the need for an integrated approach. *Int J Parasitol*. 28, 85-104.

Tibayrenc, M. (2011). *Genetics and Evolution of Infectious Diseases*. Elsevier.

Tilley, A., Welburn, S. C., Fèvre, E. M., Feil, E. J., and Hide, G. (2003). *Trypanosoma brucei*: trypanosome strain typing using PCR analysis of mobile genetic elements (MGE-PCR). *Exp. Parasitol.* 104, 26–32.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi:10.1093/bioinformatics/btp120.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7, 562–578. doi:10.1038/nprot.2012.016.

Treff, N. R., Su, J., Tao, X., Northrop, L. E., and Scott, R. T. (2011). Single-cell whole-genome amplification technique impacts the accuracy of SNP microarray-based genotyping and copy number analyses. *Mol. Hum. Reprod.* 17, 335–343. doi:10.1093/molehr/gaq103.

Troeberg, L., Morty, R. E., Pike, R. N., Lonsdale-Eccles, J. D., Palmer, J. T., McKerrow, J. H., and Coetzer, T. H. (1999). Cysteine proteinase inhibitors kill cultured bloodstream forms of *Trypanosoma brucei brucei*. *Exp. Parasitol.* 91, 349–355. doi:10.1006/expr.1998.4386.

Truc, P., Büscher, P., Cuny, G., Gonzatti, M. I., and Jannin, J. (2013). Atypical human infections by animal trypanosomes. *PLoS Negl Trop …*.

Truc, P., Lejon, V., Magnus, E., Jamonneau, V., Nangouma, A., Verloo, D., Penchenier, L., and Büscher, P. (2002). Evaluation of the micro-CATT, CATT/*Trypanosoma brucei gambiense*, and LATEX/*T b gambiense* methods for serodiagnosis and surveillance of human African trypanosomiasis in West and Central Africa. *Bulletin of the World Health Organization* 80, 882–886.

Tschudi, C., Young, A. S., Ruben, L., Patton, C. L., and Richards, F. F. (1985). Calmodulin Genes in Trypanosomes Are Tandemly Repeated and Produce Multiple Messenger-Rnas with a Common 5' Leader Sequence. *Proc. Natl. Acad. Sci. U.S.A.* 82, 3998–4002.

Turner, C. M., and Barry, J. D. (1989). High frequency of antigenic variation in *Trypanosoma brucei rhodesiense* infections. *Parasitology* 99 Pt 1, 67–75.

Tyler, K. M., Matthews, K. R., and Gull, K. (1997). The bloodstream differentiation-division of *Trypanosoma brucei* studied using mitochondrial markers. *Proc. Biol. Sci.* 264, 1481–1490. doi:10.1098/rspb.1997.0205.

Uzcategui, N. L., Szallies, A., Pavlovic-Djuranovic, S., Palmada, M., Figarella, K., Boehmer, C., Lang, F., Beitz, E., and Duszenko, M. (2004). Cloning, Heterologous Expression, and Characterization of Three Aquaglyceroporins from *Trypanosoma brucei*. *Journal of Biological Chemistry* 279, 42669–42676. doi:10.1074/jbc.M404518200.

Uzureau, P., Uzureau, S., Lecordier, L., Fontaine, F., Tebabi, P., Homblé, F., Grélard, A., Zhendre, V., Nolan, D. P., Lins, L., et al. (2013). Mechanism of *Trypanosoma brucei gambiense* resistance to human serum. *Nature* 501, 430–434. doi:10.1038/nature12516.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G.,

McKernan, K., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18, 1051–1063. doi:10.1101/gr.076463.108.

Van Den Abbeele, J., Claes, Y., Van Bockstaele, D., LE RAY, D., and COOSEMANS, M. (1999). *Trypanosoma brucei* spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. *Parasitology* 118, 469–478.

van Grinsven, K. W. A., Van Den Abbeele, J., Van den Bossche, P., van Hellemond, J. J., and Tielens, A. G. M. (2009a). Adaptations in the glucose metabolism of procyclic *Trypanosoma brucei* isolates from tsetse flies and during differentiation of bloodstream forms. *Eukaryotic Cell* 8, 1307–1311. doi:10.1128/EC.00091-09.

van Grinsven, K. W. A., van Hellemond, J. J., and Tielens, A. G. M. (2009b). Acetate:succinate CoA-transferase in the anaerobic mitochondria of *Fasciola hepatica*. *Mol. Biochem. Parasitol.* 164, 74–79. doi:10.1016/j.molbiopara.2008.11.008.

van Hellemond, J. J., and Tielens, A. G. M. (2006). Adaptations in the lipid metabolism of the protozoan parasite *Trypanosoma brucei*. *FEBS Lett.* 580, 5552–5558. doi:10.1016/j.febslet.2006.07.056.

van Luenen, H. G. A. M., Kieft, R., Mussmann, R., Engstler, M., Riet, ter, B., and Borst, P. (2005). Trypanosomes change their transferrin receptor expression to allow effective uptake of host transferrin. *Molecular Microbiology* 58, 151–165. doi:10.1111/j.1365-2958.2005.04831.x.

Vandenberghe, A. E., Meedel, T. H., and Hastings, K. E. (2001). mRNA 5'-leader trans-splicing in the chordates. *Genes & Development* 15, 294–303. doi:10.1101/gad.865401.

Vanhamme, L., Paturiaux-Hanocq, F., Poelvoorde, P., Nolan, D. P., Lins, L., Van Den Abbeele, J., Pays, A., Tebabi, P., Van Xong, H., Jacquet, A., et al. (2003). Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* 422, 83–87. doi:10.1038/nature01461.

Vanhamme, L., Postiaux, S., Poelvoorde, P., and Pays, E. (1999). Differential regulation of ESAG transcripts in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 102, 35–42.

Vanhollebeke, B., and Pays, E. (2006). The function of apolipoproteins L. *Cell. Mol. Life Sci.* 63, 1937–1944. doi:10.1007/s00018-006-6091-x.

Vanhollebeke, B., and Pays, E. (2010). The trypanolytic factor of human serum: many ways to enter the parasite, a single way to kill. *Molecular Microbiology* 76, 806–814. doi:10.1111/j.1365-2958.2010.07156.x.

Vanhollebeke, B., De Muylder, G., Nielsen, M. J., Pays, A., Tebabi, P., Dieu, M., Raes, M., Moestrup, S. K., and Pays, E. (2008). A haptoglobin-hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. *Science* 320, 677–681. doi:10.1126/science.1156296.

Vanhollebeke, B., Nielsen, M. J., Watanabe, Y., Truc, P., Vanhamme, L., Nakajima, K., Moestrup, S. K., and Pays, E. (2007). Distinct roles of haptoglobin-related protein and apolipoprotein L-I in trypanolysis by human serum. *Proc. Natl. Acad. Sci. U.S.A.* 104, 4118–4123. doi:10.1073/pnas.0609902104.

Vassella, E., Oberle, M., Urwyler, S., Renggli, C. K., Studer, E., Hemphill, A., Fragoso, C., Bütikofer, P., Brun, R., and Roditi, I. (2009). Major Surface Glycoproteins of Insect Forms of *Trypanosoma brucei* Are Not Essential for Cyclical Transmission by Tsetse. *PLoS ONE* 4, e4493. doi:10.1371/journal.pone.0004493.

Venkatesan, S., and Ormerod, W. E. (1976). Lipid content of the slender and stumpy forms of *Trypanosoma brucei rhodesiense*: a comparative study. *Comp. Biochem. Physiol., B* 53, 481–487.

Vergnaud, G. (2000). Minisatellites: Mutability and Genome Architecture. *Genome Res.* 10, 899–907. doi:10.1101/gr.10.7.899.

Vertommen, D., Van Roy, J., Szikora, J.-P., Rider, M. H., Michels, P. A. M., and Opperdoes, F. R. (2008). Differential expression of glycosomal and mitochondrial proteins in the two major life-cycle stages of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 158, 189–201. doi:10.1016/j.molbiopara.2007.12.008.

Vincendeau, P., and Bouteille, B. (2006). Immunology and immunopathology of African trypanosomiasis. *An. Acad. Bras. Cienc.* 78, 645–665.

Vincent, I. M., and Barrett, M. P. (2015). Metabolomic-Based Strategies for Anti-Parasite Drug Discovery. *Journal of Biomolecular Screening* 20, 44–55. doi:10.1177/1087057114551519.

Vincent, I. M., Creek, D., Watson, D. G., Kamleh, M. A., Woods, D. J., Wong, P. E., Burchmore, R. J. S., and Barrett, M. P. (2010). A molecular mechanism for eflornithine resistance in African trypanosomes. *PLoS Pathog* 6, e1001204. doi:10.1371/journal.ppat.1001204.

Voet, T., Kumar, P., Van Loo, P., Cooke, S. L., Marshall, J., Lin, M.-L., Esteki, M. Z., Van der Aa, N., Mateiu, L., McBride, D. J., et al. (2013). Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Research* 41, 6119–6138. doi:10.1093/nar/gkt345.

Vuyisich, M., Arefin, A., Davenport, K., Feng, S., Gleasner, C., McMurry, K., Parson-Quintana, B., Price, J., Scholz, M., and Chain, P. (2014). Facile, high quality sequencing of bacterial genomes from small amounts of DNA. *Int J Genomics* 2014, 434575. doi:10.1155/2014/434575.

Walter, R. D., and Albiez, E. J. (1981). Inhibition of Nadp-Linked Malic Enzyme From Onchocerca-Volvulus and Dirofilaria-Immitis by Suramin. *Mol. Biochem. Parasitol.* 4, 53–60.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38, e164. doi:10.1093/nar/gkq603.

Wang, Y., Utzinger, J., Saric, J., Li, J. V., Burckhardt, J., Dirnhofer, S., Nicholson, J. K., Singer, B. H., Brun, R., and Holmes, E. (2008). Global metabolic responses of mice to *Trypanosoma brucei brucei* infection. *Proceedings of the National Academy of Sciences* 105, 6127–6132. doi:10.1073/pnas.0801777105.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi:10.1038/nature01262.

Widener, J., Nielsen, M. J., Shiflett, A., Moestrup, S. K., and Hajduk, S. (2007). Hemoglobin Is a Co-Factor of Human Trypanosome Lytic Factor. *PLoS Pathog* 3, e129. doi:10.1371/journal.ppat.0030129.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243. doi:10.1038/nature07002.

Wilkerson, M. D., Cabanski, C. R., Sun, W., Hoadley, K. A., Walter, V., Mose, L. E., Troester, M. A., Hammerman, P. S., Parker, J. S., Perou, C. M., et al. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Research* 42, e107. doi:10.1093/nar/gku489.

Wootton, J. C., Feng, X. R., Ferdig, M. T., Cooper, R. A., Mu, J. B., Baruch, D. I., Magill, A. J., and Su, X. Z. (2002). Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* 418, 320–323. doi:10.1038/nature00813.

World Health Organization (2014). A global brief on vector-borne diseases.

Xia, J., Sinelnikov, I. V., Han, B., and Wishart, D. S. (2015). MetaboAnalyst 3.0-making metabolomics more meaningful. *Nucleic Acids Research* 43, W251–7. doi:10.1093/nar/gkv380.

Xong, H. V., Vanhamme, L., Chamekh, M., Chimfwembe, C. E., Van Den Abbeele, J., Pays, A., Van Meirvenne, N., Hamers, R., De Baetselier, P., and Pays, E. (1998). A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*. *Cell* 95, 839–846.

Zerbino, D. R. (2010). Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* Chapter 11, Unit 11.5. doi:10.1002/0471250953.bi1105s31.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci. U.S.A.* 89, 5847–5851.

Zhang, Z. Q., and Baltz, T. (1994). Identification of *Trypanosoma evansi, Trypanosoma equiperdum* and

Trypanosoma brucei brucei using repetitive DNA probes. *Vet. Parasitol.* 53, 197–208.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* 9, e78644. doi:10.1371/journal.pone.0078644.

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi:10.1093/bioinformatics/bts606.

Zinoviev, A., and Shapira, M. (2012). Evolutionary conservation and diversification of the translation initiation apparatus in trypanosomatids. *Comp. Funct. Genomics* 2012, 813718. doi:10.1155/2012/813718.

Appendices

Additional appendices are given in the format of excel spreadsheets for the metabolomics data and for additional information for the differential gene expression analysis in the transcriptomic data. A brief summary of each of these appendices is provided beneath.

A11. B17_metabolites_used_to_generate_heatmaps_and_pca.xlsx

This provides the LC-MS intensity data for all of the metabolites that were identified from the B17 infection, with intensity infection per individual given at all three time points. This information was used to generate the heatmap and PCA plots for the B17 infections in Chapter 4.

A12. Z310_metabolites_used_to_generate_heatmaps_and_pca.xlsx

This provides the LC-MS intensity data for all of the metabolites that were identified from the Z310 infection, with intensity infection per individual given at all three time points. This information was used to generate the heatmap and PCA plots for the Z310 infections in Chapter 4.

A13. B17_IDEOM_corrected_to_control.xlsx
This contains the intensity data, and additional information, such as the KEGG pathways associated with these metabolites for the B17 infection. Standard metabolites used for calibration are highlighted in yellow in the metabolite ID column. Averaged intensities across individuals for each metabolite are given in the comparison sheet. High intensities relative to the control are given in red. Intensities with no difference compared to the control are given in blue, and low intensity metabolites yellow.

A14. Z310_IDEOM_corrected_to_control.xlsx

This contains the intensity data, and additional information, such as the KEGG pathways associated with these metabolites for the Z310 infection. Standard metabolites used for calibration are highlighted in yellow in the metabolite ID column. Averaged intensities across individuals for each metabolite are given in the comparison sheet. High intensities relative to the control are given in red. Intensities with no difference compared to the control are given in blue, and low intensity metabolites yellow.

A18. DEG_additional_information.xlsx

This contains gene ID information as given in appendices A16 and A17 for the differentially expressed genes , and includes information such as the logged fold counts per million (logCPM), and the number of transcripts mapped to each gene ID per individual.

Appendix 1: Figures used to derive table of comparison between WGS and enrichment data at the end of Chapter 2

	WGS data				Enriched data				Percentage similarity (%) compared to WGS			
	First Design		Second Design		First Design		Second Design		First Design		Second Design	
	Z310	B17	Z310	B17	Z310	B17	Z310	B17	Z310	B17	Z310	B17
SNPs called in target region	5,086	5,045	9,360	9,202	6,898	7,138	14,307	9,049	26% increase	29% increase	35% increase	17% decrease
Number of heterozygous SNPs within the target region	1,387	1,467	2,637	2,500	1,952	2,013	4,288	1,346	29% increase	27% increase	39% increase	14% decrease
Percentage of heterozygous SNPs of total SNPs called (%)	29	29	27	27	28	28	30	15	1% decrease	1% decrease	3% increase	12% decrease
Number of homozygous SNPs within the target region	3,669	3,578	6,723	6,702	4,946	5,125	10,019	7,703	26% increase	30% increase	33% increase	13% increase
Percentage of homozygous SNPs of total SNPs called (%)	73	71	73	73	72	72	85	70	1% decrease	1% increase	12% increase	3% decrease
SNPs found in both enriched and WGS data	4,144 (82%)	4,123 (82%)	8,832 (94%)	5,918 (64%)	4,144 (60%)	4,123 (58%)	8,832 (62%)	5,918 (66%)	82% of WGS SNPs found in enrichment data	82% of WGS SNPs found in enrichment data	94% of WGS SNPs found in enrichment data	64% of WGS SNPs found in enrichment data
SNPs found in only the enriched data	N/A				2,754 (40%)	3,015 (42%)	5,475 (38%)	3,131 (35%)	40	42	38	36

Appendix 2: Mapping statistics for analysis in Figure 3.9. This is the percentage mapped to the Tb927 v8.1 genome in the first and second design enrichment data.

	Zymodeme	Strain	Number of mapped reads	Number of unmapped reads	% Mapped
Design one	Z310	B	4,760,494	2,710,694	64
		M	3,159,047	3,738,119	46
		T	1,194,061	4,609,905	21
	Z366	O	3,533,578	3,0159,86	54
	B17	K	1,445,293	3612983	29
		E	3,785,095	2,878,297	57
		N	3,463,064	2,720,802	56
Design two	Z310	B	31,648,671	6,789,877	82
		M	24,731,217	11,209,339	69
		T	39,880,822	11,344,354	78
	Z366	O	12,033,683	13,225,031	48
	B17	K	13,423,808	12,283,470	52
		E	4,328,556	25,019,710	15
		N	22,901,561	16,722,543	58
	Unknown	G1	127,092	20,983,344	1
		G2	269,602	36,472,240	1
		G3	164,378	23,394,980	1
		G4	193,342	28,911,314	1
		G5	432,178	33,611,120	1
		G6	258,136	34,348,612	1
		G7	20,984,509	5,618,953	79
		G9	842,923	34,764,987	2
		G10	356,892	42,823,280	1

Appendix 3: Gene IDs included within the first design

Tb927.1.1010
Tb927.1.1050
Tb927.1.1100
Tb927.1.1140
Tb927.1.1240
Tb927.1.1270
Tb927.1.1340
Tb927.1.1390
Tb927.1.1420
Tb927.1.1470
Tb927.1.1500
Tb927.1.1600
Tb927.1.1620
Tb927.1.1670
Tb927.1.1750
Tb927.1.1840
Tb927.1.1910
Tb927.1.1960
Tb927.1.2110
Tb927.1.2320
Tb927.1.2580
Tb927.1.2670
Tb927.1.2740
Tb927.1.2990
Tb927.1.30
Tb927.1.3070
Tb927.1.3120
Tb927.1.3170
Tb927.1.3260
Tb927.1.3450
Tb927.1.3550
Tb927.1.3830
Tb927.1.40
Tb927.1.4010
Tb927.1.4050
Tb927.1.4100
Tb927.1.4310
Tb927.1.4370
Tb927.1.4390
Tb927.1.4480
Tb927.1.4700
Tb927.1.4720
Tb927.1.4740
Tb927.1.4800
Tb927.1.5000
Tb927.1.630
Tb927.1.640
Tb927.1.760
Tb927.1.790
Tb927.1.840
Tb927.10.10050
Tb927.10.10140
Tb927.10.10170
Tb927.10.10340
Tb927.10.10650
Tb927.10.10700
Tb927.10.10770
Tb927.10.11010
Tb927.10.1110
Tb927.10.11170
Tb927.10.11310
Tb927.10.11420
Tb927.10.11480
Tb927.10.11580
Tb927.10.11670
Tb927.10.1190
Tb927.10.11910
Tb927.10.12020
Tb927.10.12120
Tb927.10.12160
Tb927.10.12310
Tb927.10.12390
Tb927.10.12470
Tb927.10.12540
Tb927.10.12640
Tb927.10.1270
Tb927.10.12810
Tb927.10.12990
Tb927.10.13070
Tb927.10.1310
Tb927.10.13150
Tb927.10.13180
Tb927.10.13240
Tb927.10.13330
Tb927.10.13430
Tb927.10.13550
Tb927.10.13620
Tb927.10.13740
Tb927.10.13820
Tb927.10.13920
Tb927.10.13960
Tb927.10.14040
Tb927.10.14140
Tb927.10.14230
Tb927.10.14340
Tb927.10.1440
Tb927.10.14470
Tb927.10.14530
Tb927.10.14570
Tb927.10.14870
Tb927.10.14950
Tb927.10.15040
Tb927.10.15080
Tb927.10.1510
Tb927.10.15190
Tb927.10.15280
Tb927.10.15310
Tb927.10.15490
Tb927.10.15610
Tb927.10.15680
Tb927.10.15800
Tb927.10.15810
Tb927.10.15940
Tb927.10.16040
Tb927.10.1630
Tb927.10.16400
Tb927.10.170
Tb927.10.1890
Tb927.10.1930
Tb927.10.2040
Tb927.10.210
Tb927.10.2130
Tb927.10.2220
Tb927.10.2300
Tb927.10.2350
Tb927.10.2440
Tb927.10.2520
Tb927.10.2550
Tb927.10.2600
Tb927.10.2630
Tb927.10.2700
Tb927.10.2760
Tb927.10.2820
Tb927.10.2900
Tb927.10.2960
Tb927.10.3010
Tb927.10.3080
Tb927.10.3140
Tb927.10.3170
Tb927.10.330
Tb927.10.3350
Tb927.10.3430
Tb927.10.3480
Tb927.10.3510
Tb927.10.3760
Tb927.10.3790
Tb927.10.3820
Tb927.10.3910
Tb927.10.4050
Tb927.10.4170
Tb927.10.420
Tb927.10.4200
Tb927.10.4520
Tb927.10.4610
Tb927.10.4670
Tb927.10.4770
Tb927.10.4800
Tb927.10.4900
Tb927.10.5140
Tb927.10.5220
Tb927.10.5250
Tb927.10.530
Tb927.10.5320
Tb927.10.5420
Tb927.10.5530
Tb927.10.5670
Tb927.10.5760
Tb927.10.5870
Tb927.10.5960
Tb927.10.5990
Tb927.10.6050

Tb927.10.610
Tb927.10.6150
Tb927.10.6230
Tb927.10.6300
Tb927.10.6360
Tb927.10.6470
Tb927.10.6550
Tb927.10.6690
Tb927.10.7010
Tb927.10.7110
Tb927.10.7130
Tb927.10.7230
Tb927.10.730
Tb927.10.7350
Tb927.10.7450
Tb927.10.7540
Tb927.10.7550
Tb927.10.7580
Tb927.10.760
Tb927.10.7630
Tb927.10.7700
Tb927.10.7870
Tb927.10.790
Tb927.10.7980
Tb927.10.8040
Tb927.10.8120
Tb927.10.8210
Tb927.10.8370
Tb927.10.8430
Tb927.10.8540
Tb927.10.870
Tb927.10.8710
Tb927.10.8770
Tb927.10.8820
Tb927.10.890
Tb927.10.8910
Tb927.10.9000
Tb927.10.9070
Tb927.10.9180
Tb927.10.9260
Tb927.10.9290
Tb927.10.9370
Tb927.10.9510
Tb927.10.9640
Tb927.10.9690
Tb927.10.970
Tb927.10.9720
Tb927.10.9840
Tb927.10.9940
Tb927.11.10110
Tb927.11.10280
Tb927.11.10370
Tb927.11.10520
Tb927.11.10630
Tb927.11.10690
Tb927.11.10750
Tb927.11.1080
Tb927.11.10870
Tb927.11.10990
Tb927.11.11020
Tb927.11.11180
Tb927.11.11310
Tb927.11.11430
Tb927.11.1150
Tb927.11.11510
Tb927.11.11720
Tb927.11.11780
Tb927.11.11810
Tb927.11.11900
Tb927.11.12000
Tb927.11.12110
Tb927.11.1230
Tb927.11.12320
Tb927.11.12430
Tb927.11.12530
Tb927.11.12580
Tb927.11.12660
Tb927.11.1280
Tb927.11.12850
Tb927.11.12930
Tb927.11.13060
Tb927.11.13150
Tb927.11.13250
Tb927.11.13290
Tb927.11.13360
Tb927.11.1340
Tb927.11.13500
Tb927.11.13730
Tb927.11.13770
Tb927.11.1400
Tb927.11.14060
Tb927.11.14110
Tb927.11.14190
Tb927.11.14370
Tb927.11.14450
Tb927.11.14530
Tb927.11.14560
Tb927.11.1460
Tb927.11.14630
Tb927.11.14720
Tb927.11.14890
Tb927.11.14920
Tb927.11.14990
Tb927.11.15100
Tb927.11.15190
Tb927.11.15310
Tb927.11.15330
Tb927.11.15390
Tb927.11.15400
Tb927.11.15490
Tb927.11.1560
Tb927.11.15610
Tb927.11.15730
Tb927.11.15800
Tb927.11.15940
Tb927.11.16010
Tb927.11.16110
Tb927.11.16210
Tb927.11.16300
Tb927.11.16400
Tb927.11.16570
Tb927.11.16650
Tb927.11.16700
Tb927.11.16770
Tb927.11.16840.1
Tb927.11.16930
Tb927.11.170
Tb927.11.17020
Tb927.11.1740
Tb927.11.1900
Tb927.11.1920
Tb927.11.2020
Tb927.11.2140
Tb927.11.2210
Tb927.11.2280
Tb927.11.2490
Tb927.11.2530
Tb927.11.260
Tb927.11.2740
Tb927.11.2800
Tb927.11.2830
Tb927.11.2920
Tb927.11.3010
Tb927.11.3120
Tb927.11.3190
Tb927.11.3240
Tb927.11.3270
Tb927.11.3310
Tb927.11.3380
Tb927.11.3420
Tb927.11.3470
Tb927.11.3550
Tb927.11.360
Tb927.11.3650
Tb927.11.3710
Tb927.11.3770
Tb927.11.3840
Tb927.11.3910
Tb927.11.4060
Tb927.11.4210
Tb927.11.4280
Tb927.11.4350
Tb927.11.440
Tb927.11.4430
Tb927.11.4510
Tb927.11.460
Tb927.11.4610
Tb927.11.4620
Tb927.11.4660
Tb927.11.4690
Tb927.11.4760
Tb927.11.4800
Tb927.11.4950
Tb927.11.5000
Tb927.11.5110
Tb927.11.5210
Tb927.11.5330

Tb927.11.5370
Tb927.11.540
Tb927.11.5420
Tb927.11.5490
Tb927.11.5590
Tb927.11.5700
Tb927.11.5750
Tb927.11.5800
Tb927.11.5910
Tb927.11.6020
Tb927.11.6100
Tb927.11.6130
Tb927.11.6250
Tb927.11.6460
Tb927.11.6550
Tb927.11.6630
Tb927.11.6720
Tb927.11.6810
Tb927.11.6830
Tb927.11.6920
Tb927.11.700
Tb927.11.7050
Tb927.11.7130
Tb927.11.7211
Tb927.11.7218
Tb927.11.7320
Tb927.11.7330
Tb927.11.7400
Tb927.11.7590
Tb927.11.780
Tb927.11.7890
Tb927.11.8010
Tb927.11.8050
Tb927.11.810
Tb927.11.8150
Tb927.11.8210
Tb927.11.8310
Tb927.11.8390
Tb927.11.8430
Tb927.11.8830
Tb927.11.8930
Tb927.11.8960
Tb927.11.9050
Tb927.11.910
Tb927.11.9150
Tb927.11.9220
Tb927.11.9290
Tb927.11.9330
Tb927.11.9350
Tb927.11.9420
Tb927.11.9590
Tb927.11.9690
Tb927.11.980
Tb927.11.9810
Tb927.11.9840
Tb927.11.9950
Tb927.2.1380
Tb927.2.1600
Tb927.2.1700
Tb927.2.1730
Tb927.2.1780
Tb927.2.1860
Tb927.2.1920
Tb927.2.2020
Tb927.2.2060
Tb927.2.2180
Tb927.2.2230
Tb927.2.2240
Tb927.2.2270
Tb927.2.2370
Tb927.2.2410
Tb927.2.2490
Tb927.2.2540
Tb927.2.2550
Tb927.2.2580
Tb927.2.2650
Tb927.2.2720
Tb927.2.2830
Tb927.2.2950
Tb927.2.3080
Tb927.2.3340
Tb927.2.3720
Tb927.2.3730
Tb927.2.3780
Tb927.2.3910
Tb927.2.4000
Tb927.2.4130
Tb927.2.4200
Tb927.2.4210
Tb927.2.4380
Tb927.2.4460
Tb927.2.4470
Tb927.2.4550
Tb927.2.4670
Tb927.2.4720
Tb927.2.4840
Tb927.2.4950
Tb927.2.5020
Tb927.2.5050
Tb927.2.5130
Tb927.2.5150
Tb927.2.5240
Tb927.2.5660
Tb927.2.5750
Tb927.2.5810
Tb927.2.5820
Tb927.2.5890
Tb927.2.5900
Tb927.2.5930
Tb927.2.6050
Tb927.2.6080
Tb927.2.6130
Tb927.3.1070
Tb927.3.1090
Tb927.3.1140
Tb927.3.1200
Tb927.3.1220
Tb927.3.1260
Tb927.3.1320
Tb927.3.1540
Tb927.3.1550
Tb927.3.1590
Tb927.3.1710
Tb927.3.1800
Tb927.3.1810
Tb927.3.1840
Tb927.3.1850
Tb927.3.1860
Tb927.3.1880
Tb927.3.1910
Tb927.3.1940
Tb927.3.2220
Tb927.3.2240
Tb927.3.2380
Tb927.3.2460
Tb927.3.2660
Tb927.3.2680
Tb927.3.2690
Tb927.3.2790
Tb927.3.2860
Tb927.3.3000
Tb927.3.3020
Tb927.3.3050
Tb927.3.3090
Tb927.3.3130
Tb927.3.3220
Tb927.3.3300
Tb927.3.3410
Tb927.3.3460
Tb927.3.3540
Tb927.3.3560
Tb927.3.3670
Tb927.3.3850
Tb927.3.3910
Tb927.3.4150
Tb927.3.4220
Tb927.3.4270
Tb927.3.4370
Tb927.3.4420
Tb927.3.4490
Tb927.3.4550
Tb927.3.4610
Tb927.3.4630
Tb927.3.4680
Tb927.3.4850
Tb927.3.4950
Tb927.3.4960
Tb927.3.5000
Tb927.3.5050
Tb927.3.5120
Tb927.3.5220
Tb927.3.5360
Tb927.3.5420
Tb927.3.5510
Tb927.3.5540
Tb927.3.5620
Tb927.3.5660

Tb927.3.580
Tb927.3.600
Tb927.3.630
Tb927.3.640
Tb927.3.670
Tb927.3.700
Tb927.3.740
Tb927.3.770
Tb927.3.790
Tb927.3.820
Tb927.3.880
Tb927.3.920
Tb927.3.950
Tb927.3.970
Tb927.4.1040
Tb927.4.1160
Tb927.4.1230
Tb927.4.1280
Tb927.4.1390
Tb927.4.1440
Tb927.4.1510
Tb927.4.1570
Tb927.4.1670
Tb927.4.1750
Tb927.4.1840
Tb927.4.1950
Tb927.4.1990
Tb927.4.2020
Tb927.4.2140
Tb927.4.2210
Tb927.4.2340
Tb927.4.2410
Tb927.4.2490
Tb927.4.2560
Tb927.4.2620
Tb927.4.2680
Tb927.4.2750
Tb927.4.2850
Tb927.4.2940
Tb927.4.3030
Tb927.4.3040
Tb927.4.3120
Tb927.4.3170
Tb927.4.320
Tb927.4.3480
Tb927.4.3730
Tb927.4.3770
Tb927.4.380
Tb927.4.3840
Tb927.4.3980
Tb927.4.4170
Tb927.4.4230
Tb927.4.450
Tb927.4.4500
Tb927.4.4590
Tb927.4.4640
Tb927.4.4680
Tb927.4.4760
Tb927.4.5130
Tb927.4.520
Tb927.4.570
Tb927.4.610
Tb927.4.680
Tb927.4.760
Tb927.4.770
Tb927.4.820
Tb927.4.880
Tb927.4.930
Tb927.5.1030
Tb927.5.1090
Tb927.5.1130
Tb927.5.1200
Tb927.5.1260
Tb927.5.1310
Tb927.5.1490
Tb927.5.1560
Tb927.5.1600
Tb927.5.1680
Tb927.5.1770
Tb927.5.1910
Tb927.5.1970
Tb927.5.2010
Tb927.5.2050
Tb927.5.2120
Tb927.5.2270
Tb927.5.2300
Tb927.5.2340
Tb927.5.2470
Tb927.5.2510
Tb927.5.2570
Tb927.5.2650
Tb927.5.270
Tb927.5.2800
Tb927.5.2820
Tb927.5.2930
Tb927.5.3030
Tb927.5.3150
Tb927.5.3170
Tb927.5.3200
Tb927.5.3230
Tb927.5.3260
Tb927.5.3310
Tb927.5.3370
Tb927.5.3450
Tb927.5.3540
Tb927.5.3620
Tb927.5.3690
Tb927.5.3810
Tb927.5.3820
Tb927.5.3870
Tb927.5.3950
Tb927.5.4090
Tb927.5.4160
Tb927.5.4290
Tb927.5.4360
Tb927.5.4400
Tb927.5.4480
Tb927.5.4560
Tb927.5.530
Tb927.5.560
Tb927.5.570
Tb927.5.620
Tb927.5.690
Tb927.5.760
Tb927.5.830
Tb927.5.840
Tb927.5.920
Tb927.5.960
Tb927.6.1090
Tb927.6.1120
Tb927.6.1220
Tb927.6.1470
Tb927.6.1550
Tb927.6.1660
Tb927.6.1770
Tb927.6.1810
Tb927.6.1880
Tb927.6.2020
Tb927.6.2080
Tb927.6.210
Tb927.6.2150
Tb927.6.2230
Tb927.6.2390
Tb927.6.2510
Tb927.6.2630
Tb927.6.2770
Tb927.6.2840
Tb927.6.2960
Tb927.6.3050
Tb927.6.3170
Tb927.6.3260
Tb927.6.3430
Tb927.6.3490
Tb927.6.3580
Tb927.6.360
Tb927.6.3710
Tb927.6.3820
Tb927.6.3980
Tb927.6.4070
Tb927.6.410
Tb927.6.4150
Tb927.6.4200
Tb927.6.4340
Tb927.6.440
Tb927.6.4480
Tb927.6.4600
Tb927.6.4730
Tb927.6.4770
Tb927.6.4830
Tb927.6.4970
Tb927.6.5060
Tb927.6.5100
Tb927.6.5150
Tb927.6.660
Tb927.6.700
Tb927.6.740
Tb927.6.810

Tb927.6.860
Tb927.6.900
Tb927.6.950
Tb927.7.1090
Tb927.7.1150
Tb927.7.1190
Tb927.7.1410
Tb927.7.1490
Tb927.7.1610
Tb927.7.1770
Tb927.7.2080
Tb927.7.2160
Tb927.7.2290
Tb927.7.2320
Tb927.7.2460
Tb927.7.2560
Tb927.7.2680
Tb927.7.2760
Tb927.7.290
Tb927.7.2960
Tb927.7.3040
Tb927.7.3090
Tb927.7.3190
Tb927.7.3320
Tb927.7.3360
Tb927.7.340
Tb927.7.3460
Tb927.7.3550
Tb927.7.3740
Tb927.7.3850
Tb927.7.3970
Tb927.7.4020
Tb927.7.4120
Tb927.7.4150
Tb927.7.4330
Tb927.7.4480
Tb927.7.4530
Tb927.7.460
Tb927.7.4640
Tb927.7.4760
Tb927.7.4860
Tb927.7.4910
Tb927.7.4950
Tb927.7.500
Tb927.7.5060
Tb927.7.5140
Tb927.7.5220
Tb927.7.5240
Tb927.7.5360
Tb927.7.540
Tb927.7.5460
Tb927.7.5550
Tb927.7.5560
Tb927.7.5590
Tb927.7.5720
Tb927.7.5790
Tb927.7.590
Tb927.7.5920
Tb927.7.6090
Tb927.7.6230
Tb927.7.6300
Tb927.7.6400
Tb927.7.6440
Tb927.7.6690
Tb927.7.680
Tb927.7.6930
Tb927.7.6990
Tb927.7.7010
Tb927.7.7300
Tb927.7.7370
Tb927.7.770
Tb927.7.810
Tb927.7.860
Tb927.8.1140
Tb927.8.1150
Tb927.8.1250
Tb927.8.1270
Tb927.8.1380
Tb927.8.1440
Tb927.8.1550
Tb927.8.1700
Tb927.8.1770
Tb927.8.1820
Tb927.8.2050
Tb927.8.2190
Tb927.8.2350
Tb927.8.2390
Tb927.8.2480
Tb927.8.2580
Tb927.8.2640
Tb927.8.2670
Tb927.8.2710
Tb927.8.2760
Tb927.8.2800
Tb927.8.2810
Tb927.8.2870
Tb927.8.2960
Tb927.8.3010
Tb927.8.3050
Tb927.8.3150
Tb927.8.3240
Tb927.8.3310
Tb927.8.3350
Tb927.8.3400
Tb927.8.3530
Tb927.8.3660
Tb927.8.3730
Tb927.8.3830
Tb927.8.3870
Tb927.8.4150
Tb927.8.4190
Tb927.8.4210
Tb927.8.4260
Tb927.8.4350
Tb927.8.4390
Tb927.8.4500
Tb927.8.4520
Tb927.8.4620
Tb927.8.4690
Tb927.8.4780
Tb927.8.4870
Tb927.8.5040
Tb927.8.5050
Tb927.8.5090
Tb927.8.5140
Tb927.8.5240
Tb927.8.5310
Tb927.8.5350
Tb927.8.5380
Tb927.8.540
Tb927.8.5410
Tb927.8.5510
Tb927.8.5530
Tb927.8.5600
Tb927.8.5740
Tb927.8.5790
Tb927.8.5870
Tb927.8.590
Tb927.8.5980
Tb927.8.6090
Tb927.8.6200
Tb927.8.6250
Tb927.8.6350
Tb927.8.6410
Tb927.8.6520
Tb927.8.6560
Tb927.8.6580
Tb927.8.6690
Tb927.8.670
Tb927.8.6920
Tb927.8.720
Tb927.8.7200
Tb927.8.7280
Tb927.8.7570
Tb927.8.7770
Tb927.8.7840
Tb927.8.790
Tb927.8.7970
Tb927.8.8010
Tb927.8.8030
Tb927.8.8130
Tb927.8.830
Tb927.8.8310
Tb927.8.880
Tb927.8.910
Tb927.9.10040
Tb927.9.10160
Tb927.9.10200
Tb927.9.10440
Tb927.9.10490
Tb927.9.10530
Tb927.9.10580
Tb927.9.10670
Tb927.9.10690
Tb927.9.10840
Tb927.9.11050
Tb927.9.11110

Tb927.9.11250
Tb927.9.11540
Tb927.9.11670
Tb927.9.11830
Tb927.9.11890
Tb927.9.12090
Tb927.9.12280
Tb927.9.12390
Tb927.9.12440
Tb927.9.12500
Tb927.9.12700
Tb927.9.12900
Tb927.9.12980
Tb927.9.13010
Tb927.9.13150
Tb927.9.1320
Tb927.9.13330
Tb927.9.13440
Tb927.9.13510
Tb927.9.13610
Tb927.9.13780
Tb927.9.1400
Tb927.9.14190
Tb927.9.14330
Tb927.9.14440
Tb927.9.14620
Tb927.9.14960
Tb927.9.15090
Tb927.9.15290
Tb927.9.1530
Tb927.9.15400
Tb927.9.15620
Tb927.9.15690
Tb927.9.15850
Tb927.9.16010
Tb927.9.1640
Tb927.9.1770
Tb927.9.18100
Tb927.9.1940
Tb927.9.1970
Tb927.9.2070
Tb927.9.2110
Tb927.9.2220
Tb927.9.2240
Tb927.9.2260
Tb927.9.2320
Tb927.9.2390
Tb927.9.2560
Tb927.9.2590
Tb927.9.2650
Tb927.9.2700
Tb927.9.2760
Tb927.9.2900
Tb927.9.3080
Tb927.9.3400
Tb927.9.3460
Tb927.9.3470
Tb927.9.3680
Tb927.9.3820
Tb927.9.3860
Tb927.9.4080
Tb927.9.4130
Tb927.9.4300
Tb927.9.4370
Tb927.9.4500
Tb927.9.4560
Tb927.9.4640
Tb927.9.4760
Tb927.9.4900
Tb927.9.5170
Tb927.9.5210
Tb927.9.5410
Tb927.9.5520
Tb927.9.5620
Tb927.9.5900
Tb927.9.6110
Tb927.9.6320
Tb927.9.6430
Tb927.9.6530
Tb927.9.6580
Tb927.9.6720
Tb927.9.7030
Tb927.9.7150
Tb927.9.7200
Tb927.9.7290
Tb927.9.7690
Tb927.9.7720
Tb927.9.7740
Tb927.9.7760
Tb927.9.7810
Tb927.9.7830
Tb927.9.7960
Tb927.9.8000
Tb927.9.8160
Tb927.9.8260
Tb927.9.8400
Tb927.9.8510
Tb927.9.8700
Tb927.9.8950
Tb927.9.9000
Tb927.9.9120
Tb927.9.9220
Tb927.9.9310
Tb927.9.9600
Tb927.9.9720
Tb927.9.9810

Appendix 4: Gene IDs used in design two

Tb927.11.12850	Tb927.10.12310	Tb927.10.3510
Tb927.1.1010	Tb927.10.12390	Tb927.10.3760
Tb927.1.1050	Tb927.10.12470	Tb927.10.3790
Tb927.1.1140	Tb927.10.12640	Tb927.10.3820
Tb927.1.1240	Tb927.10.1270	Tb927.10.3910
Tb927.1.1270	Tb927.10.12810	Tb927.10.4050
Tb927.1.1390	Tb927.10.12990	Tb927.10.4170
Tb927.1.1420	Tb927.10.13070	Tb927.10.420
Tb927.1.1470	Tb927.10.1310	Tb927.10.4200
Tb927.1.1500	Tb927.10.13180	Tb927.10.4520
Tb927.1.1600	Tb927.10.13240	Tb927.10.4610
Tb927.1.1620	Tb927.10.13330	Tb927.10.4670
Tb927.1.1750	Tb927.10.13430	Tb927.10.4770
Tb927.1.1840	Tb927.10.13740	Tb927.10.5140
Tb927.1.1910	Tb927.10.13820	Tb927.10.5220
Tb927.1.1960	Tb927.10.13960	Tb927.10.5250
Tb927.1.2110	Tb927.10.14140	Tb927.10.530
Tb927.1.2320	Tb927.10.14230	Tb927.10.5320
Tb927.1.2580	Tb927.10.14470	Tb927.10.5420
Tb927.1.2670	Tb927.10.14530	Tb927.10.5530
Tb927.1.2740	Tb927.10.14870	Tb927.10.5670
Tb927.1.2990	Tb927.10.14950	Tb927.10.5760
Tb927.1.3070	Tb927.10.15040	Tb927.10.5870
Tb927.1.3120	Tb927.10.15080	Tb927.10.5960
Tb927.1.3170	Tb927.10.1510	Tb927.10.5990
Tb927.1.3450	Tb927.10.15190	Tb927.10.6050
Tb927.1.3550	Tb927.10.15310	Tb927.10.610
Tb927.1.3830	Tb927.10.15490	Tb927.10.6150
Tb927.1.4010	Tb927.10.15610	Tb927.10.6360
Tb927.1.4050	Tb927.10.15680	Tb927.10.6470
Tb927.1.4100	Tb927.10.15800	Tb927.10.6550
Tb927.1.4310	Tb927.10.15810	Tb927.10.7110
Tb927.1.4370	Tb927.10.15940	Tb927.10.7130
Tb927.1.4480	Tb927.10.16040	Tb927.10.7230
Tb927.1.4700	Tb927.10.1630	Tb927.10.730
Tb927.1.4720	Tb927.10.16400	Tb927.10.7350
Tb927.1.4740	Tb927.10.1890	Tb927.10.7450
Tb927.1.4800	Tb927.10.1930	Tb927.10.7540
Tb927.1.5000	Tb927.10.2220	Tb927.10.7550
Tb927.1.630	Tb927.10.2300	Tb927.10.7580
Tb927.1.640	Tb927.10.2350	Tb927.10.760
Tb927.1.760	Tb927.10.2520	Tb927.10.7630
Tb927.1.790	Tb927.10.2550	Tb927.10.7700
Tb927.1.840	Tb927.10.2600	Tb927.10.7870
Tb927.10.10050	Tb927.10.2630	Tb927.10.790
Tb927.10.10650	Tb927.10.2700	Tb927.10.7980
Tb927.10.10700	Tb927.10.2760	Tb927.10.8040
Tb927.10.10770	Tb927.10.2820	Tb927.10.8120
Tb927.10.11010	Tb927.10.2900	Tb927.10.8210
Tb927.10.1110	Tb927.10.2960	Tb927.10.8370
Tb927.10.11170	Tb927.10.3010	Tb927.10.8430
Tb927.10.11420	Tb927.10.3080	Tb927.10.8540
Tb927.10.11480	Tb927.10.3140	Tb927.10.870
Tb927.10.11580	Tb927.10.3170	Tb927.10.8710
Tb927.10.11670	Tb927.10.3350	Tb927.10.8770
Tb927.10.12020	Tb927.10.3430	Tb927.10.8820
Tb927.10.12120	Tb927.10.3480	Tb927.10.890

Tb927.10.8910
Tb927.10.9000
Tb927.10.9070
Tb927.10.9180
Tb927.10.9290
Tb927.10.9370
Tb927.10.9640
Tb927.10.9690
Tb927.10.970
Tb927.10.9720
Tb927.10.9840
Tb927.11.1080
Tb927.11.1150
Tb927.11.1230
Tb927.11.1280
Tb927.11.1340
Tb927.11.1400
Tb927.11.1460
Tb927.11.1560
Tb927.11.1900
Tb927.11.2020
Tb927.11.2140
Tb927.11.2210
Tb927.11.2280
Tb927.11.2490
Tb927.11.2530
Tb927.11.2800
Tb927.11.2830
Tb927.11.3010
Tb927.11.3120
Tb927.11.3270
Tb927.11.3310
Tb927.11.3420
Tb927.11.3550
Tb927.11.3650
Tb927.11.3710
Tb927.11.3770
Tb927.11.3840
Tb927.11.3910
Tb927.11.4210
Tb927.11.4280
Tb927.11.4350
Tb927.11.440
Tb927.11.460
Tb927.11.4610
Tb927.11.4620
Tb927.11.4660
Tb927.11.4690
Tb927.11.4760
Tb927.11.4800
Tb927.11.4950
Tb927.11.5000
Tb927.11.5210
Tb927.11.5330
Tb927.11.540
Tb927.11.5420
Tb927.11.5490
Tb927.11.5590
Tb927.11.5700
Tb927.11.5750
Tb927.11.5800
Tb927.11.5910
Tb927.11.6020
Tb927.11.6100
Tb927.11.6130
Tb927.11.6250
Tb927.11.6460
Tb927.11.6550
Tb927.11.6630
Tb927.11.6720
Tb927.11.6810
Tb927.11.6830
Tb927.11.700
Tb927.11.7050
Tb927.11.7130
Tb927.11.7211
Tb927.11.7320
Tb927.11.7330
Tb927.11.7400
Tb927.11.7590
Tb927.11.780
Tb927.11.7890
Tb927.11.8150
Tb927.11.8210
Tb927.11.8310
Tb927.11.8390
Tb927.11.8430
Tb927.11.8830
Tb927.11.8930
Tb927.11.8960
Tb927.11.910
Tb927.11.980
Tb927.2.1380
Tb927.2.1600
Tb927.2.1700
Tb927.2.1730
Tb927.2.1780
Tb927.2.1860
Tb927.2.1920
Tb927.2.2020
Tb927.2.2060
Tb927.2.2180
Tb927.2.2230
Tb927.2.2240
Tb927.2.2270
Tb927.2.2370
Tb927.2.2410
Tb927.2.2490
Tb927.2.2580
Tb927.2.2650
Tb927.2.2720
Tb927.2.2830
Tb927.2.2950
Tb927.2.3080
Tb927.2.3340
Tb927.2.3720
Tb927.2.3730
Tb927.2.3780
Tb927.2.3910
Tb927.2.4000
Tb927.2.4130
Tb927.2.4200
Tb927.2.4210
Tb927.2.4380
Tb927.2.4460
Tb927.2.4470
Tb927.2.4670
Tb927.2.4840
Tb927.2.4950
Tb927.2.5130
Tb927.2.5660
Tb927.2.5750
Tb927.2.5810
Tb927.2.5820
Tb927.2.5890
Tb927.2.5900
Tb927.2.5930
Tb927.2.6050
Tb927.2.6080
Tb927.2.6130
Tb927.3.1070
Tb927.3.1090
Tb927.3.1140
Tb927.3.1200
Tb927.3.1220
Tb927.3.1260
Tb927.3.1320
Tb927.3.1540
Tb927.3.1550
Tb927.3.1710
Tb927.3.1800
Tb927.3.1810
Tb927.3.1840
Tb927.3.1850
Tb927.3.1860
Tb927.3.1910
Tb927.3.1940
Tb927.3.2240
Tb927.3.2380
Tb927.3.2460
Tb927.3.2660
Tb927.3.2680
Tb927.3.2690
Tb927.3.2860
Tb927.3.3000
Tb927.3.3020
Tb927.3.3050
Tb927.3.3090
Tb927.3.3130
Tb927.3.3220
Tb927.3.3410
Tb927.3.3540
Tb927.3.3560
Tb927.3.3670
Tb927.3.3850
Tb927.3.4150
Tb927.3.4220

Tb927.3.4270
Tb927.3.4420
Tb927.3.4550
Tb927.3.4610
Tb927.3.4630
Tb927.3.4680
Tb927.3.4950
Tb927.3.4960
Tb927.3.5000
Tb927.3.5120
Tb927.3.5220
Tb927.3.5420
Tb927.3.5510
Tb927.3.5540
Tb927.3.580
Tb927.3.600
Tb927.3.630
Tb927.3.640
Tb927.3.670
Tb927.3.700
Tb927.3.740
Tb927.3.770
Tb927.3.790
Tb927.3.820
Tb927.3.880
Tb927.3.920
Tb927.3.950
Tb927.3.970
Tb927.4.1040
Tb927.4.1230
Tb927.4.1390
Tb927.4.1510
Tb927.4.1570
Tb927.4.1670
Tb927.4.1750
Tb927.4.1840
Tb927.4.1950
Tb927.4.1990
Tb927.4.2140
Tb927.4.2210
Tb927.4.2340
Tb927.4.2410
Tb927.4.2490
Tb927.4.2620
Tb927.4.2680
Tb927.4.2750
Tb927.4.2850
Tb927.4.3030
Tb927.4.3040
Tb927.4.3120
Tb927.4.3170
Tb927.4.320
Tb927.4.3480
Tb927.4.3730
Tb927.4.3770
Tb927.4.380
Tb927.4.3980
Tb927.4.4170
Tb927.4.4230
Tb927.4.450
Tb927.4.520
Tb927.4.570
Tb927.4.680
Tb927.4.760
Tb927.4.770
Tb927.4.820
Tb927.4.930
Tb927.5.1030
Tb927.5.1090
Tb927.5.1130
Tb927.5.1260
Tb927.5.1310
Tb927.5.1490
Tb927.5.1560
Tb927.5.1600
Tb927.5.1680
Tb927.5.1770
Tb927.5.1910
Tb927.5.1970
Tb927.5.2010
Tb927.5.2050
Tb927.5.2120
Tb927.5.2270
Tb927.5.2340
Tb927.5.2470
Tb927.5.2510
Tb927.5.2570
Tb927.5.2650
Tb927.5.270
Tb927.5.2800
Tb927.5.2820
Tb927.5.2930
Tb927.5.3030
Tb927.5.3150
Tb927.5.3200
Tb927.5.3230
Tb927.5.3260
Tb927.5.3310
Tb927.5.3370
Tb927.5.530
Tb927.5.560
Tb927.5.570
Tb927.5.620
Tb927.5.760
Tb927.5.830
Tb927.5.840
Tb927.5.960
Tb927.6.1090
Tb927.6.1120
Tb927.6.1220
Tb927.6.1550
Tb927.6.1660
Tb927.6.1770
Tb927.6.1810
Tb927.6.1880
Tb927.6.210
Tb927.6.2230
Tb927.6.2510
Tb927.6.2630
Tb927.6.2770
Tb927.6.2840
Tb927.6.2960
Tb927.6.3050
Tb927.6.3170
Tb927.6.3260
Tb927.6.3430
Tb927.6.3490
Tb927.6.3580
Tb927.6.360
Tb927.6.3710
Tb927.6.3820
Tb927.6.3980
Tb927.6.4070
Tb927.6.410
Tb927.6.4150
Tb927.6.4200
Tb927.6.4340
Tb927.6.440
Tb927.6.4480
Tb927.6.4600
Tb927.6.4730
Tb927.6.4770
Tb927.6.4830
Tb927.6.4970
Tb927.6.5060
Tb927.6.5100
Tb927.6.5150
Tb927.6.660
Tb927.6.700
Tb927.6.810
Tb927.6.860
Tb927.6.950
Tb927.7.1090
Tb927.7.1150
Tb927.7.1190
Tb927.7.1410
Tb927.7.1610
Tb927.7.1770
Tb927.7.2080
Tb927.7.2320
Tb927.7.2460
Tb927.7.2560
Tb927.7.2680
Tb927.7.2960
Tb927.7.3040
Tb927.7.3090
Tb927.7.3190
Tb927.7.3320
Tb927.7.3460
Tb927.7.3550
Tb927.7.3740
Tb927.7.3970
Tb927.7.4020
Tb927.7.4120
Tb927.7.4150
Tb927.7.4330
Tb927.7.4480

Tb927.7.4530
Tb927.7.4640
Tb927.7.4760
Tb927.7.4860
Tb927.7.4910
Tb927.7.4950
Tb927.7.500
Tb927.7.5060
Tb927.7.5220
Tb927.7.5360
Tb927.7.540
Tb927.7.5550
Tb927.7.5560
Tb927.7.5720
Tb927.7.5790
Tb927.7.590
Tb927.7.5920
Tb927.7.6090
Tb927.7.6230
Tb927.7.6300
Tb927.7.6400
Tb927.7.6440
Tb927.7.6690
Tb927.7.680
Tb927.7.6930
Tb927.7.6990
Tb927.7.7010
Tb927.7.7300
Tb927.7.7370
Tb927.7.770
Tb927.7.860
Tb927.8.1140
Tb927.8.1150
Tb927.8.1250
Tb927.8.1270
Tb927.8.1550
Tb927.8.1700
Tb927.8.1770
Tb927.8.1820
Tb927.8.2190
Tb927.8.2350
Tb927.8.2390
Tb927.8.2640
Tb927.8.2670
Tb927.8.2870
Tb927.8.2960
Tb927.8.3010
Tb927.8.3050
Tb927.8.3240
Tb927.8.3310
Tb927.8.3350
Tb927.8.3400
Tb927.8.3530
Tb927.8.3660
Tb927.8.3730
Tb927.8.3830
Tb927.8.3870
Tb927.8.4150
Tb927.8.4190
Tb927.8.4210
Tb927.8.4260
Tb927.8.4350
Tb927.8.4390
Tb927.8.4500
Tb927.8.4520
Tb927.8.4620
Tb927.8.4690
Tb927.8.4780
Tb927.8.4870
Tb927.8.5050
Tb927.8.5090
Tb927.8.5140
Tb927.8.5240
Tb927.8.5310
Tb927.8.5350
Tb927.8.5380
Tb927.8.5510
Tb927.8.5530
Tb927.8.5740
Tb927.8.5790
Tb927.8.5870
Tb927.8.6090
Tb927.8.6200
Tb927.8.6250
Tb927.8.6350
Tb927.8.6410
Tb927.8.6520
Tb927.8.6580
Tb927.8.6690
Tb927.8.6920
Tb927.8.720
Tb927.8.7200
Tb927.8.7570
Tb927.8.7770
Tb927.8.7840
Tb927.8.790
Tb927.8.7970
Tb927.8.8010
Tb927.8.8030
Tb927.8.8130
Tb927.8.830
Tb927.8.8310
Tb927.8.880
Tb927.8.910
Tb927.9.10040
Tb927.9.10160
Tb927.9.10200
Tb927.9.10440
Tb927.9.10490
Tb927.9.10530
Tb927.9.10580
Tb927.9.10670
Tb927.9.10690
Tb927.9.10840
Tb927.9.11050
Tb927.9.11110
Tb927.9.11540
Tb927.9.11670
Tb927.9.11830
Tb927.9.11890
Tb927.9.12090
Tb927.9.12280
Tb927.9.12390
Tb927.9.12440
Tb927.9.12500
Tb927.9.12700
Tb927.9.12900
Tb927.9.12980
Tb927.9.13010
Tb927.9.13150
Tb927.9.13330
Tb927.9.13440
Tb927.9.13510
Tb927.9.13610
Tb927.9.13780
Tb927.9.14190
Tb927.9.14330
Tb927.9.14440
Tb927.9.14960
Tb927.9.15090
Tb927.9.15400
Tb927.9.15620
Tb927.9.15690
Tb927.9.15850
Tb927.9.16010
Tb927.9.1640
Tb927.9.1770
Tb927.9.18100
Tb927.9.2070
Tb927.9.2110
Tb927.9.2220
Tb927.9.2240
Tb927.9.2260
Tb927.9.2320
Tb927.9.2390
Tb927.9.2560
Tb927.9.2590
Tb927.9.2650
Tb927.9.2700
Tb927.9.2760
Tb927.9.2900
Tb927.9.3080
Tb927.9.3400
Tb927.9.3460
Tb927.9.3470
Tb927.9.3820
Tb927.9.4080
Tb927.9.4130
Tb927.9.4300
Tb927.9.4500
Tb927.9.4640
Tb927.9.4760
Tb927.9.4900
Tb927.9.5170
Tb927.9.5210
Tb927.9.5410
Tb927.9.5520

Tb927.9.5620
Tb927.9.5900
Tb927.9.6110
Tb927.9.6320
Tb927.9.6430
Tb927.9.6530
Tb927.9.6580
Tb927.9.7150
Tb927.9.7200
Tb927.9.7290
Tb927.9.7690
Tb927.9.7720
Tb927.9.7740
Tb927.9.7760
Tb927.9.7810
Tb927.9.7830
Tb927.9.8000
Tb927.9.8160
Tb927.9.8260
Tb927.9.8400
Tb927.9.8510
Tb927.9.8700
Tb927.9.8950
Tb927.9.9000
Tb927.9.9120
Tb927.9.9220
Tb927.9.9310
Tb927.9.9720
Tb927.9.9810

Appendix 5: Raw parasite counts from tail snip bleeds during infections used for microscopy screening and QPCR data. These were counts taken from 5µl volumes, per ml parasitaemia is 200 x these total counts for example a total count of 50 parasites represents 1 x 10^4 parasites per ml. Short stumpy is represented by SS, long slender by LS, and intermediate by I. Raw counts are given per stage, the percentage of parasites in this stage is also given. Counts beneath 20 are shown in bold and represent a parasitaemia of less than 4000 parasites per ml. Individuals B1-B5 were infected with Z310 parasites, strain B, E1-E5, with B17 parasites, strain E. Instances where no parasites were counted are indicated.

Day post infection		Zymodeme infected with									
		Z310					B17				
		B1	B2	B3	B4	B5	E1	E2	E3	E4	E5
3	SS counts	2	5	3	1	1	No parasites detected				
	SS percentage	6.45	8.93	7.69	3.45	2.56					
	LS counts	29	51	35	25	34					
	LS percentage	93.54	91.07	89.74	86.21	87.18					
	I counts	0	0	1	3	4					
	I percentage	0	0	2.56	10.34	10.26					
	Total count	31	56	39	29	39					
4	SS counts	3	3	0	0	3					
	SS percentage	5.56	6.67	0	0	7.5					
	LS counts	51	42	50	38	36					
	LS percentage	94.44	93.33	100	100	90					
	I counts	0	0	0	0	1					
	I percentage	0	0	0	0	2.5					
	Total count	54	45	50	38	40					
5	SS counts	2	2	3	3	3	6	0	0	3	0
	SS percentage	5.71	5.71	7.69	8.57	10	85.71	0	0	75	0
	LS counts	32	32	33	31	25	1	0	0	0	0
	LS percentage	91.43	91.43	84.62	88.57	83.33	14.29	0	0	0	0
	I counts	1	1	3	1	2	0	0	0	1	0
	I percentage	2.86	2.86	7.69	2.86	6.67	0	0	0	25	0
	Total count	35	35	39	35	30	**7**	0	0	**4**	0
6	SS counts	3	1	3	1	0	21	1	7	37	34
	SS percentage	13.63	14.28	7.69	9.09	0	63.64	100	87.5	90.24	91.89
	LS counts	19	6	33	10	1	6	0	1	4	3
	LS percentage	86.36	85.71	84.62	90.91	100	18.18	0	12.5	9.76	8.11
	I counts	0	0	3	0	0	6	0	0	0	0
	I percentage	0	0	7.69	0	0	18.18	0	0	0	0
	Total count	22	**7**	39	**11**	**1**	33	**1**	**8**	41	37
7	SS counts	0	No parasites detected				30	25	35	50	50
	SS percentage	0					90.91	73.53	85.37	100	94.34
	LS counts	1					3	5	3	0	3
	LS percentage	100					9.09	14.71	7.32	0	5.66
	I counts	0					0	4	3	0	0
	I percentage	0					0	11.76	7.32	0	0
	Total count	**1**					33	34	41	50	53
8	SS counts	7	5	No parasites detected			0	30	0	2	7
	SS percentage	46.67	71.43				0	96.77	0	100	100
	LS counts	7	1				0	1	0	0	0
	LS percentage	46.67	14.29				0	3.23	0	0	0
	I counts	1	1				0	0	0	0	0
	I percentage	6.67	14.29				0	0	0	0	0
	Total count	**15**	**7**				0	31	0	**2**	**7**
9	SS counts	7	5	4	1	4	1	No parasites detected			
	SS percentage	53.85	15.63	28.57	50	11.76	100				
	LS counts	6	20	10	1	30	0				
	LS percentage	46.15	62.5	71.43	50	88.24	0				
	I counts	0	7	0	0	0	0				
	I percentage	0	21.88	0	0	0	0				
	Total count	**13**	32	**14**	**2**	34	**1**				

Appendix 6: Short stumpy (SS), long slender (LS) and intermediate (I) relative percentages per individual and averaged are given below and based on data from appendix 5. Individuals B1-B5 were infected with Z310 parasites, strain B, E1-E5, with B17 parasites, strain E. Instances where no parasites were counted are indicated. *No parasites were detected in some of the individuals, and so the percentages for each stage were averaged from across only individuals in which parasites were counted. **Only one parasite was counted across all individuals, and so this was treated as no parasites were detected because two few were counted to determine relative stage abundances.

Day post infection		Zymodeme infected with									
		Z310					B17				
		B1	B2	B3	B4	B5	E1	E2	E3	E4	E5
3	SS percentage	6.45	8.93	7.69	3.45	2.56	No parasites detected				
	Average SS percentage	5.8									
	LS percentage	93.54	91.07	89.74	86.21	87.18					
	Average LS percentage	89.6									
	I percentage	0	0	2.56	10.34	10.26					
	Average I percentage	4.6									
4	SS percentage	5.56	6.67	0	0	7.5					
	Average SS percentage	4.0									
	LS percentage	94.44	93.33	100	100	90					
	Average LS percentage	95.5									
	I percentage	0	0	0	0	2.5					
	Average I percentage	0.5									
5	SS percentage	5.71	5.71	7.69	8.57	10	85.71	0	0	75	0
	Average SS percentage	7.5					80.4*				
	LS percentage	91.43	91.43	84.62	88.57	83.33	14.29	0	0	0	0
	Average LS percentage	87.9					7.1*				
	I percentage	2.86	2.86	7.69	2.86	6.67	0	0	0	25	0
	Average I percentage	4.6					12.5*				
6	SS percentage	13.63	14.28	7.69	9.09	0	63.64	100	87.5	90.24	91.89
	Average SS percentage	8.9					86.7				
	LS percentage	86.36	85.71	84.62	90.91	100	18.18	0	12.5	9.76	8.11
	Average LS percentage	89.6					9.7				
	I percentage	0	0	7.69	0	0	18.18	0	0	0	0
	Average I percentage	1.5					3.6				
7	SS percentage	No parasites detected** (One SS counted in one B1 mouse)					90.91	73.53	85.37	100	94.34
	Average SS percentage						88.8				
	LS percentage						9.09	14.71	7.32	0	5.66
	Average LS percentage						7.4				
	I percentage						0	11.76	7.32	0	0
	Average I percentage						3.8				
8	SS percentage	46.67	71.43	No parasites detected			0	96.77	0	100	100
	Average SS percentage	59.1*					98.9				
	LS percentage	46.67	14.29				0	3.23	0	0	0
	Average LS percentage	30.5*					1.1				
	I percentage	6.67	14.29				0	0	0	0	0
	Average I percentage	10.4*					0				
9	SS percentage	53.85	15.63	28.57	50	11.76	No parasites detected** (One SS counted in one E1 mouse)				
	Average SS percentage	32.0									
	LS percentage	46.15	62.5	71.43	50	88.24					
	Average LS percentage	63.6									
	I percentage	0	21.88	0	0	0					
	Average I percentage	4.4									

Appendix 7: Parasitaemia per individual and averaged values are given below for the infected animals discussed in Chapter 2, which were used to generate QPCR and metabolomic data and observe the relative abundances of SS:LS stages. Individuals B1-B5 were infected with Z310 parasites, strain B, E1-E5, with B17 parasites, strain E. Parasitaemia calculated from raw counts from thin films in appendix 5.

		Parasites per ml									
		Zymodeme infected with									
		Z310					B17				
Days post infection		B1	B2	B3	B4	B5	E1	E2	E3	E4	E5
3		6200	11200	7800	5800	7800	0	0	0	0	0
	Mean	7760					0				
4		10800	9000	10000	7600	8000	0	0	0	0	0
	Mean	9080					0				
5		7000	7000	7800	7000	6000	1400	0	0	800	0
	Mean	6960					440				
6		4400	1400	7800	2200	200	6600	200	1600	8200	7400
	Mean	3200					4800				
7		200	0	0	0	0	6600	6800	8200	10000	10600
	Mean	40					8440				
8		3000	1400	0	0	0	0	6200	0	400	1400
	Mean	880					1600				
9		2600	6400	2800	400	6800	200	0	0	0	0
	Mean	3800					40				

Appendix 8: SNP counts and percentages used to generate Figure 3.14A

	SNP impacts													
	High		Low		Moderate		Modifier		Missense		Nonsense		Silent	
	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
SNPs common to all B17 and Z310 strains	14	0.027	4,593	8.789	3,309	6.332	44,430	84.852	3,314	41.865	9	0.114	4,593	58.022
SNPs common to all Z366, B17 and Z310 strains	11	0.023	4,178	8.884	2,942	6.256	39,899	84.837	2,947	41.327	6	0.084	4,178	58.589
SNPs unique to Z310 strains	2	0.02	545	5.403	425	4.213	9,115	90.364	425	43.724	2	0.206	545	56.07
SNPs unique to B17 strains	-	-	223	7.981	166	5.941	2,405	86.077	166	42.674	-	-	223	57.326
SNPs unique to Z366 strain	15	0.132	608	5.332	635	5.569	10,144	88.967	636	50.556	14	1.113	608	48.331
SNPs unique to B strain	2	0.016	535	4.298	440	3.535	11,470	92.151	440	45.036	2	0.205	535	54.759
SNPs unique to E strain	5	0.102	286	5.817	323	6.569	4,303	87.513	324	52.769	4	0.651	286	46.58

Appendix 9: SNP counts and percentages used to generate Figure 3.14B

Type		SNPs common to all B17 and Z310 strains	SNPs common to all Z366, B17 and Z310 strains	SNPs unique to Z310 strains	SNPs unique to B17 strains	SNPs unique to Z366 strain	SNPs unique to B strain	SNPs unique to E strain
Downstream	Count	20,849	18,722	4,088	1,071	4,748	5,140	1,889
	%	39.898	39.809	40.527	38.332	41.642	41.295	38.418
Intergenic	Count	3,261	2,934	1,174	209	1,082	1,718	456
	%	6.24	6.239	11.639	7.48	9.49	13.803	9.274
Missense	Count	3,309	2,942	425	166	635	440	323
	%	6.332	6.256	4.213	5.941	5.569	3.535	6.569
Start lost	Count	-	-	-	-	1	-	1
	%	-	-	-	-	0.009	-	0.02
Stop gained	Count	9	6	2	-	14	2	4
	%	0.017	0.013	0.02	-	0.123	0.016	0.081
Stop lost	Count	5	5	-	-	-	-	-
	%	0.01	0.011	-	-	-	-	-
Stop retained	Count	9	9	-	-	4	-	1
	%	0.017	0.019	-	-	0.035	-	0.02
Synonymous	Count	4,584	4,169	545	223	604	535	285
	%	8.772	8.865	5.403	7.981	5.297	4.298	5.796
Upstream	Count	20,230	18,243	3,854	1,125	4,314	4,612	1,958
	%	38.713	38.79	38.198	40.265	37.835	37.053	39.821

Appendix 10: SNP counts and percentages used to generate Figure 3.14C

Region	Downstream		Exon		Intergenic		Upstream	
	Count	%	Count	%	Count	%	Count	%
SNPs common to all B17 and Z310 strains	20,849	40	7,908	15	3,261	6	20,230	39
SNPs common to all Z366, B17 and Z310 strains	18,722	40	7,123	15	2,934	6	18,243	39
SNPs unique to Z310 strains	4,088	40	972	10	1,174	12	3,853	38
SNPs unique to B17 strains	1,071	38	389	14	209	8	1,125	40
SNPs unique to Z366 strain	4,748	41	1,253	11	1,082	9	4516	39
SNPs unique to B strain	5,140	41	977	8	1,718	14	4,612	37
SNPs unique to E strain	1,889	39	612	12	456	9	1,958	40

Appendix 15: Intensity values averaged across each pathway and across individuals for each stage at days three and eight post infection. Intensities are relative to the pre-infection data for each strain, which is set to zero. This data was used to generate the heatmap Figure 4.2.

Pathway	Z310 Day 3	Z310 Day 8	B17 Day 3	B17 Day 8
Xenobiotics Metabolism	0.92	1.32	5.2	12.56
Tri-peptide Metabolism	1.06	0.47	22.35	19.02
Di-peptide Metabolism	1.53	0.59	53.67	31.12
Di-peptide Metabolism	2.37	1.95	17.7	23.97
Peptide Metabolism	36.85	12.91	81.59	26.57
Nucleotide Metabolism	16.73	0.71	55.97	57.86
Cofactors and Vitamins Metabolism	1.07	0.47	30.98	45.54
Sterol lipid Metabolism	4.15	2.66	55.31	175.43
Sphingolipid Metabolism	1.45	1.63	37.76	26.56
Polyketides Metabolism	2.04	0.75	1.51	2.72
Glycerophospholipid Metabolism	2.61	1.64	26.61	17.17
Glycerolipid Metabolism	1.07	0.82	2.32	1.85
Ganglioside Metabolism	0	5.59	1.25	7.06
Fatty Acyl Metabolism	21.37	17.57	38.07	52.45
Lipid Metabolism	1.65	1.27	10.65	9.98
Energy Metabolism	2.58	0.82	8.43	5.01
Carbohydrate Metabolism	1.01	0.67	74.73	25.09
Secondary Metabolites Biosynthesis	8.34	0.56	3.7	3
Polyketides and Nonribosomal Peptide Biosynthesis	0	0.22	0	0
Amino Acid Metabolism	12.89	1.49	162.76	97.74

Appendix 16: Mapping statistics used to generate Figure 4.14

Sample ID	Total reads	Reads mapped	Reads unmapped	Percentage mapped (%)
Z310				
B1_1_3	113,912,156	91,344,446	22,567,710	80
B1_3_2_7	67,409,372	33,659,766	33,749,606	50
B4_1_2	76,946,188	32,737,410	44,208,778	43
B5_1_1	8,1581,426	64,033,896	17,547,530	79
B1_4_2_6	130,625,146	93,556,915	37,068,231	72
B17				
E1_1_2_5	91,557,138	73,457,264	18,099,874	80
E4_1_4	77,343,632	53,644,089	23,699,543	69
E3_1_5	97,804,766	73,025,999	24,778,767	75
E1_1_7	78,493,866	61,148,015	17,345,851	78
E2_1_6	92,821,216	63,419,857	29,401,359	68

Appendix 16: Gene IDs of differentially expressed genes, which were more transcriptionally active in the Z310 strain

Gene ID	Fold Change	Gene ID	Fold Change	Gene ID	Fold Change
Tb927.3.2600	15.09	Tb927.5.4180	4.72	Tb927.9.740	1.23
Tb927.11.13000	13.66	Tb927.11.17180	4.70	Tb927.8.6710	1.21
Tb927.3.3200	13.55	Tb927.10.7920	4.58	Tb927.9.6170	1.18
Tb927.3.1900	12.78	Tb11.v5.0932	4.42	Tb927.11.14840	1.17
Tb927.9.5000	11.64	Tb927.7.440	4.41	Tb927.8.6660	1.17
Tb927.3.5400	11.54	Tb927.3.3400	4.39	Tb927.3.5370	1.17
Tb927.9.340	11.44	Tb927.4.5740	4.38	Tb927.8.6240	1.15
Tb927.9.970	11.33	Tb927.10.660	4.19	Tb927.3.2310	1.14
Tb927.9.960	10.96	Tb927.9.3000	4.12	Tb927.5.4010	1.14
Tb927.9.5800	10.87	Tb927.10.2770	4.10	Tb927.4.4690	1.14
Tb927.11.17050	10.86	Tb927.2.6280	4.05	Tb927.1.2670	1.14
Tb927.10.12050	10.79	Tb09.v4.0024	4.00	Tb927.8.6940	1.13
Tb927.9.9500	10.76	Tb08.27P2.400	3.67	Tb927.10.13100	1.13
Tb927.9.10100	10.47	Tb927.7.150	3.60	Tb927.4.2740	1.13
Tb927.8.260	10.46	Tb927.7.2020	3.46	Tb927.7.3420	1.11
Tb927.11.13040	10.40	Tb927.10.5700	3.41	Tb927.10.1040	1.10
Tb927.9.11700	10.24	Tb05.5K5.420	3.35	Tb927.2.5660	1.10
Tb927.9.1230	10.20	Tb11.v5.0790	3.19	Tb927.4.2450	1.10
Tb927.10.13080	10.17	Tb927.9.16460	3.18	Tb927.10.13110	1.07
Tb927.3.380	10.15	Tb927.5.1420	3.06	Tb927.5.1740	1.06
Tb927.11.20260	9.84	Tb927.6.360	2.90	Tb927.4.4700	1.06
Tb927.8.120	9.78	Tb927.11.17170	2.89	Tb927.10.14500	1.03
Tb927.5.410	9.72	Tb927.6.5530	2.75	Tb927.8.710	1.02
Tb927.11.17080	9.37	Tb05.5K5.240	2.71	Tb927.11.9860	0.99
Tb927.11.20400	9.33	Tb11.v5.0487	2.60	Tb927.10.10140	0.99
Tb927.9.9400	9.10	Tb11.v5.0675	2.58	Tb927.11.7470	0.99
Tb927.8.190	9.08	Tb927.1.1820	2.50	Tb927.10.12820	0.98
Tb927.1.3000	9.06	Tb927.11.950	2.43	Tb927.11.10940	0.98
Tb927.3.1000	9.06	Tb927.8.7890	2.42	Tb927.8.6430	0.97
Tb927.11.2070	8.97	Tb927.5.4470	2.41	Tb927.8.2780	0.97
Tb927.11.20730	8.73	Tb927.11.780	2.30	Tb927.1.3830	0.96
Tb927.10.500	8.72	Tb927.6.3390	2.27	Tb927.11.11210	0.96
Tb927.8.340	8.67	Tb927.11.12020	2.21	Tb927.9.6130	0.96
Tb927.3.250	8.58	Tb927.10.10590	2.14	Tb927.11.16550	0.96
Tb927.9.1150	8.58	Tb927.9.16640	2.13	Tb927.11.14880	0.95
Tb927.10.13010	8.27	Tb11.v5.0710	2.07	Tb927.5.2940	0.94

Tb927.11.20570	8.26	Tb927.6.350	2.05	Tb927.11.7350	0.93
Tb927.9.160	8.24	Tb927.8.3500	1.93	Tb927.3.5010	0.93
Tb927.11.19050	8.21	Tb927.8.5440	1.89	Tb927.11.10810	0.93
Tb927.5.4630	7.83	Tb927.6.440	1.87	Tb927.10.830	0.92
Tb927.5.4380	7.77	Tb927.7.1990	1.87	Tb927.11.3250	0.92
Tb927.11.18220	6.59	Tb927.10.10920	1.86	Tb927.11.1430	0.92
Tb927.6.140	6.50	Tb05.5K5.210	1.83	Tb927.11.14300	0.92
Tb927.1.4000	6.42	Tb927.3.5830	1.80	Tb927.3.3040	0.92
Tb927.10.380	6.32	Tb927.9.5130	1.77	Tb927.3.930	0.91
Tb10.v4.0172	6.31	Tb927.10.14160	1.70	Tb927.6.400	0.91
Tb927.11.20090	6.12	Tb927.8.6760	1.56	Tb927.8.3850	0.91
Tb927.10.750	6.08	Tb927.9.4730	1.52	Tb927.10.5400	0.90
Tb927.11.19430	5.65	Tb927.6.410	1.47	Tb927.3.3270	0.90
Tb927.9.12700	5.63	Tb927.1.2680	1.46	Tb927.4.4040	0.89
Tb927.6.5550	5.58	Tb927.10.14140	1.45	Tb927.10.8930	0.88
Tb927.4.250	5.47	Tb927.8.7970	1.40	Tb927.1.4890	0.87
Tb927.11.17120	5.43	Tb927.8.5460	1.38	Tb927.4.870	0.86
Tb927.9.16920	5.24	Tb927.7.6840	1.37	Tb927.10.9570	0.86
Tb927.10.16530	5.23	Tb927.1.3470	1.32	Tb927.3.2960	0.85
Tb927.9.17850	5.17	Tb927.8.5010	1.29	Tb927.7.6690	0.85
Tb11.v5.1046	4.95	Tb927.2.2770	1.29	Tb927.11.6550	0.84
Tb927.5.280	4.90	Tb927.11.7710	1.29	Tb927.11.4180	0.84
Tb927.3.550	4.86	Tb927.10.2190	1.28	Tb927.10.7930	0.80
Tb927.9.17390	4.83	Tb927.11.14030	1.27	Tb927.7.1310	0.80
Tb09.v4.0200	4.76	Tb927.11.3630	1.26	Tb927.3.5020	0.75
Tb09.v4.0031	4.75	Tb927.4.4580	1.25	Tb927.4.3740	0.72
Tb10.v4.0088	4.73	Tb927.1.4830	1.23		

Appendix 17: Gene IDs of differentially expressed genes, which were more transcriptionally active in the B17 strain

Gene ID	Fold Change	Gene ID	Fold Change	Gene ID	Fold Change
Tb927.9.1010	15.21	Tb927.6.5490	4.45	Tb927.8.730	1.65
Tb927.9.490	12.68	Tb927.3.520	4.29	Tb927.11.1300	1.65
Tb927.11.20150	12.64	Tb927.4.5580	4.25	Tb927.1.1000	1.62
Tb927.10.5200	12.04	Tb927.5.230	4.24	Tb927.5.2160	1.60
Tb927.7.140	12.01	Tb10.v4.0227	4.05	Tb927.7.4270	1.54
Tb927.11.20700	11.86	Tb927.9.16220	3.94	Tb927.11.7060	1.52
Tb927.11.19060	11.86	Tb927.10.15070	3.80	Tb927.7.5950	1.50
Tb927.8.440	11.84	Tb10.v4.0174	3.77	Tb927.10.2360	1.50
Tb927.9.430	11.80	Tb927.9.7340	3.76	Tb927.11.5860	1.48
Tb927.7.110	11.62	Tb10.v4.0214	3.72	Tb927.3.4180	1.45
Tb927.9.300	11.61	Tb09.v4.0082	3.64	Tb927.8.1390	1.44
Tb927.9.280	10.76	Tb927.7.170	3.40	Tb927.7.2640	1.43
Tb927.9.1240	10.63	Tb927.9.730	3.32	Tb927.4.3940	1.42
Tb927.11.17070	10.54	Tb927.6.5280	3.20	Tb927.3.3750	1.39
Tb927.8.490	10.45	Tb927.11.2690	3.08	Tb927.7.6330	1.37
Tb927.11.19020	10.28	Tb927.9.7470	3.00	Tb927.7.6340	1.35
Tb927.9.830	10.22	Tb927.8.6750	2.96	Tb927.8.6930	1.34
Tb927.8.100	10.20	Tb09.v4.0151	2.93	Tb927.7.7430	1.30
Tb927.1.10	10.18	Tb927.7.5930	2.92	Tb927.8.760	1.29
Tb927.3.390	10.10	Tb927.10.3210	2.91	Tb927.11.12040	1.25
Tb927.5.210	10.00	Tb927.9.7370	2.89	Tb927.6.2740	1.24
Tb927.11.20160	9.59	Tb927.7.5940	2.89	Tb927.5.3950	1.22
Tb927.11.18050	9.58	Tb927.11.10530	2.75	Tb927.10.7700	1.21
Tb927.9.440	9.41	Tb927.9.7450	2.75	Tb927.10.14860	1.19
Tb927.9.510	9.39	Tb927.10.16430	2.73	Tb927.11.9820	1.17
Tb927.8.430	9.32	Tb927.11.11680	2.72	Tb927.10.12840	1.17
Tb927.11.20550	9.06	Tb927.11.18980	2.66	Tb927.2.3370	1.16
Tb927.11.20210	8.85	Tb927.9.7460	2.61	Tb927.6.2790	1.16
Tb927.3.330	8.84	Tb927.9.5910	2.55	Tb927.2.4370	1.14
Tb927.2.6220	8.66	Tb927.7.2660	2.52	Tb927.3.1380	1.12
Tb927.11.15060.2	8.66	Tb927.11.12710	2.50	Tb927.10.7570	1.12
Tb927.11.19710	8.57	Tb927.9.680	2.48	Tb927.9.10400	1.10
Tb927.3.180	8.39	Tb927.3.2230	2.39	Tb927.10.12700	1.10
Tb927.3.350	8.15	Tb927.3.560	2.39	Tb927.8.7530	1.09
Tb927.11.20110	8.02	Tb927.10.7410	2.33	Tb927.10.2370	1.08
Tb927.11.20230	7.77	Tb927.4.2240	2.33	Tb927.7.4390	1.08
Tb927.10.5220	7.31	Tb927.11.2010	2.26	Tb927.8.5450	1.07
Tb927.9.330	7.29	Tb927.6.1520	2.20	Tb927.6.4750	1.07
Tb927.9.1050	7.09	Tb927.3.5760	2.18	Tb927.8.8070	1.06

Tb927.10.5250	7.09	Tb927.3.1790	2.17	Tb927.7.180	1.05
Tb927.3.490	7.09	Tb927.11.9980	2.16	Tb927.11.4700	1.04
Tb927.3.190	7.00	Tb927.8.1130	2.16	Tb927.7.3520	1.04
Tb927.9.16040	6.74	Tb927.9.15310	2.14	Tb927.9.8580	1.02
Tb927.4.5530	6.43	Tb927.11.9030	2.14	Tb927.8.8320	1.02
Tb927.11.19730	6.43	Tb927.10.12260	2.10	Tb927.8.6060	1.02
Tb10.v4.0139	6.41	Tb927.11.1450	2.09	Tb927.8.7150	1.00
Tb927.3.480	6.28	Tb927.9.5920	2.08	Tb927.9.6310	0.99
Tb927.11.1100	6.20	Tb927.9.5900	2.07	Tb927.8.3770	0.99
Tb927.11.19720	6.00	Tb927.9.1520	2.07	Tb927.10.11220	0.99
Tb927.4.5790	5.89	Tb927.9.720	2.06	Tb927.10.12240	0.98
Tb927.9.770	5.86	Tb927.7.3970	2.04	Tb927.9.7830	0.98
Tb927.1.5340	5.70	Tb927.9.15290	2.01	Tb927.3.3330	0.98
Tb927.3.270	5.47	Tb927.5.2260	1.98	Tb927.8.6450	0.97
Tb927.11.20220	5.43	Tb927.9.5950	1.91	Tb927.11.7900	0.95
Tb927.6.110	5.33	Tb927.3.5820	1.87	Tb927.7.6600	0.95
Tb927.9.16880	5.26	Tb927.10.2350	1.81	Tb927.7.5550	0.95
Tb927.11.20320	5.11	Tb927.7.7500	1.80	Tb927.11.6280	0.94
Tb927.4.5400	5.07	Tb927.9.15300	1.78	Tb927.8.6770	0.93
Tb927.5.5400	4.91	Tb927.7.210	1.77	Tb927.8.1780	0.92
Tb927.9.12500	4.70	Tb927.9.5940	1.75	Tb927.10.1030	0.90
Tb11.v5.0518	4.65	Tb927.11.16730	1.74	Tb927.10.3770	0.89
Tb927.11.20330	4.57	Tb927.10.14630	1.69	Tb927.4.2630	0.84
Tb927.4.170	4.50	Tb11.v5.0752	1.69	Tb927.9.7670	0.82