Genomic approaches to understanding host resistance and parasite virulence in *Trypanosoma* parasites

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

by

Ian Barry Goodhead

May 2011

UNIVERSITY OF LIVERPOOL

Acknowledgements

To Neil and Harry, thanks for the supervision and for the opportunity to undertake such a project. Thanks also go to the Centre for Genomic Research, to Steve Paterson and Mark Caddick, and to others in the Institute for helpful discussions and help throughout this PhD. Particular thanks go to J. Wendi Bailey, without whose astonishing attention to detail this would never have been possible. Also to Tanya Beament, Jason Dowling, Mark Taylor and Michael Chance (LSTM) for the preliminary work on the B17 and Z310 strain phenotyping, which was the forerunner to this work. Thanks also to Pam Pask, Stephen Johnson and Jo Sanders for all of their help at the Biomedical Services Unit. Thanks also to our collaborators Annette MacLeod, Liam Morrison and to the rest of her team at the University of Glasgow for all of their help throughout the project, particularly with microsatellite genotyping and PCR, culturing (and for the curry).

On a personal note: Si Reeves, thanks for introducing me to the joys of listening to vinyl and, arguably more importantly, for giving me the opportunity to move to Cambridge and take up my first job at the Sanger; Matt Ife, thanks to you I took up whisky. Both of you introduced me to finger-style guitar. Endlessly. Yeah, cheers. Thanks, too, go to my co-habitants since: to Katherine, for managing to live with me for the last three years and to Cheryl, for hopefully managing to for a few more.

To my bosses Carol and Karen who, arguably, started all this with the phrase: "*How would you like a trip to America?*" A heartfelt thanks to you both, and all of my friends and colleagues at Hinxton and Gt. Chesterford, particularly to those that gave me that final push to finally get out of there and do this PhD.

Thanks to the various coffee houses (particularly Bold St. coffee!) and to the many bars that have fulfilled my football addiction and to those friends that have accompanied me along the way.

I reserve last, and pride of, place for my family, and particularly, of course, to my Mum and Dad. Thank you for your never-ending support from wherever in the world you happen to be at the time.

Ian

Table of Contents

Index of Tables and Figures

List of Major Abbreviations

Abbreviation	Description
(K/M) bp	Base Pairs
AA	Amino Acid
aCGH	Array Comparative Genomic Hybridisation
AIL	Advanced Intercross Line
ALAT	Alanine aminotransferase
AMP	Adenosine Monophosphate
ApoA	Apolipoprotein A
ASAT	Aspartate aminotransferase
B(number) eg. B17	Busoga (zymodeme strain group)
BAPS	Bayesian Analysis of Population Structure (Software)
BBB	Blood/Brain Barrier
CAPS	Cleaved Amplified Polymorphic Sequence
Cd244	Cluster of Differentiation 244
CHIP	Chromatin Immunoprecipitation
cM	CentiMorgans
CNS	Central Nervous System
CNV	Copy Number Variation / Variant
CSF	Cerebro-Spinal Fluid
EDTA	Ethylenediaminetetraacetic acid
FACS	Fluorescence Activated Cell Sorting
GO	Gene Ontology
GOT	Glutamate oxaloacetate transaminase
GPI	Glycosylphosphatidylinositol
gRNA	Guide RNA
HAT	Human African Trypanosomiasis
HPR	Haptoglobin Receptor
IACUC	Institutional Animal Care and Use Committee
ICD	Isocitrate Dehydrogenase
IFNg	Interferon Gamma
IgM	Immunoglobin M
ILRI	International Livestock Research Institute (Kenya)
kDNA	Kinetoplast DNA
LPS	Lipopolysaccharide
MCMC	Markov Chain Monte-Carlo
MDH	Malate dehydrogenase
MHC	Major Histocompatibility Complex
MLEE	Multi Locus Enzyme Electrophoresis
MLMT	Multi Locus Microsatellite Typing
MLST	Multi Locus Sequence Typing
MY	Million Years
NHD	Nucleoside hydrolase
NHD	Nucleoside Hydrolase (Deoxyinosine)

Abbreviation	Description
NHI	Nucleoside Hydrolase (Inosine)
NJ	Neighbour Joining
NO	Nitric Oxide
nsSNP	non synonymous SNP
PCR	Polymerase Chain Reaction
PGM	Phosphoglucomutase
Pram1	PML-RARA-regulated adapter molecule 1
QTG	Quantitative Trait Gene
QTL	Quantitative Trait Locus / Loci
rRNA	Ribosomal RNA
SIT	Sterile Insect Technique
SNP	Single Nucleotide Polymorphism(s)
snRNA	small nuclear RNA
SOD	Super Oxide Dismutase
SRA	Serum Resistance Associated (gene)
SSU	Small Subunit
TDH	Threonine dehydrogenase
Tir (1/2/3a-c)	Trypanosoma Infection Response (QTL)
Tir1AA	Mice Congenic for A/J (AA) alleles at Tir1
Tir1CC	Mice Congenic for C57BL/6 (CC) alleles at Tir1
TLF(1/2)	Trypanosome Lytic Factor
Tlr4	Toll-like Receptor 4
TNFa	Tumour Necrosis Factor a
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
VSG	Variable Surface Glycoprotein
WHO	World Health Organisation
Z(number) eg. Z310	Zambesi (zymodeme strain group)

Declaration

This thesis has been prepared with the intention of submitting them for publication, and as such the chapters are independently introduced and discussed. Chapters one and six are provided as a general introduction and discussion, respectively. The thesis is my own work and contains nothing that is the outcome of work done in collaboration with others except as specified in the text. Any additional analyses important to the overall discussion of the data that were in association with supervisors or collaborators are available as supplementary material within the appendix and are highlighted in the text.

Abstract

Roughly one-third of cattle in sub-Saharan Africa are at risk of contracting "Nagana" – a disease caused by *Trypanosoma* parasites similar to those that cause human "Sleeping Sickness". Laboratory mice can also be infected by trypanosomes, and different mouse breeds show varying levels of susceptibility to infection, similar to what is seen between breeds of cattle. We have applied next-generation technologies to identify shared polymorphisms between susceptible mice, and annotated these for potential function alongside publicly available SNP data sets. By so doing, short lists of genes at the QTL have been created to aid functional testing in cattle. This includes two promising 'candidate genes': *Pram1* and *Cd244*, which can now be tested to confirm their effect on response to trypanosome infection.

The human-infective parasite *Trypanosoma brucei rhodesiense* generally causes an acute form of "sleeping sickness" across Eastern Africa, compared to the more chronic *T. b. gambiense* infections found in Western Africa. The 1988–1993 Ugandan *T. b. rhodesiense* outbreak constituted infections by parasites with differences in their clinical manifestation. Two such subtypes, termed *Busoga* 17 (B17) and *Zambesi* 310 (Z310), caused more acute, and more chronic infections, respectively. In order to investigate whether the major QTL that regulates survival in *T. congolense* infections (*Tir1*) does so in a similar manner in *T. b. rhodesiense*, mice congenic for the C57BL/6 allele (Tir1CC) at *Tir1* were infected with Z310 and B17 zymodeme *T. b. rhodesiense* parasites. Whilst *Tir1* was not found to have a significant effect on survival, all mice had a significantly shorter mean survival time when infected with B17 (~10.7 days) than those infected with Z310 (~15.6 days), in line with previous observations of human infections.

In order to identify genetic loci that might underlie differences in virulence between *T. b. rhodesiense* zymodemes, cluster analysis was performed on the microsatellite genotypes of 31 *T. b. rhodesiense* isolates that represented nine different zymodemes. Despite STRUCTURE identifying three population clusters, the Z310 and B17 parasite populations could not be distinguished, suggesting that either multiple genes control virulence, that there is gene flow between similar parasite populations, or that the microsatellite genotyping is insufficient to distinguish between different parasite populations.

Finally, we present the first whole-genome sequences of *T. b. rhodesiense* field isolates, one each of Z310 and B17. Genomic analysis of east African *T. b. rhodesiense* and west African *T. b. gambiense* has suggested that recombination may be occurring between them. SNP genotyping of 32 *T. b. rhodesiense* isolates showed that differences in clinical phenotypes were associated with differences in alleles on chromosome 8. The genome sequence suggests that chromosome 8 is heterozygous for alleles of west African origin in the more virulent strain, suggesting that recombination may be associated with parasite virulence. This suggests that the human subspecies of *T. brucei* are not genetically distinct, which has major implications for the control of the parasite, the spread of drug resistance and understanding the variation in virulence and the emergence of human infectivity. Further genetic analysis of *T. b. brucei* populations from Western, central and Eastern Africa may be necessary to ascertain whether recombination is occurring directly between human-infective subspecies, or in the underlying animal-infective population.

Chapter One: Introduction

African Trypanosomiasis is a neglected disease that has a wide-ranging impact across sub-Saharan Africa [1, 2]. The causative agents, the trypanosomes, are parasites from the Kinetoplastida order of protozoa, so-called due to the presence of a distinctive structure, the kinetoplast, situated within the single mitochondrion at the base of the flagellum [3]. The Kinetoplastida order also includes the closely-related *Leishmania* genus, that causes a range of diseases in tropical regions [4].

The molecular phylogeny of trypanosomes

Mammalian trypanosomes are split into two groups based on the location of their development within their insect vectors: the Stercoraria develop in the midgut or the hindgut (e.g. *T. cruzi* develop in both the midgut and hindgut of the triatomine bug); the Salivaria develop in the midgut and/or salivary glands (e.g. *T. brucei* develops within the midgut and salivary gland of the tsetse fly).

A study of the sequences of 34 small subunit (SSU) rRNA sequences from different isolates of *Trypanosoma* suggested an ancient separation into five distinct clades that correlated with host/parasite co-evolution (Figure 1.1; [5]). Salivarian trypanosomes split from the other trypanosomes approximately 500MY ago, possibly linked to a common ancestor developing the antigen-switching strategy to survive for long periods within the mammalian bloodstream. Fish and amphibian trypanosomes appear to have split from their bird-infecting counterparts some 130MY ago, although there is evidence for some host-switching having taken place [5].

Trypanosomes exist in both intra- and extra-cellular forms in a wide range of species, including man [6]. Fortunately, few species are human-infective: Firstly, *Trypanosoma cruzi* causes "Chagas Disease" across South America [7]; Secondly, two subspecies of *Trypanosoma brucei* cause Human African Trypanosomiasis (HAT), commonly termed "Sleeping Sickness" [8].

Three species of African trypanosome are medically and economically important: the two human infective sub-species of *T. brucei*: *T. b. gambiense* and *T. b. rhodesiense*, and the major animal infective species: *T. vivax* and *T. congolense* [6]. All medically important African trypanosomes are transmitted via the bites of different subspecies of infected Tsetse flies (genus *Glossina*). Table 1.1 shows a list of the principle African trypanosomes that infect man, other primates and wildlife. Indicated are the species of the parasite, mode of transmission (and associated vector) and host species.

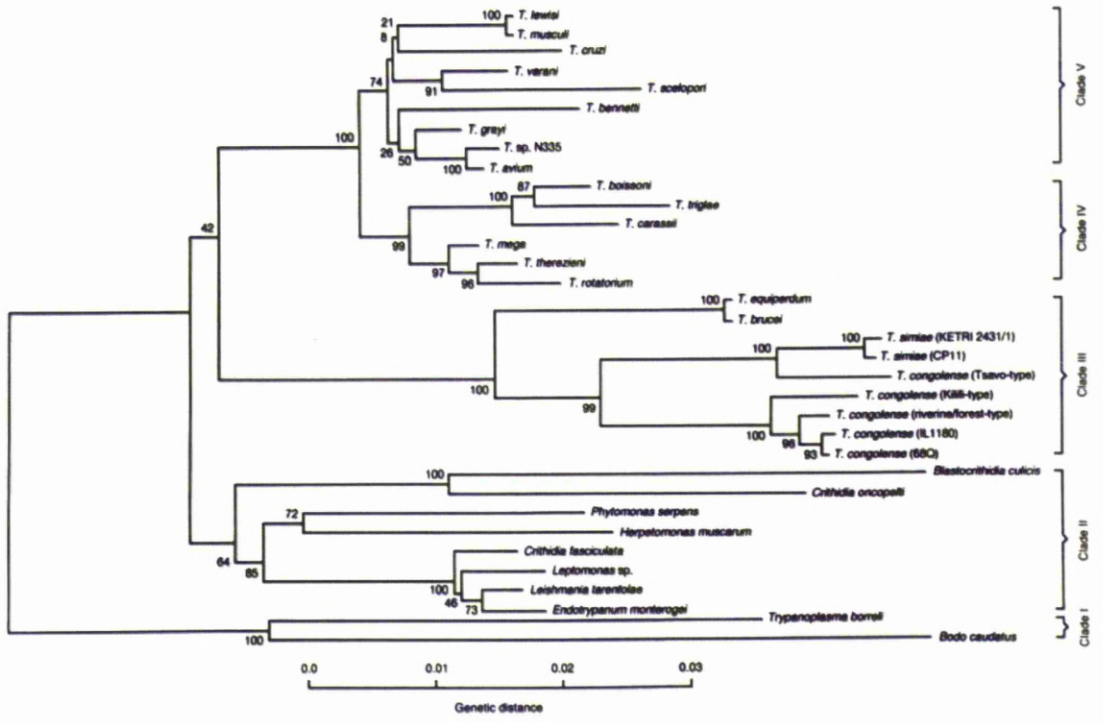


Figure 1.1: Molecular phylogeny of *Trypanosoma* based on SSU rRNA sequence data. Phylogenetic tree represents a neighbour-joining tree of SSU rRNA sequences for 34 *Trypanosoma* species and sub-species. Five clades are represented with high bootstrap values. Reproduced with permission from Haag *et al* (1998) [5].

Table 1.1: (+) Trypanosome species that infect man, other primates and domesticated and other wildlife. (-) Animal infective trypanosomes are unable to infect man and no known trypanosomes infect baboons. **(U)** Evidence for animal infections of *T. b. gambiense* is increasing, but the importance of an animal reservoir is unknown [9]. (§) *T. vivax, T. evansi* and *T. equiperdum* have spread beyond sub-Saharan Africa: *T. vivax* is found in countries across South America [10]; *T. evansi* is found in Asian countries such as China and Indonesia [11]. Vector information from Ford (1970) and Hoare (1972) [3, 12]. **(L)** *T. equiperdum* is principally an equine disease, although laboratory-adapted strains can infect mice and domestic animals [11]. (#) *T. simiae* causes an important disease in domestic pigs, but is otherwise difficult to distinguish from *T. congolense*, which does not. The species of infecting tsetse may have an impact upon the pathogenesis of the disease [13].

Species	**Vector / Mode of transmission**	**Man**	**Baboon**	**Other primates**	**Domestic animals**	**African Wildlife**
T. vivax (§)	Various: Includes *G. palpalis*; *G. mortisans*, *G. tachinoides*	-	-	+	+	+
T. congolense	Various: Includes *G. mortisans*; *G. palpalis*; *G. longipalis*; *G. pallidipes*; *G. austeni*	-	-	+	+	+
T. simiae	*G. mortisans*; *G. tachinoides* (#)	-	-	-	+	+
T. b. brucei	Most *Glossina spp.*	-	-	+	+	+
T. b. rhodesiense	*G. morsitans*; *G. pallidipes*; *G. fuscipes*	+	-	+	+	+
T. b. gambiense	*G. palpalis*; *G. tachinoides*; *G. fuscipes*	+	-	+	- (U)	- (U)
T. equiperdum	Sexually transmitted	-	-	-	+ (L)	+
T. evansi	Various bloodsucking insects, including: *Tabanus*; *Stomoxys*; *Atylotus* and *Lyperosia spp.*	-	-	-	+	+

Animal trypanosomiasis: "Nagana"

African animal trypanosomiasis affects over ten million square kilometres of Africa and some thirty percent of Africa's 160 million cattle are at risk of infection. Some species of African trypanosome cause potentially fatal disease, such as *T. simiae* in pigs [14], *T. evansi* in camels [15], and *T. b. rhodesiense* in cattle [16]. *T. vivax* and *T. congolense* also cause significant disease in livestock, whereas *T. brucei brucei*, causes mild symptoms, or is asymptomatic [17, 18].

Keeping livestock, and cattle in particular, provides many benefits for rural Africans. Cattle are not only used for milk and food, but also aid crop production as they provide power for ploughing, manure and transport [Reviewed [19]]. The symptoms associated with animal trypanosomiasis to susceptible livestock include weight-loss, anaemia and cachexia. As such, as the disease renders the animals unsuitable for these uses, the losses associated with unproductive livestock and decreases in crop production are estimated at over $1 billion per annum [20, 21]. It has been estimated that a six percent reduction in disease in livestock equates to the ability to feed an additional 250 million people [14].

Human African trypanosomiasis: "sleeping sickness"

The two human infective *T. brucei* subspecies, and a third animal infective form, are morphologically identical. They follow similar life cycles in the insect and mammalian hosts but cause three distinct diseases. *Trypanosoma brucei brucei*, as described above, can be found across Sub-Saharan Africa and is not human infective. Secondly, *Trypanosoma brucei gambiense* can be found in Western and Central Africa and represents approximately 90% of reported cases of sleeping sickness in humans. This parasite causes a chronic infection and can take months or even years to develop into an advanced stage [6]. Molecular studies have revealed two sub-types of *T. b. gambiense*: Type 1 is genetically distinct from other *T. brucei* subspecies whereas Type 2 *T. b. gambiense* is more similar to *T. b. brucei* [22]. Thirdly, *Trypanosoma brucei rhodesiense* represents around 10% of reported cases in humans and causes an often acute disease that rapidly develops into an advanced stage [23]. It is found in Eastern and Southern

Africa, where incidences are increasing, particularly in endemic countries such as Uganda [24].

Human African trypanosomiasis (HAT) affects more than 36 countries across Sub-Saharan Africa wherever tsetse flies are present. Figure 1.2 shows a map of the incidences of HAT in 2009 [2]. Thought to have been almost eradicated in the 1960s, the World Health Organization now estimates that there are around half a million cases of disease each year, killing around 66,000 people and disabling 100,000, thanks in part to a lack of new treatments and the difficulties associated with treatment in war-torn areas [21]. Epidemics are now re-emerging in countries such as the Democratic Republic of Congo [25]; Sudan [26]; Uganda and Tanzania [2, 27].

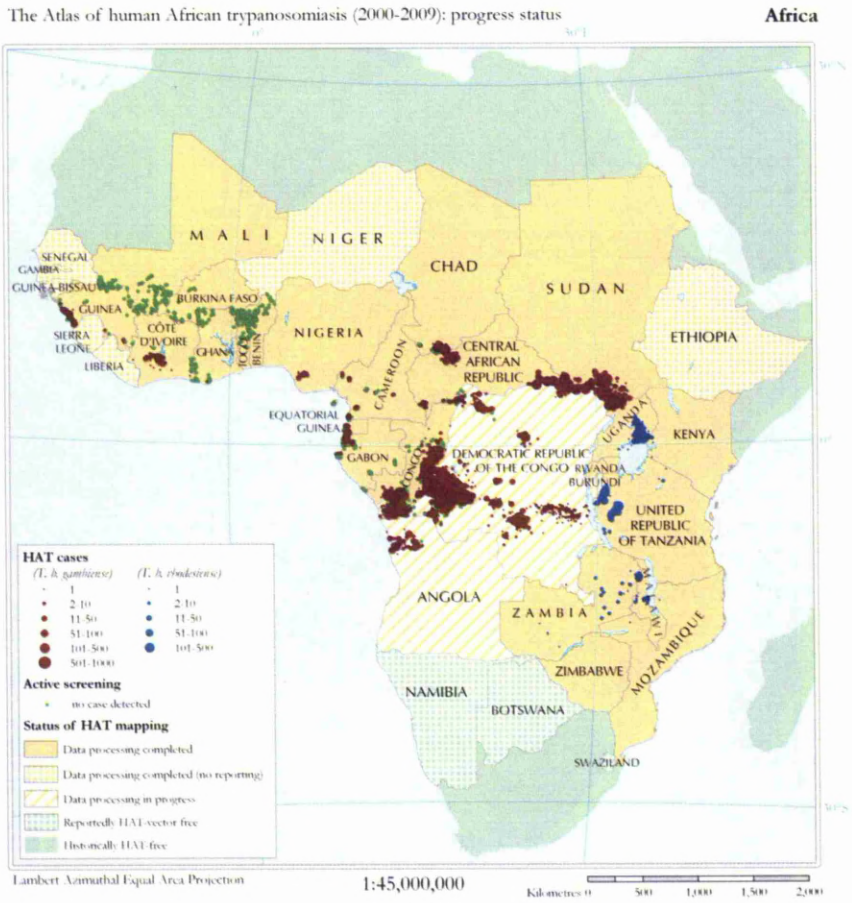


Figure 1.2: A map showing incidences of HAT (or "Sleeping Sickness") between 2001-2009. Red circles represent the "Western" *T. b. gambiense*–mediated and blue circles the "Eastern" *T. b. rhodesiense*–mediated sleeping sickness. Map reproduced from Simarro (2010); © World Health Organization/Food and Agriculture Organization of the United Nations [2].

The life cycle of the African trypanosome

A schematic depicting the trypanosome life cycle is shown in Figure 1.3. After inoculation by an infected tsetse fly (1), disease onset (also termed "early stage" or "stage 1 disease") is represented by long-slender trypomastigotes developing and multiplying within the blood, lymph and subcutaneous tissues (2-3). Some of these diverge to a morphologically-similar form, the short-stumpy trypomastigote (4), which is insect-infective. During this stage, further tsetse flies can become infected if the host is bitten (5). The short-stumpy forms thereafter transform into procyclic trypomastigotes within the midgut of the tsetse, wherein they undergo further replication (6). Subsequent migration to the insect salivary gland is accompanied by transformation to the epimastigote form (7) and mammalian-infective metacyclic trypomastigotes (8).

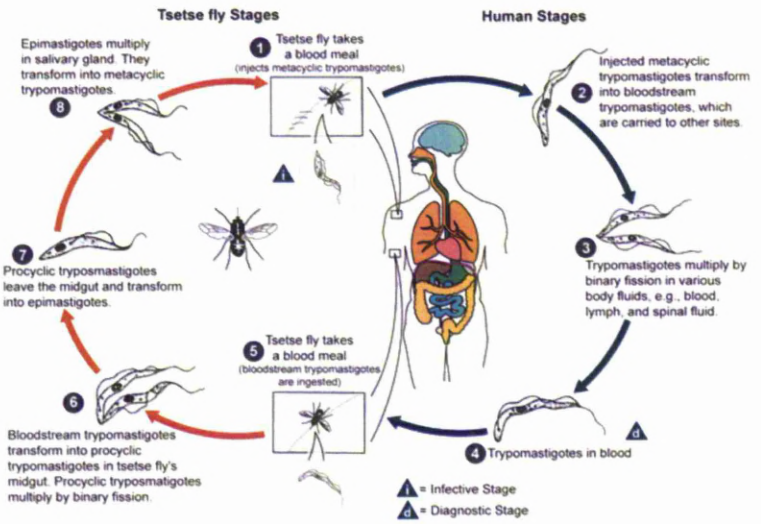


Figure 1.3: Life cycle of *Trypanosoma brucei ssp.* showing the insect and human stages. Image reproduced from the US Centre for Disease Control Department of Parasitic Diseases.

Disease and symptoms

If left untreated, the parasites eventually cross the blood-brain barrier to infect the central nervous system. In *T. b. rhodesiense* infections, this can occur as quickly as between three weeks to two months after initial infection, or can take up to several years in the case of *T. b. gambiense* [28]. It is at this stage (often termed "late stage" or "Stage 2" disease) where neurological symptoms can develop, including lethargy, lack of

coordination and the inverted sleep pattern that gives the disease its name. Patients tend to die in coma from heart failure, malnutrition or from secondary diseases [29, 30].

As trypanosomes have evolved alongside several host species, some hosts exist in which no quantifiable symptoms manifest [31]. This presumably increases the likelihood of transmission as the parasites survive for longer inside the bloodstream, increasing the chance of being taken up by the vector at its next blood-meal. This is of particular significance to the zoonotic *T. b. rhodesiense*, which also infects wild animals (e.g. antelope and boar), and livestock (e.g. domesticated pigs and cattle) in which it shows few symptoms [32]. As such, this represents a problem in East Africa where humans and animals live in close proximity. This increases the likelihood of animal to human transmission and underlies the problem in controlling the disease as all vectors and potential hosts need to be monitored and/or treated [33, 34].

Immune response and antigenic variation

As trypanosomes are extra-cellular parasites, they are constantly under attack from the various components of the host immune system. Studies using the mouse model have shown that during early infection, trypanosomes initiate the innate immune response through the triggering of pattern recognition receptors (such as the toll-like receptors) that, in turn, signal for the production of inflammatory cytokines such as tumour necrosis factor alpha (TNFa), interferon gamma (IFNg) and nitric oxide (NO) [35] [30]. Triggering the IFNg [36] and TNFa [37] cytokine responses, confers a growth advantage to the parasite, despite TNFa having a trypanolytic effect [38].

The parasite's primary form of defence relies on the expression of a variant surface glycoprotein (VSG), which creates a coat that acts as a physical barrier to attack by the adaptive immune system. Whilst the mammalian hosts produce specific antibodies (IgG, IgM) against this coat, the parasite has two associated evasion strategies: Firstly, the parasites are able to internalise and recycle the VSG on the surface, removing attached antibodies as it does so [39]; Secondly, they are able to shed the original coat and change the expressed VSG gene [40]. Taking place once in every approximately 100 generations, this process occurs through recombination with a catalogue of several hundred 'silent' gene components located around the genome [41, 42]. The parasites

that carry an antigenically distinct coat avoid targeting whilst the older generations are removed, which gives rise to 'waves' of parasitemia as shown in Figure 1.4.

Importantly, animal-infective trypanosomes are unable to infect humans due to the presence of a 'trypanosome lytic factor' (TLF) in human serum. TLF is present in the host plasma and is a minor component of the high-density lipoprotein, made up of three proteins: apolipoprotein A (APOA), apolipoprotein L-I (APOL-I), and haptoglobin-related protein (HPR). TLF has been further characterised to consist of two forms: TLF-1, made up of HDL-bound APOA-I, APOA-II and HPR; and TLF-2, which is made up of an APOA-I/HPR/IgM complex [43]. APOL-I (and the associated TLF-1) is arguably the more important molecule for immunity. It is endocytosed by the trypanosome whereupon it disrupts the lysosomal membrane, releasing the contents into the cytoplasm and triggering auto-lysis [44]. The role of TLF-2 *in vivo* remains controversial: Whilst Raper *et al* have suggested that TLF2-based haptoglobin may be the main anti-trypanocidal molecule [45], studies have observed a reduction in lytic activity with increasing levels of TLF-2 [46] [Reviewed [47]]. Nevertheless, both of the human-infective forms of the parasite have evolved independent mechanisms to evade these processes, and thus infect man: Details of the process in *T. b. gambiense* are becoming clear, being due to the reduced expression of the haptoglobin receptor and a corresponding reduced uptake of TLF-1 [48]. *T. b. rhodesiense* has evolved the "Serum Resistance Associated" (SRA) gene, which appears to have been derived from a VSG gene [49, 50], and works by binding APOL-I [51]. Interestingly, baboons are resistant to human-infective parasites (Table 1.1) due to the presence of a specific two amino-acid motif that renders SRA unable to bind to their version of APOL-I [52].

Treatment and prevention

Whilst a number of drugs are available to treat the disease, the type of drug used and their efficacy are largely dependant on the form of the disease (i.e. either *T. b. gambiense* or *T. b. rhodesiense*) and whether the parasites are present in the bloodstream or in the CNS. The five drugs listed in Table 1.2 are the most wide-spread, each with a particular target in terms of strain and stage of infection.

Table 1.2: Drugs available for HAT treatment and their use in the different forms of the disease (Adapted from [53]). *T. b. g.* = *T. b. gambiense*; *T. b. r.* = *T. b. rhodesiense.*

Drug	First Marketed	Disease targeted	Stage of Disease	Notes
Pentamidine	1937	*T. b. gambiense*	Stage I	Treatment failures
Suramin	1922	*T. b. rhodesiense*	Stage I Stage II	Can be used for *T.b.g.* but not recommended
Melarsoprol	1949	*T. b. gambiense* *T. b. rhodesiense*	Stage II	Treatment failures 2-12% mortality
Eflornithine	1981	*T. b. gambiense*	Stage II	Large dose required. Difficult to administer
Nifurtimox	1960	*T. b. gambiense* (*T.cruzi*)	Stage II	Treats Chagas disease; Not approved for HAT. Effects on *T.b.r.* unknown

If diagnosed early, Suramin and Pentamidine are the drugs of choice, as they have relatively few side effects (there is a possibility of anaphylactic shock with Suramin). Late stage treatment is more complicated due to the drug having to cross the blood/brain barrier. Melarsoprol, an arsenic-based drug that inhibits parasite glycolysis, was, until recently, the only drug available for this, however it has significant risk of side effects [54]. Furthermore, there are signs of resistance to it in Central Africa [55].

Drug production, transport, storage and administration remain a major hurdle in the fight against HAT. Pharmaceutical companies had planned to halt production of anti-trypanosomatid drugs, however the WHO and Médecins Sans Frontiers campaigned against this, and secured a five-year, $25 million commitment from Aventis and Bayer to supply these five drugs over five years, ending in 2006 [6].

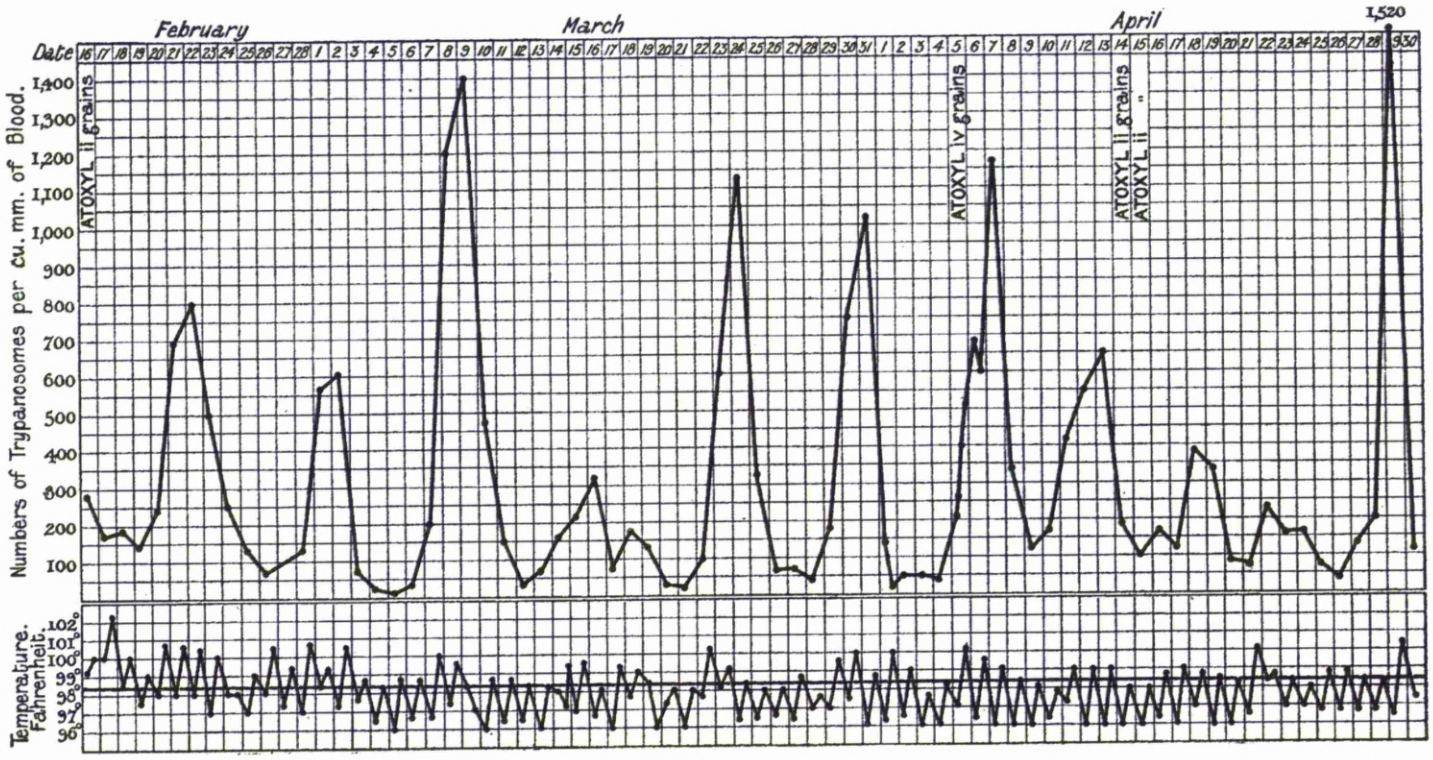


Figure 1.4: Numbers of trypanosomes in the blood, and the temperature of, a 26-year-old male patient that was infected with trypanosomes (likely *T. b. rhodesiense*) in 1909. Patient was treated in Liverpool with Atoxyl and Mercury at the times specified on the graph. The rise and fall of total parasitemia can now be attributed to the antigenic variation caused by the VSG repertoire of the trypanosome. Data and graph from Ross and Thompson, 1910 [56].

Due to the difficulties of treating the disease after infection and of immunising against Sleeping Sickness, the primary focus is on HAT prevention and therefore targeting the vector, the tsetse fly. Reducing the number of potentially infective flies in the wild is aided by the unusual life-cycle of tsetse – they are relatively slow to reproduce and lay larvae, commonly in ground-based leaf-litter. This renders them particularly susceptible to insecticides: the easiest and cheapest control programmes utilise pyrethroid-based insecticides on the ground, on odour-baited nets, or on cattle [57]. Additionally, sterile insect technique (SIT), previously used to control New World Screwworm in North Africa, Mexico and Central America, has been adapted to the tsetse fly, and has shown some positive results on *G. austeni* on Unguja island near Zanzibar [58]. SIT involves irradiating male flies, rendering them sterile, and releasing them into the wild to mate with the wild-type, female population, resulting in fewer offspring being born, however is only particularly useful after the application of more conventional techniques have already significantly reduced the population size. Nevertheless, simpler techniques such as the trapping of tsetse flies in nets and the spraying of insecticides (Figure 1.5) are more commonplace as they are relatively inexpensive, it is easier to transport and supply the necessary equipment and to train people in their effective use [59].

Figure 1.5: Vector control in sub-Saharan Africa. The image on the left shows tsetse nets being laid out and set-up; top-right shows a helicopter spraying insecticides; bottom-right shows caught tsetse flies. Images are provided courtesy of the Wellcome Trust.

Trypanosome genetics

Genome structure

The genome of the trypanosome is split into two distinct "units": the nuclear and the kinetoplast, or mitochondrial, genome. The nuclear genome is approximately 35Mbp in length and can be categorized into three sets of chromosomes based on broad size-ranges from their migration pattern in a pulsed field gel: The largest "megabase" chromosomes (1-6Mbp); The intermediate chromosomes (200-900Kbp) and the smallest "mini" chromosomes (50-150Kbp) [60].

At the time of publishing in 2005, the *T. b. brucei* genome sequence (GeneDB; version 3) had a total of 9068 putative annotated genes, 908 of which were related to VSG production [42, 61]. A further 907 pseudogenes have also been annotated, although the traditional term 'pseudogene' perhaps doesn't apply to trypanosomes; Defined as: "A DNA sequence that was derived originally from a functional protein-coding gene that has lost its function" [33] it has been suggested that modular recombination can allow parts of these genes to come together to form new versions of existing "active" genes, rendering them immensely variable and less "inactive" than the classical definition suggests [62].

Megabase chromosomes

T. brucei ssp have 11 pairs of diploid chromosomes >1Mbp in length (termed "megabase" chromosomes). The architecture of the chromosomes has been shown to be similar to that of other lower eukaryotes, sharing the same "three-tier" organisation of genes: a central core of housekeeping genes, proximal sub-telomeric domains containing species-specific genes and distal sub-telomeric domains containing variable surface glycoprotein (VSG) expression sites. Indeed, whilst most chromosomes have been shown to have differences along the length of the alternative homologues, the greatest variability in homologue diversity has been shown to occur at these sub-telomeric ends, presumably due to the necessity for the recombination associated with antigenic variation [63-65].

Unusual to *T. brucei* is the organisation of the genes along the chromosomes into "directional gene clusters" [66]. These regions are long stretches of coding sequences

along the same DNA strand separated by "strand-switch regions" [67]. As such, transcription in trypanosomes, mediated by RNA polymerase II, is polycistronic: hundreds of genes are co-transcribed and then regulated post-transcriptionally [reviewed [68]].

Small and intermediate chromosomes

The small "minichromosomes" (~100x; 10-20% of the total genome) predominantly consist of an interior, tandem array of 177-bp repeats and the same (TTAGGG)n telomeric repeats shared with the largest chromosomes. It has been suggested that the presence of these, plus VSG or VSG-like elements, may have a role in antigen-switching through interaction with the other copies on the main chromosomes [69]. There are also several intermediate chromosomes of uncertain ploidy, which vary in number and size depending on the strain. These have been also shown to contain VSG-like sequences and so may play a similar antigen-switching role to their smaller counterparts [70].

The kinetoplast

Contained within the single mitochondrion, the kinetoplast is arranged as a complex network of interwoven, circular kDNA and makes up around 20% (10Mbp) of the total DNA content of the organism. There are two types of these circles: minicircles (0.6-5.0K bp) and maxicircles (~22K bp), together encoding similar products to those of standard mitochondrial DNA in higher eukaryotes, including rRNAs and respiratory complexes [71]. Additionally, maxicircles also contain a number of so-called "cryptogenes" whose mRNA transcripts are edited from genes that initially don't have enough information to encode their respective proteins, to ones that can. This is achieved via a process directed by a set of guide (g)RNAs contained predominantly on the minicircles [72].

Population structure

Monitoring *T. brucei* epidemics since 1900 has revealed changes between and within the causative subspecies of the particular outbreak. Certainly early outbreaks of chronic Ugandan *T. b. gambiense* gave way to acute *T. b. rhodesiense* after treatment of the *G. fuscipes* vector with insecticide [73]. Nevertheless, pauses in said treatments allowed for subsequent re-emergence, possibly due to the spread from animal-based reservoirs of the disease that harboured the disease throughout a decline in tsetse numbers [74].

Understanding the epidemiology and population structure of sleeping sickness is therefore important to understand the problems underlying treatment, re-emergence and monitoring the possible spread of insecticide and/or drug-resistance throughout *T. brucei* populations. Understanding gene flow between populations, the spread of virulence and potential barriers to such spreading (such as gaps in geographical foci) can help inform clinicians treating the disease and scientists generate new models in aiding in its control [Reviewed [75]].

In terms of population structure, current knowledge suggests that *T. b. rhodesiense* is clonal. *T. b. brucei* is either panmictic, or epidemic in nature [76]. Smith had previously reported that *T. b. rhodesiense* structure may be epidemic with rapid and extreme expansions of particularly virulent sub-types [77].

Whilst genetic exchange is now widely accepted to occur during the insect salivary-gland stage of the life cycle of *T. b. brucei ssp.*, as proven by experimental mixed infections in tsetse [78, 79], its occurrence in natural populations remains a contentious issue, particularly for the human infective forms [80], but an important one due to the possibility of the spread of drug resistance [76], or the spread of human infectivity [81]. It has been shown that human serum resistance can be passed between stocks of *T. b. rhodesiense* and *T. b. brucei* by transfection of the SRA gene [82]. This raises the possibility that if *T. b. brucei* and *T. b. rhodesiense* interact, the larger genotypic diversity of *T. b. brucei* in the African wildlife and livestock population [83] could serve as a pool of potentially virulent genotypes, which could become infective to humans upon gaining the serum resistance trait [84].

Differences in clinical presentation and severity of infection can be found amongst different south-Eastern Ugandan Trypanosomiasis foci [85]. Similar studies show that these differences can even exist sympatrically; Bailey and colleagues reported that HAT patients presenting to Ugandan medical clinics between 1987 and 1993 showed different clinical manifestations of *T. b. rhodesiense* infection. In this study, one set of patients presented with short clinical histories with early symptoms such as fever and a chancre at the site of the Tsetse bite; others presented later, due to a lack of initial symptoms, and had a more chronic disease onset with symptoms that were more HIV/AIDS-like, including a lack of co-ordination and general malaise [86]. The existence of differing forms of *T. b. rhodesiense*-mediated sleeping sickness has been previously documented: It was observed that the symptoms in rats infected with *T. b. rhodesiense* differed based upon the geographical location of the origin of the parasite, and concluded that those strains that are more infective in man (from former epidemic areas) develop more slowly with lower levels of parasitemia in the rodent model [87].

Furthermore, Bailey reported that differential symptoms could be attributed to different genetic forms of the disease, as detected by isoenzyme analysis. Isoenzymes, or enzymes of similar function with different isoforms can be used to classify strains according to the assessment of their mobility on a thin-layer starch gel. At least two studies have attempted to classify *T. b. rhodesiense* according to similarities in their multilocus enzyme electrophoresis (MLEE) patterns [88, 89]. Several patterns can be grouped together to form a 'barcode' that uniquely identifies an individual or a group of isolates, and similarly, zymodemes can be grouped together into 'strain groups'; classically "*Busoga*" and "*Zambesi*" strain groups were identified from the regions where the zymodemes predominated. Little is known about the underlying genetic differences: *T. brucei* genome sequences currently only exist for *T. b. brucei* strain TREU927 [42], and more recently, *T. b. gambiense* [90]. *T. b. rhodesiense* is now thought to be a host-range variant of *T. b. brucei*, with the acquisition of the SRA gene, which confers human serum resistance on the parasite [122].

Many studies have attempted to characterise the population structure of, and the occurrence of genetic exchange between, *T. brucei* subspecies [91]. Whilst strains of *T. b. rhodesiense* have classically been characterised by MLEE [22, 89, 92], more recent studies have used neutral polymorphic markers such as micro- and mini-satellites to do so [93, 94]. This has allowed a genome-wide approach to study similarities between isolates from the same, and between different foci. Certainly, it is known that *T. b. brucei* isolates are much more genetically diverse than *T. b. gambiense*, which displays little diversity across much of its geographical range [95]. Outbreaks of *T. b. rhodesiense* in countries such as Zimbabwe, Tanzania, Malawi, Kenya and Uganda, have been recorded since the 1980s [32], each with distinct genotypes. Similarly, three different outbreaks in Uganda have been described as each having different genotypes [96].

Genetic exchange between *T. b. brucei* and *T. b. rhodesiense* is a possible cause for the increased genetic diversity of the latter [Reviewed [97]], the occurrence of which has been established in the laboratory setting [98]. Field isolates of *T. b. rhodesiense* are more closely related to sympatric *T. b. brucei* isolates than *T. b. rhodesiense* isolates from elsewhere, although *T. b. brucei* and *T. b. rhodesiense* isolates from the same focus remained distinguishable by a single minisatellite marker [76, 99].

Project objectives and methodology

This project seeks to utilise the improvements in genomic technology to identify candidate genes involved in parasite virulence and host resistance to *Trypanosoma* infections.

Identifying genes that regulate host response to trypanosomiasis

Some animals are tolerant to the African animal trypanosomiasis, remaining productive despite infection and maintain their body-weight and better control parasitemia. Symptoms in susceptible livestock include weight-loss, anaemia and cachexia, which render the animals unsuitable for farming. The local term for the animal trypanosomiasis – 'Nagana' - comes from the Zulu word for 'depressed' [100]. The wider introduction of trypanotolerant breeds of livestock, such as the West African N'Dama (*Bos taurus*), have been suggested as a method of controlling the effects of the

disease across Africa in place of vector control, or where drugs have been rendered ineffective by drug-resistance [101]. Whilst indigenous cattle are thought to be as productive as their trypano-susceptible counterparts [102], the latter are favoured by African farmers, hindering efforts to introduce tolerant breeds and thus control the disease in high tsetse challenge areas [103].

Mapping this trait in cattle has revealed 19 QTL on 17 Bovine chromosomes in N'Dama and Boran crosses [104]. The relationship between genotype and resistance to disease is complex, however, as resistance alleles were present in both breeds, suggesting that a synthetic breed involving alleles from both N'Dama and Boran would be more resistant than either parent [104].

Susceptibility and tolerance to Trypanosomiasis in the animal model

Scientists are aided in the identification of candidate genes and pathways by the presence of a mouse model of trypanotolerance. C57BL/6 mice are relatively resistant to the disease and survive for a relatively long period after infection with *T. congolense* strain IL1180 (110 days). Other strains are more susceptible, such as A/J (16 days), BALB/c (49 days) and C3H/HeJ (59 days) [105-107].

Generally, the major symptoms associated with experimental murine infections are loss of body weight, splenomegaly and hepatomegaly and anaemia, the latter of which is shared with both animal (eg. cattle) and human disease [108]. Anaemia during trypanosomiasis has been suggested to be due to the macrophage-mediated phagocytosis of both parasites and red blood cells [109]. Trypanotolerance in cattle is associated with the ability to control the associated anaemia and thereby remain productive despite persistent infection, which suggests that tolerant animals are better at controlling the pro-inflammatory and cytokine responses with a corresponding anti-inflammatory response [110].

Murine trypanotolerance has been shown to have a genetic basis, for which three QTL, *Tir1*, *Tir2* and *Tir3* respectively for *Trypanosoma Infection Response*, have been identified in crosses between resistant and susceptible inbred mice. Whilst the QTL regions have been reasonably well defined, moving from QTL regions to QTL genes remains a

major challenge: A review by Flint *et al* suggested that, at the time of publication, over 2,750 such quantitative trait loci (QTL) had been mapped in mice and rats but only twenty causative genes had been characterised [111]. A more recent search of the MGI database lists 3,963 QTL in mice (http://www.informatics.jax.org/; Accessed August 2010). The murine QTL described here still contain several hundred genes, any one of which, either individually, or in combination may be driving the resistance or the susceptibility phenotype.

Genome sequencing

Sanger and colleagues first developed the chain-termination method of DNA sequencing, and applying them to sequence the first microbial genome – that of the bacteriophage phiX174 in 1977 [112]. Despite alternative methods having already been developed [113], "Sanger sequencing" has been the bedrock for subsequent genomic studies, including that of the human genome project, completed in 2003 [114, 115]. The *T. b. brucei* TREU 927/4 sequencing project took several years to complete, using a chromosome-by-chromosome Sanger sequencing approach that involved separating the genome into its respective chromosomes and sequencing them individually. Whilst reducing the complexity of the assembly process by not having to assemble on a 'genome-wide' scale, this resulted in the intermediate and small chromosomes and the kinetoplast genome not being sequenced. The sequence of Type 1 *T. b. gambiense* (DAL972) was completed in 2010 using a Sanger-sequencing whole genome shotgun sequencing strategy [90].

Next-generation, (or better 'second-generation' sequencing instrumentation, as higher-throughput technologies are currently being released to market) now exist that allow for the sequencing of genomes in less than one week, trading decreased time and cost for decreased read-length (the length of a single stretch of a sequenced DNA fragment) [116]. Next-generation sequencing is not without its problems. Early "Sanger-sequenced" projects, such as that of the original *T. b. brucei* genome, were achieved due to the availability of relatively long read lengths sequenced in pairs. Such 'paired-ends' allow for sequence assembly across repetitive regions due to the ability to anchor a read in 'unique' sequence, even if its mate is within a repeat. Next-generation sequencing technologies are more difficult to assemble '*de-novo*' in this way, as read-lengths are

much shorter, and paired-ends are more difficult to achieve without large amounts of starting material (>10ug in some cases). The highest throughput capillary-based Sanger sequencing machines generally produce 96 reads every 100 minutes at read-lengths of between 500-800bp (http://www6.appliedbiosystems.com; Life Technologies website. Accessed October 2010). By comparison, at the end of 2010, the highest throughput sequencer available, the Illumina HiSeq 2000, generates around two billion reads in 8 days at read-lengths of around 100bp (http://www.illumina.com; Illumina Inc. website. Accessed October 2010).

As such, many next-generation sequencing projects rely on a high-quality, often Sanger-sequenced, reference to which to map the sequence reads. Such 'aligning' allows for resequencing and SNP and small insertion / deletion (indel) discovery, but precludes the discovery of large-scale insertions and differences from the reference on which the data are based. Nevertheless, these technologies offer the opportunity to rapidly resequence isolates of the same strain of parasite, or even similar species and accurately identify differences between them. Whilst alignments such as those described are much faster and require less computer power, it is often preferable to assemble reads *de novo* without the use of prior information as described as this imparts no bias on the overall outcome. The ability to sequence and assemble reads in this manner is, however, dependent on a number of considerations, such as the available computing power and the read-length obtained by the sequencing technology, as shorter read lengths tend to preclude *de novo* assembly [116]. Generally, however, the read-length of the given technology used often dictates the experiment, namely, in the presence of a reference sequence, an ABI SOLID or Illumina instrument (with a shorter read length, but greater output) might be chosen; in the absence of such a reference, a longer-read instrument such as the Roche 454 GS-FLX might be used.

These technologies can also be applied to the host: Several human [117] and mouse [118] genome sequences are already publicly available. The size of mammalian genomes requires a large amount of sequencing, even on next-generation sequencing instruments, with only the newest instruments able to generate an entire human genome in a single run. Technologies have also been developed to focus the large number of reads generated on smaller, targeted regions. So called 'targeted-resequencing' can be achieved by the amplification of specific regions by PCR [119] or

by preferentially capturing regions of interest on microarrays [120], and subsequent sequencing on a next-generation instrument.

Inference of population structure in differentially virulent isolates of *T. brucei rhodesiense*

Various methods have been developed to elucidate the population structure of sleeping sickness due to the importance of understanding instances of re-emergence or the emergence of new virulent epidemics. It is clear that there is variability within and between sub-species of *T. brucei*, established by a number of molecular methods including MLEE [121], minisatellite genotyping [76], microsatellite genotyping or DNA sequencing (of the SRA gene) [122].

There are numerous software tools that can aid researchers in understanding the population structure derived from genotyping or sequence data. Arguably the gold standard is STRUCTURE [123], which uses a Markov chain Monte Carlo (MCMC) method: By iteratively increasing the predicted number of populations it can simulate data based on a number of prior assumptions and assess the similarity of genotyping datasets to the said model. In this manner, it can estimate the proportion of each individual's genotype that is derived from each of a set of pre-determined number of populations (e.g. given six populations, whether an individual is made up equally of said six populations, or there is more bias towards any particular population). STRUCTURE, however, is known to falter given hierarchical datasets: predictions of the number of populations within datasets tends towards the lowest number of populations regardless of whether some sub-structuring is present in the data. For instance, given a set of two closely related subspecies, STRUCTURE may preferentially report two populations, whereupon further analysis would be required to reveal any further populations within the data. As such, methods have also been established to further identify the number of populations within STRUCTURE-derived datasets [124]. Balmer *et al*, used such methods to identify eleven populations present within *T. brucei* subspecies across Africa, including several mixed populations of *T. b. brucei* and *T. b. rhodesiense*, lending further weight to the hypothesis that *T. b. rhodesiense* is a genotypically varied host range variant of *T. b. brucei* [122]. Software also exists that negates the need for prior assumptions: BAPS (Bayesian Analysis of Population

Structure) adds Bayesian statistics to MCMC to automatically identify the most likely number of populations [125].

Sympatric (samples from the same focus) zymodemes of *T. b. rhodesiense* have shown differential virulence in humans and mice. It may be possible to infer the population structure of different zymodemes from the 1980's/early 1990's epidemic in Uganda and identify genetic loci that underlie differences in virulence between zymodemes. Furthermore, as *T. b. rhodesiense* is thought to be a host range variant of the Sanger sequenced (and well finished) *T. b. brucei* [42], the *T. b. rhodesiense* genome should be well suited for next-generation sequencing and mapping against the reference strain.

Chapter Two

The identification of candidate genes that may be responsible for regulating survival in mice after infection with *Trypanosoma congolense* strain IL1180

Abstract

About one-third of cattle in sub-Saharan Africa are at risk of contracting "Nagana" – a disease caused by *Trypanosoma* parasites similar to those that cause human "Sleeping Sickness". Laboratory mice can also be infected by trypanosomes, and different mouse breeds show varying levels of susceptibility to infection, similar to what is seen between different breeds of cattle. Survival time after infection is controlled by the underlying genetics of the mouse breed, and previous studies have localised three genomic regions that regulate this trait. These three "Quantitative Trait Loci" (QTL), which have been called *Tir1*, *Tir2* and *Tir3a-c* (for *Trypanosoma Infection Response 1-3a-c*) are well defined, but nevertheless still contain over one thousand genes, any number of which may be influencing survival.

By systematically combining mapping in an additional susceptible breed, next-generation DNA capture and sequencing and SNP annotation we have developed a strategy that can generate a short list of polymorphisms in candidate QTL genes that can be functionally tested. Mapping loci regulating survival time after *T. congolense* infection in an additional cross revealed that susceptible C3H/HeJ mice have alleles that reduce survival time after infection at *Tir1* and *Tir3* QTL, but not at *Tir2*. Next-generation resequencing of a 6.2 Mbp region of mouse chromosome 17, which includes *Tir1*, identified 1,632 common single nucleotide polymorphisms (SNP) including a non-synonymous SNP in *Pram1* (PML-RAR alpha-regulated adaptor molecule 1), which is an intracellular adaptor involved in T-cell signalling. The protein is important for neutrophil function and shares structural homology with adaptor proteins involved in integrin activation, which is essential in leukocyte adhesion and subsequent cytotoxicity and inflammation in response to infection. The non-synonymous SNP was predicted to be 'probably-damaging' and as such, *Pram1* is the most plausible candidate QTL gene in *Tir1*.

Introduction

Animal African trypanosomiasis in livestock is mainly caused by two species of trypanosomes: *T. vivax* and *T. congolense*. The disease affects over ten million Km^2 of Africa and it is estimated that some thirty percent of Africa's 160 million cattle are at risk of infection. Losses of livestock and crop production are estimated at over $1 billion per annum [20].

Scientists are aided by a mouse model of trypanotolerance, as African trypanosomes also infect laboratory mice in which susceptibility is measured by survival time after infection, which varies between inbred lines. Whilst C57BL/6 mice survive for a relatively long period after infection with *T. congolense* IL1180 (110 days), some other strains, such as A/J (16 days), 129/J (23 days), BALB/c (49 days) and C3H/HeJ (59 days) mice are relatively susceptible [105-107]. Mapping studies, initially undertaken in two independent F2 crosses: C57BL/6JOlaHSD (C57BL/6) × BALB/cOlaHsd (BALB/c) and C57BL/6JOlaHSD × A/JOlaHsd (A/J), identified three major QTL regulating survival time [126]. These were mapped to mouse chromosomes 17, 5 and 1 and have been designated *Tir1*, *Tir2* and *Tir3* respectively for *Trypanosoma Infection Response*.

There are a number of strategies available for the high-resolution mapping of QTL, including backcrosses, recombinant inbred lines, congenic mice and advanced intercross lines (AILs) (Reviewed [111]). AILs, which are generated by first producing an F2 intercross from an initial F1 parental cross and then randomly intercrossing for a number of generations, have been estimated to produce confidence intervals that are N/2 smaller than similar results from the F2 progeny (where 'N' is the number of additional intercrossed generations) [127]. Using this method, trypanotolerance loci were refined using F6 crosses between A/J and BALB/cJ strains with resistant C57BL/6J [128]. This reduced the size of the confidence intervals for *Tir1* to less than 1 cM, and *Tir2* and *Tir3*, to within 5-12 cM and resolved *Tir3* into three smaller regions, termed *Tir3a*, *Tir3b* and *Tir3c*. These lines were subsequently extended to F12 generations, in which the sizes of the 95% confidence intervals of each of the QTL were substantially reduced to between 0.9 and 7.2cM. [129].

The refinement of candidate gene numbers through comparative genomics

The completion of the C57BL/6J mouse reference genome sequence in 2002 [130] provided a backbone on which subsequent comparative studies could be performed. Initially, research was restricted to comparison with other, finished, vertebrate genome sequences – predominantly the human genome, which had been completed in the previous year [114]. Advances in array-based resequencing technology (Affymetrix) allowed Perlegen to rapidly resequence the genomes of an additional fifteen inbred strains of mouse. The technology used, however, was limited to identifying discordant bases in 25bp probes at an equivalent of 1.5X coverage if performed by Sanger sequencing. As such, the Perlegen dataset of approximately eight million polymorphisms has been estimated to be about 45% complete [131].

The polymorphisms discovered were not randomly distributed across the mouse genomes. Regions of high and low single nucleotide polymorphism (SNP) density can be found in a comparison between any two strains as a consequence of the mosaic of ancestral genomes from which laboratory mice are derived: High SNP density occurring in regions where the two strains had distinct ancestors and low SNP density occurring in regions of common ancestry. It has been estimated that at any given position there are usually two ancestral haplotypes [132]. Perlegen made available a haplotype map showing breakpoints between haplotype blocks derived from different ancestral strains for each of the sixteen available genomes, alongside the SNP data (http://mouse.perlegen.com/mouse/index.html; Accessed August 2010).

By assigning genomic regions to different ancestral haplotypes, it is possible to identify those that reside within regions of similar ancestry across strains with similar phenotypes [133, 134]. For instance, in the case of trypanotolerance, with resequencing data from a number of susceptible breeds across *Tir1*, the QTL can be refined to regions of shared ancestry that are different to that of the resistant C57BL/6.

Next-generation sequencing technologies

New sequencing technologies are now making it possible to identify a large proportion of the differences between common inbred mouse strains. At present this is possible for defined areas of the genome (so called "targeted resequencing"), but public data sets will soon be available for whole genomes. Currently underway is the Mouse Genomes Project at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk /resources/mouse/genomes), which is resequencing the whole genomes of seventeen key laboratory strains of inbred mouse utilising the Illumina (Solexa) GA sequencing platform. Their first publication, in October 2009, coincidentally compared the A/J and CAST/Ei mouse chromosome 17 (which includes *Tir1*) to the C57BL/6 reference sequence and identified candidate genes regulating triglyceride levels in the liver [135]. The benefits of large-scale resequencing are two-fold: Firstly, one can identify novel SNP that are shared between susceptible lines and that may have a functional effect and thus may be driving differences in phenotype; and secondly SNP density can be increased across one, or several, QTL so as to increase the resolution of haplotype block assignment to candidate genes.

Next-generation sequencing technologies are not without their problems, and are particularly susceptible to false-positives and negatives under regions of low sequence coverage. Technologies also have problems specific to the particular technology: The Roche/454 technology is pyrosequencing based, and as such cannot reliably predict the number of bases within a homopolymeric tract. Nevertheless, 454 sequencing benefits from having longer read-lengths than other second-generation sequencers (mean >400bp), which enables both *de novo* assembly and high-confidence alignment against a reference. The proprietary software distributed with the Roche/454 platform – Newbler – was specifically developed with this in mind, and is adept at dealing with and correcting for 454-sequencing problems associated with long homopolymeric tracts.

Aims and Objectives

Despite the mapping of the three *Tir* QTL, there still remains >1000 genes within the predicted boundaries of the QTL, any number of which may underlie the response to infection with *T. congolense* IL1180. In order to reduce this number, it may be possible to apply next-generation sequencing technologies and SNP annotation to novel, and pre-existing, publicly available, sequencing datasets to prioritise genes for subsequent functional testing, and to discount genes without significant polymorphisms.

Assuming that where QTL overlap in multiple susceptible strains, that this effect is due to shared polymorphisms, then the number of candidate genes within a QTL can be reduced through the removal of any genes in a susceptible strain that share a similar allele, or ancestral haplotype, with the resistant (C57BL/6) strain. Although it is known that C3H/HeJ mice are relatively susceptible to *T. congolense* infections [105], it is unknown whether this is due to the presence of the *Tir* QTL. To that end, we have mapped QTL in a C57BL/6 x C3H/HeJ cross so that we now know whether a total of four mouse strains carry either the susceptible or the resistant allele at each *Tir* QTL. Furthermore, we have sequenced one of the QTL regions in three additional strains of susceptible mice to identify SNP that correlate with phenotype and have used Polyphen to identify the non-synonymous SNP in the QTL regions that are most likely to alter the activity or function of a candidate gene. A fourth mouse strain, 129P3, was included in the resequencing to assess whether the *Tir* QTL should be mapped in that strain.

By combining the resequencing data with publicly available SNP, SNP density at *Tir1* can be increased so as to improve haplotype block assignment. Genes residing in haplotype blocks with similar ancestry to the resistant strain can then be removed from lists of potential candidates.

Materials and Methods

Care and use of laboratory animals

All C3H/HeJ x C57BL/6 animal work was undertaken at ILRI, Kenya and was performed under IACUC ref no 2003.19. The ILRI IACUC complies voluntarily with the UK Animals (Scientific Procedures) Act 1986 that contains guidelines and codes of practice for the housing and care of animals used in scientific procedures. All animals on survival experiments were regularly monitored to check for signs of terminal illness, and any showing such signs were euthanised by UK Schedule 1 procedures.

Identification of the *Tir* QTL in C3H/HeJ mice

C3H/HeJ x C57BL/6 cross

Peris Amwayi and Fuad Iraqi (ILRI, Kenya), performed the cross of C3H/HeJ x C57BL/6 mice and subsequent phenotyping as follows: C57BL/6JOlaHSD (C57BL/6) and C3H/HeJ mice were obtained from Harlan Laboratories. Mice were infected with 4×10^4 *T. congolense* strain IL1180 intraperitoneally (i/p) as previously described [126]. Any mice that did not develop a microscopically proven parasitaemia were removed from the study. 345 F2 C3H/HeJ x C57BL/6 mice were phenotyped for survival time after infection with *T. congolense* strain IL1180.

Genotyping of the C3H/HeJ x C57BL/6 cross

All markers used and associated primer sequences are listed in Appendix I: Table 1. PCR reactions were performed using Reddymix (Thermo) with 20ng of template DNA. Cycling conditions were as follows: 95°C, 50secs; [Tm -5]°C, 50secs; 65°C, 50secs; 30x cycles. PCR products, including negative controls, were resolved by ethidium bromide stained agarose-gel electrophoresis and visualised under UV-light. Microsatellite primers incorporated a 5'-fluorescent label, which enabled the accurate sizing of SNP on an ABI-3130XL capillary sequencer. SNP were genotyped by directly sequencing PCR products as follows: Unincorporated primers and residual nucleotides were degraded using ExoSAP-IT (USB Corp, Ohio, USA) and sequencing products generated using Big-Dye v3.1 terminators (Applied Biosystems, Foster City, USA).

Cycle sequencing products were ethanol precipitated and electrophoresed on an Applied Biosystems ABI-3130XL capillary sequencer. Microsatellite and SNP genotyping data were viewed using PeakScanner (Applied Biosystems) and GAP4 [136] respectively. Mean survival for mice at each marker for each given genotype is shown in Appendix I.

Tir1, *Tir2* and *Tir3*(a-c)

345 F2 C3H/HeJ × C57BL/6 mice were phenotyped for survival time after infection with *T. congolense* strain IL1180 by Fuad Iraqi and Peris Amwayi at ILRI, Kenya.
Choosing only the extremes of a phenotypic distribution for subsequent genotyping reduces genotyping costs with little loss of power to detect QTL, however it does give exaggerated estimates of effect sizes [137]. The 94 animals that had the most extreme survival times (<62 days and >141 days; Figure 2.1) were genotyped at eighteen microsatellite and SNP loci across the three *Tir* loci.

Tlr4

C3H/HeJ mice carry a Pro to His mutation at position 712 of the *Tlr4* gene that makes this mouse strain insensitive to LPS. This spontaneous *Tlr4* null mutant makes it possible to test whether TLR4 is involved in regulating survival after infection with *T. congolense* [138]. As such, the mice were genotyped at the *Tlr4* locus using a closely related microsatellite marker (D4mit178). Additionally, the functional SNP (rs3023006) was sequenced as previously described.

U4 / *U6*

U4, a small nuclear RNA (snRNA) that is a member of the spliceosome complex [139], and *U6*, with which it forms a functional duplex, were identified to be candidate genes at two different QTL (*Tir2* and *Tir3b*, respectively). Microsatellite data for 676 F6 AIL C3H/HeJ x C57BL/6 crossed mice was obtained for four markers, two at each of the U4 and U6 snRNA [129]. These were used to identify significant linkage disequilibrium (chi-square test) between the two loci: D5mit113 and D5mit10 (*U4*); and D1mit102 and D1mit425 (*U6*).

Targeted resequencing of *Tir1* in susceptible mice

DNA capture and sequencing

Genomic DNA for BALB/cJ (Jackson #000651), 129P3/J (Jackson #000690), A/J (Jackson #000646) and C3H/HeJ (Jackson #000659) were obtained from the Jackson Laboratories and submitted to Nimblegen for sequence capture [120]. Capture probes were designed to cover 4.5Mbp of non-repetitive sequence between 30,637,692bp and 36,837,814bp on mouse chromosome 17 (NCBI37). 385,000 60mer probes were tiled at approximately 5bp intervals leading to a mean of 12 probes over each base.

Captured DNA was sequenced on a Roche 454 FLX Genome Sequencer using Titanium chemistry (Roche) according to the manufacturer's protocols by Dr. Margaret Hughes, University of Liverpool. The four sequencing libraries (one for each mouse breed) were each sequenced on one region of a PicoTitrePlate© (PTP), using a total of two PTPs for the four experiments.

Sequence assembly and SNP calling were performed using the Newbler mapping algorithm, which aligned 454 reads against the Ensembl C57BL/6 mouse reference (NCBI37) and outputs lists of SNP and associated coverage metrics in a tab-delimited format. As pyrosequencing is known to miscall sequences either across, or either side of, homopolymeric tracts (long stretches of a single nucleotide), differences were removed from subsequent analysis if they were within 13bp of a homopolymeric tract ≥5bp [118] using a bespoke perl script (Appendix IX: Additional data file 1). SNP were subsequently entered into a MySQL database wherein they were additionally filtered to those with at least eight-fold coverage and occurring in at least 87.5% of the reads sequenced across any polymorphic position.

14,440 high-confidence genotypes were submitted to dbSNP with SSIDs ss159831440-ss159845897. 454 reads were submitted to the European Short Read Archive under Accession number ERA000179.

A 24-bp insertion in *Mdc1* in susceptible strains was verified by PCR amplification and subsequent agarose gel electrophoresis and capillary-based dideoxynucleotide sequencing as previously described using the primers in Appendix Table A2.1.

Functional SNP identification

SNP were aligned against coding sequences and non-synonymous SNP were identified using a bespoke perl script (Appendix IX: Additional data file 2). nsSNP were extracted by substituting the SNP into the reference coding sequence, translating the sequence to the associated amino acid code and identifying changes in this sequence. Similarly, BLOSUM scores [140] were obtained for any changes observed in this manner. SNP positions were compared to the mouse regulatory build to test for SNP that may alter transcription factor binding sites or promoter regions [141, 142]. nsSNP were manually annotated with Polyphen [143].

Publicly available functional SNP confirmation

Functional SNP at *Tir2* and *Tir3* in the Perlegen dataset, for which genotypes for all five mouse strains were not available, were confirmed in C57BL/6, A/J, BALB/cJ and 129P3 mice using PCR and dideoxynucleotide sequencing as described for genotyping. Sequences that showed evidence of multiple copies were cloned using TOPO-TA cloning kit (Invitrogen) according to the manufacturer's protocols and sequenced as previously described.

Results

Identification of *Tir1* and *Tir3* QTL in C3H/HeJ mice

By increasing the number of breeds known to carry susceptible alleles at the QTL, candidate gene lists can be refined to remove those genes that are in QTL for *T. congolense* infection response but have the same ancestral haplotype as the resistant strain in at least one susceptible mouse breed. The three major *Tir* QTL have only been identified in C57BL/6, A/J and BALB/c mice, with C57BL/6 carrying the resistant allele at each locus. To that end, we measured survival after infection in an inter-cross between another susceptible breed, C3H/HeJ, and C57BL/6 mice. For the cross, the mean survival times of parental founder lines for the C3H/HeJ × C57BL/6 F2 cross were 63 days for C3H/HeJ and 87 days for C57BL/6. Out of the 345 F2 C3H/HeJ x C57BL/6 mice that were phenotyped, we selectively genotyped the 94 mice (51♂ and 43♀ p=0.41) that had the most extreme survival times (Figure 2.1) with microsatellite and SNP markers across the three known QTL. Figure 2.2 & Table 2.1 show that C3H/HeJ carries alleles that reduce survival time at the *Tir3* QTL on Mmu1 and the *Tir1* QTL on Mmu17. No QTL was discovered on Mmu5 in the region of *Tir2*.

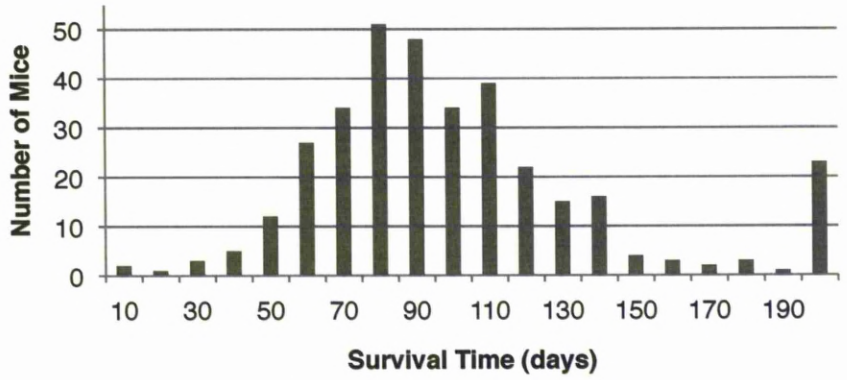


Figure 2.1: Distribution of survival times of C3H/HeJ x C57BL/6 F2 mice after infection with *T. congolense*. Labels on the X axis indicate the start of each interval; hence the interval labeled 10 includes animals that died between days 10-19. The peak at 210 days is for mice that were surviving at the end point of the experiment. These mice were marked as having survived for 141 days and were included in the genotyping.

A null allele of the *Tlr4* gene in C3H/HeJ does not affect survival

A functional *Tlr4* gene is necessary for maximal control of *Trypanosoma cruzi* in mice [144] and there is evidence that the GPI anchor of *T. brucei* VSG has endotoxin like properties that could stimulate *Tlr4* [145]. C3H/HeJ has a polymorphism in the toll like receptor 4 (*Tlr4*) gene, on mouse chromosome four, that ablates its function, making these mice insensitive to LPS [138]. This spontaneous mutation was used to discover whether *Tlr4* was as important in the response to *T. congolense* as to *T. cruzi*. Since all previous mapping had been done in mice with intact *Tlr4* genes, no QTL could have been detected at this locus even if *Tlr4* is involved in the response to infection. The C3H/HeJ × C57BL/6 mapping population could therefore be used to discover whether this gene (or a closely linked one) is involved in the regulation of survival time after infection. Mice were genotyped with a microsatellite marker linked to the functional polymorphism and sequenced across the polymorphic position. There was no significant association with either of these markers and survival time, indicating that the *Tlr4* pathway does not affect survival after *T. congolense* infection in mice (Figure 2.2; Table 2.1 and Appendix I: Table A1.6.2).

U4 and *U6* at *Tir2* and *Tir3b* do not interact

U4 is a component of the spliceosome in which it forms a duplex with *U6*. It was interesting to observe that both members of the U4/U6 duplex appear under different QTL, at *Tir2* and *Tir3b* respectively, however *U6* is one of 582 other similar sequences in the mouse genome and analysis of mapping data for an F6 C3H/HeJ x C57BL/6 AIL showed no evidence for an interaction between the *U6* and *U4* loci at *Tir3b* and *Tir2* respectively (Table 2.2).

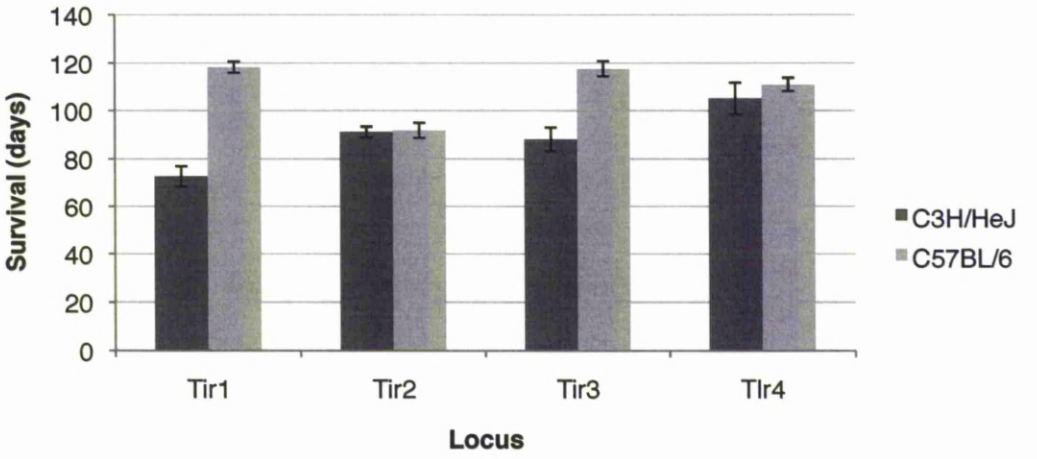


Figure 2.2: Mean survival of C3H/HeJ x C57BL/6 mice infected with *T. congolense* strain IL1180 at three trypanotolerance QTL (*Tir1-3*) and the *Tlr4* gene. Bar chart of mean survival (days ± standard error) of 94 F2 C3H/HeJ x C57BL/6 crossed mice grouped by genotype at three trypanotolerance QTL and at the *Tlr4* gene. Results are displayed grouped by resistant (C57BL/6) or susceptible (C3H/HeJ) genotypes for each locus, alongside the number of markers tested at each trypanotolerance QTL.

Table 2.1: Genotypes and associated survival of C3H/HeJ x C57BL/6 crossed mice. Differences in mean survival between homozygous genotypes were significantly different at *Tir1* and *Tir3*, but not at *Tir2* (T-test)

QTL /Gene	Genotype count			# Markers tested	Mean survival (days) grouped by genotype			Δ mean survival between homozygous genotypes (days)	P-value (T-test) for significant difference in mean survival
	C3H/HeJ	Heterozygous	C57BL/6		C3H/HeJ	Heterozygous	C57BL/6		
Tir1	82	247	244	6	72.69	102.79	118.20	45.51**	<0.001
Tir2	321	168	180	8	104.14	112.01	104.97	7.87	0.348
Tir3	75	167	129	4	88.12	104.25	117.52	29.40§	<0.001
Tlr4	38	107	47	2	105.06	103.73	110.99	5.93	0.502

Table 2.2: Genotyping of 676 F6 AIL C3H/HeJ x C57BL/6 F2 cross to identify interaction between U4 and U6 at *Tir2* and *Tir3b*. P-values (Chi-squared test) for whether alleles at U4 (D5mit113; D5mit10) and U6 (D1mit102; D1mit425) are inherited independently from each other in 676 crossed F6 C3H/HeJ x C57BL/6 mice [129]. $P < 0.05$ indicates that the observed and expected frequency of genotypes do not differ significantly.

Genotype	D5mit113/ D1mit102	D5mit113/ D1mit425	D5mit10/ D1mit102	D5mit10/ D1mit425
Alleles inherited within Hardy-Weinberg equilibrium (P-value; Chi Square)	0.70	0.26	0.38	0.40

Sequence capture and sequencing of *Tir1*

DNA from across the *Tir1* QTL was sequenced in order to characterise novel SNP and to improve the identification of alternate alleles for each haplotype block. DNA from four mouse breeds: 129P3, A/J, BALB/c and C3H/HeJ; was captured on Nimblegen arrays with probes for a 6.2Mbp region of mouse chromosome 17 between 30,637,692 and 36,837,814 (NCBI37). 1.7Mbp of repetitive sequence was excluded. Captured DNA was sequenced on a Roche 454 Genome Sequencer FLX using Titanium chemistry. 1,308,175 reads were mapped to the C57BL/6 reference sequence giving an average ~15X coverage of each sequenced strain (read length ~282bp; total sequence ~370Mbp). Plotting sequence coverage across the resequencing region revealed an increase in coverage at the proximal end of the target region in A/J (mean 3-fold), C3H/HeJ (mean 2-fold) and 129P3/J (mean 2-fold). No similar increase in coverage was observed for BALB/cJ (Figure 2.3).

SNP extraction and filtering

Three filters were applied to exclude false-positive SNP: Firstly SNP were excluded in genomic regions only covered by sequence coverage <8X; Secondly, SNP were excluded if <87.5% of the reads at a given position did not display the SNP; Finally, SNP were excluded if they were within a 13bp window of a homopolymeric tract ≤5bp. In this manner, 14,440 SNP loci were identified, 3,618 of which were not in dbSNP build 128. 1,588 loci were common to A/J, BALB/c and C3H/HeJ, but differed from C57BL/6. Furthermore, upon adding data for 129P3, there were 466 SNP loci common to all four sequenced mouse strains. Summary statistics for all SNP are available in Table 2.3. Figure 2.4 shows a circular plot of all SNP called by the Roche/454 mapping algorithm (Newbler) against the C57BL/6 reference. Haplotype blocks can be seen as clusters of high-densities and low-densities of SNP. Whilst at this resolution it is not easy to see haplotype blocks in the A/J, BALB/c or C3H/HeJ data, one haplotype block stands out in the 129P3 data where 81 common SNP clustered within a 430 Kbp region (33,245,853bp–33,675,688bp).

Table 2.3: Summary statistics for the 454 GS-FLX (Titanium) sequencing of Nimblegen array captured material from Mmu17 (30,637,692bp–36,837,814bp; NCBI37) in four breeds of mouse. Filtered SNP are those SNP remaining after filtering to remove SNP within a 13bp window of homopolymeric tracts and outside capture probe regions. SNP must have sequence coverage >8X, 7 of which must match the alternative allele. Novel genotypes are SNP either not previously characterised, or disagree with previous genotypes (dbSNP128).

	A/J	BALB/cJ	C3H/HeJ	Common Loci (3 strains)	129P3/J	Common Loci (4 strains)
Total SNP	19950	19136	12890		5616	
Filtered SNP	7969	7435	6046	1588	2160	466
Novel Genotype	1615	7327	4207	150	2160	36
% Novel genotypes at novel loci	94%	25%	30%	9%	19%	8%

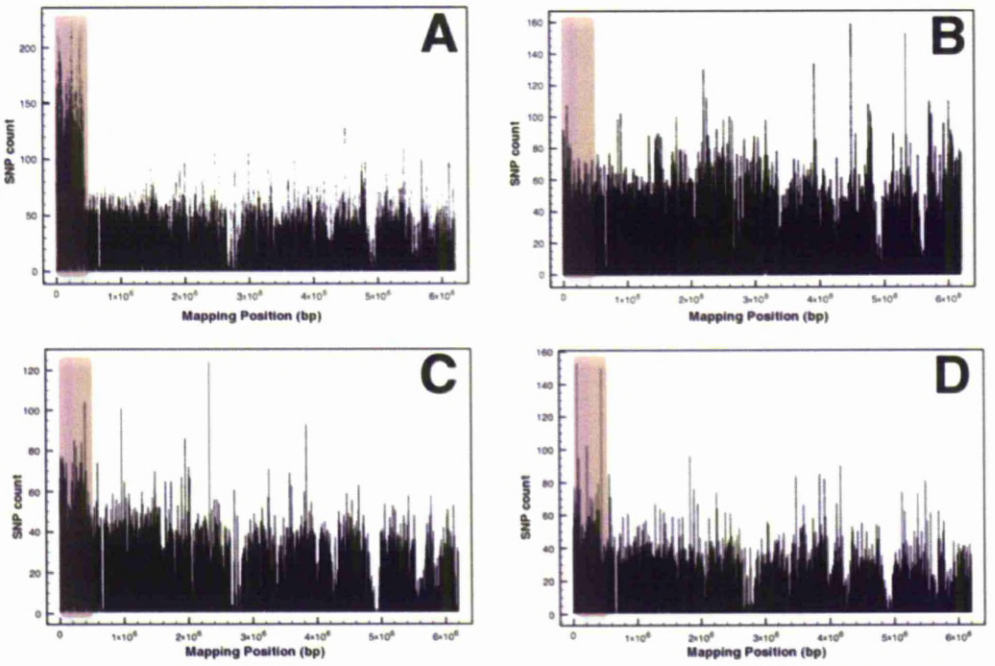


Figure 2.3: Sequence coverage of 454 resequencing of four breeds of inbred mouse. Sequence coverage across the *Tir1* resequencing region (Mmu17: 30,637,692bp–36,837,814bp) for four strains of experimental mice: A/J (A); BALB/cJ (B); C3H/HeJ (C) and 129P3/J (D). Increased read coverage can be seen at the proximal end (highlighted) for A/J (mean 3-fold increase), C3H/HeJ (mean 2-fold increase) and 129P3/J (mean 2-fold increase), which correlates with a copy number variation (CNV) at the *Glo1* locus in these strains [146]. Positions displayed are relative to the start of the resequenced region.

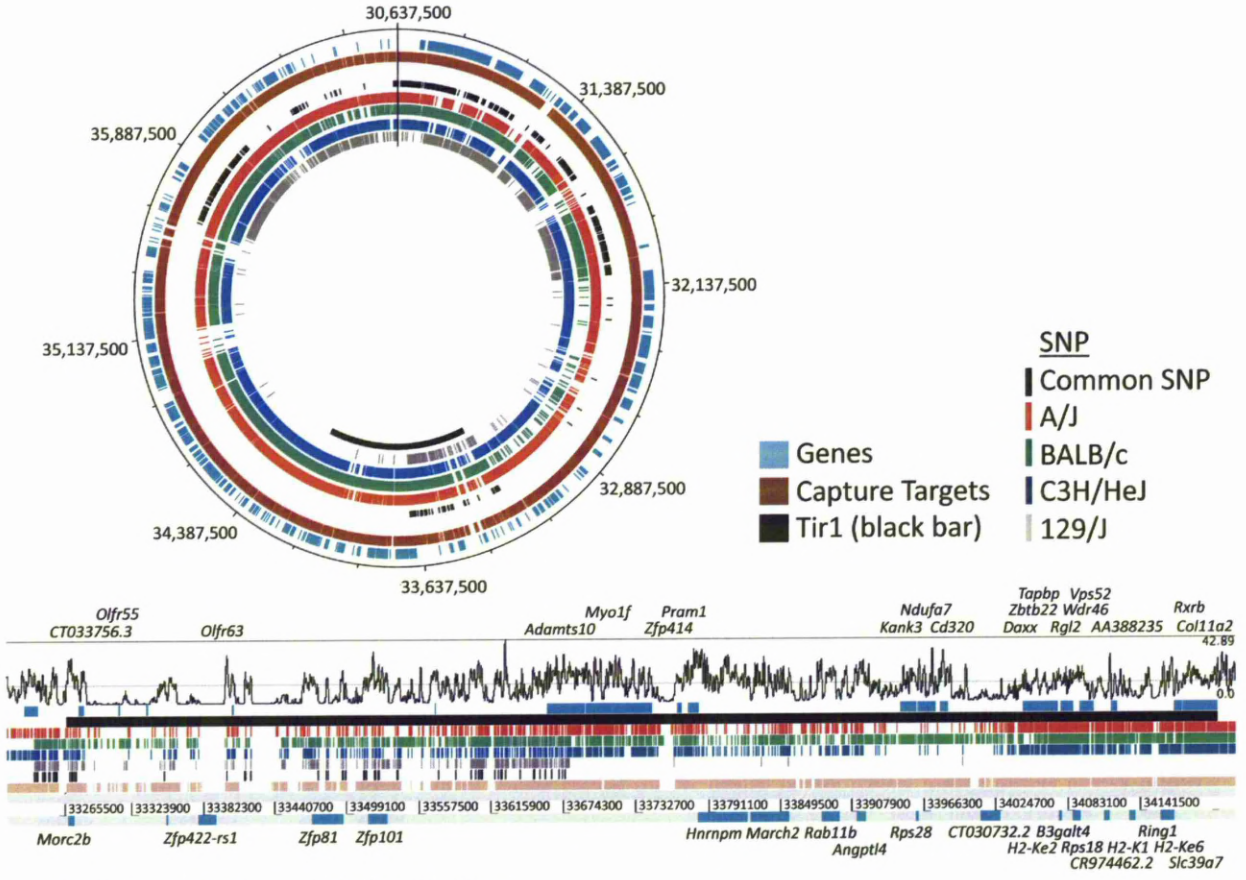


Figure 2.4: Targeted resequencing of *Tir1* in susceptible breeds of mice

Array-based sequence capture and next generation sequencing of a 6.2Mbp region of Mmu17 in four breeds of mice: A/J; BALB/c; C3H/HeJ and 129/J (Mmu17:30,637,692bp–36,837,814bp). Plot is circular for ease of display. *Tir1* is highlighted in black on the inside track. Genomic positions are in Mbp. The outer tracks (blue and brown) show genes and designed capture probes, respectively. The four, coloured, inner tracks show SNP called in each of the four sequencing experiments, with the black tick marks highlighting areas of common SNP. Haplotype blocks can clearly be seen as clustering of high- and low- density regions of SNP. Close-up region around *Tir1* is displayed underneath the circular plot. Tracks are identically coloured and includes a moving average (window 1 Kbp) of sequence read coverage across the region (top). Genes in the region are described for the forward strand (above) and reverse strand (below).

Susceptible mice have a 24bp insertion within the 3'-UTR of *Mdc1* relative to C57BL/6 mice

454 pyrosequencing identified a 24bp insertion in A/J, BALB/cJ, C3H/HeJ and 129P3/J mice relative to the C57BL/6 reference that mapped to the 3'-UTR region of *Mdc1* (Mediator of DNA damage checkpoint protein 1; Mmu17: 35,981,380-35,996,614bp). The insertion was confirmed by PCR, by which the insertion could be clearly seen as two different sized bands between C57BL/6 and susceptible mice (data not shown). The insertion was further confirmed by Sanger dideoxynucleotide sequencing. Figure 2.5 shows an alignment of the DNA sequences, with the deletion in C57BL/6 clearly marked by gaps in the alignment (indicated by asterisks) highlighted by a blue background. Positions on the figure are relative to the start of the DNA coding sequence.

Figure 2.5: Confirmation of a 24bp insertion in the 3'-UTR region of the *Mdc1* gene
An alignment of sequencing reads from PCRs across a predicted 24bp deletion by 454 pyrosequencing in C57BL/6. The deletion can be seen clearly marked by gaps in the alignment (indicated by asterisks) highlighted by a blue background. Positions on the figure are relative to the start of the DNA coding sequence.

Non-synonymous polymorphisms

Using all available data, the *Tir1* region contained 70 nsSNP loci that correlated with phenotype (Appendix II: Table A2.2). Analysing both the 454, and publicly available data, there were seven 'possibly damaging' nsSNP and three 'probably damaging' (Polyphen) nsSNP in: PML-retinoic acid receptor alpha regulated adaptor molecule 1 (*Pram1*) (rs33399614); *Rgl2* (Ral guanine nucleotide dissociation stimulator-like 2); and CR974462 (Table 2.4).

Table 2.4: A list of non-synonymous SNP loci within *Tirl* that are predicted to be damaging according to Polyphen. Genes are within *Tirl* (Mmu17: 33,271,855bp–34,203,529bp) with damaging nsSNP that correlate with survival phenotype and are classified as either "possibly" or "probably" damaging according to Polyphen [143]. A full list of annotated SNP is available in Appendix II: Table A2.2.

	Allele							
Position (bp)	**C57BL/6**	**A/J**	**BALB/c**	**C3H/HeJ**	**PhastCons**	**Gene**	**Polyphen Consequence (Damaging)**	**Peptide shift**
33,283,941	A		G	G	<0.1	*Zfp421*	possibly	Y/C
33,781,645	T	C	C	C	<0.1	*Pram1*	probably	L/P
33,956,791	T		C		<0.1	*Kank3*	possibly	S/P
34,069,285	C	T	T	T	0.928	*Rgl2*	probably	H/Y
34,112,420	T	C	C	C	<0.1	*CR974462.5*	probably	H/R
34,114,833	C	-	-		<0.1	*CR974462.5*	possibly	G/R
34,119,278	G	A			<0.1	*AA388235*	possibly	R/H
34,119,383	G	A			<0.1	*AA388235*	possibly	G/D
34,119,473	T	C			0.337	*AA388235*	possibly	F/S
34,134,481	T	C		C	<0.1	*H2-K1*	possibly	H/R

Regulatory polymorphisms

Differences between the susceptible strains and C57BL/6 were aligned to the Ensembl mouse regulatory build (NCBI37: Ensembl 54). Ten differences were predicted to fall within regions of accessible chromatin and may affect transcription factor binding regions. Furthermore, 13 differences mapped to within 2500bp of the upstream region of genes that may be associated with promoter regions. In total, 14 genes may be affected by SNP in this way (Table 2.5). Of the 13 genes for which microarray data were available, only phosphodiesterase 9A (*Pde9a*) showed any differences in gene expression, and these correlated with alleles of a SNP (rs33223038). A/J differed from C57BL/6 and BALB/c at this locus in both SNP genotype and *Pde9a* expression, but since this did not correlate with phenotype, it was discounted as a candidate SNP.

Validation of predicted nsSNP in Public Data

SNP in *Tir2* and *Tir3* that were not resequenced by 454 pyrosequencing, and were predicted to be functional and have damaging effects upon the protein by Polyphen, were validated using PCR and Sanger sequencing. In the most part, SNP had been discovered on either Celera data [147], for which only A/J data are available (from the strains used in this study), or on Perlegen (Affymetrix) data. The Perlegen dataset, which covers fifteen common strains of inbred mouse, has been predicted to be inaccurate and incomplete due to, in part, the low coverage generated, and partly because the algorithm by which SNP were detected was designed to focus on a low 'false-positive' rate [148]. As such, it is possible that the SNP actually reside in areas of repetitive DNA or high copy number regions, which could be interpreted as a SNP by base calling algorithms. Table 2.6 displays the loci that were resequenced in this manner. Blanks in the table represent a failure to basecall the sequences due to the presence of multiple sequences within the sequence "trace", indicating that the SNP was probably miscalled due to multiple copies of the gene in question. nsSNP that correlated with phenotype were confirmed for rs13468876, which has a single base insertion rather than the T>A substitution in the public dataset. Further nsSNP that correlated with phenotype were confirmed for rs45643169, ENSMUSSNP3206521, rs50073880, rs51259593 and ENSMUSSNP3208701.

Table 2.5: Putative regulatory SNP within an extended definition of *Tirl*. A list of SNP at *Tirl* (Mmu17: 33,271,855bp–34,203,529bp) matching regions of accessible chromatin according to the Ensembl murine functional genomics database. SNP matched identically in all four strains of susceptible mice after mapping against the resistant C57BL/6 reference with at least 50% of the reads agreeing at a given consensus position. Differences are either within 2.5Kbp 'upstream' in possible promoter binding regions; or are in accessible chromatin regions within coding regions of genes that may be associated with transcription factor binding sites.

Difference Position (bp)	Reference Allele	Alternative Allele	Data Type	GeneID	Region of Gene
31,011,574	G	A	H3K4me3:ESHyb	*AC165951.3-1*	Upstream
31,194,080	T	G	H3K4me3:ES	*Abcg1*	Upstream
31,194,238	G	C	H3K4me3:ES	*Abcg1*	Upstream
31,522,166	A	G	DNase1:ES	*Pde9a*	Upstream
31,979,290	T	C	DNase1:ES	*Snf1lk*	Upstream
35,669,957	C	A	DNase1:ES	*Psors1c2*	Upstream
35,771,262	T	C	DNase1:ES	*Dpcr1*	Upstream
35,772,620	T	C	H3K4me3:ES	*Dpcr1*	Upstream
35,790,848	T	G	DNase1:ES	*Vars2*	Upstream
35,790,864	T	G	DNase1:ES	*Vars2*	Upstream
35,835,340	T	C	H3K4me3:ES	*U6*	Upstream
31,046,212	G	A	DNase1:ES	*Glp1r*	Within
31,991,185	A	G	DNase1:ES	*Snf1lk*	Within
33,276,221	T	C	DNase1:ES	*Morc2b*	Within
35,692,027	T	C	H3K4me3:ESHyb	*Cdsn*	Within
35,692,042	T	C	H3K4me3:ESHyb	*Cdsn*	Within
35,692,339	T	C	H3K4me3:ESHyb	*Cdsn*	Within
35,705,160	T	C	DNase1:ES	*2300002M23Rik*	Within
35,773,710	T	C	DNase1:ES	*Dpcr1*	Within
35,819,115	A	T	DNase1:ES	*Ddr1*	Within
35,819,115	A	T	DNase1:ES	*Ddr1*	Within
36,326,297	A	G	H3K27me3:ES	*H2-T3*	Within

Table 2.6: Validation of publicly identified "potentially damaging" nsSNP at *Tir2* and *Tir3*. A list of nsSNP from public datasets with incomplete data across multiple susceptible strains of mice. Genotypes were confirmed by PCR and Sanger sequencing as described using primers in Appendix Table A2.1. Blanks represent a failure to basecall sequences due to multiple sequences within the "trace", indicating that the SNP was probably miscalled due to multiple copies of the gene in question. rs13468876 has a single base insertion rather than the T>A substitution in the public dataset.

SNP ID	Gene	Predicted Genotype	C57BL/6	A/J	BALB/c	C3H/HeJ
Confirmed SNP that correlate with phenotype						
Tir3						
rs13468876	*Apoa2*	T/A	T	GT	GT	GT
rs45643169	*Klhdc9*	T/C	T	C	C	C
ENSMUSSNP3206521	*AC083892.1*	A/G	A	G	G	G
rs50073880	*Slamf8*	A/G	A	G	G	G
rs51259593	*Darc*	G/A	G	A	A	A
ENSMUSSNP3208701	*E430029J22Rik*	G/T	G	T	T	T
SNP unconfirmed / do not correlate with phenotype						
Tir2						
rs13471968	*Srp72*	T/A	T	T	T	T
rs46908277	*Srp72*	C/G	C	G		
Tir3						
rs13469260	*Pla2g4a*	C/A	C			C
rs46572905	*Dusp12*	C/T	C	T	C	T
rs32565724	*Fcrlb*	T/C	T	C		
rs31938776	*Fcgr2b*	G/T	G	G	T	G
rs47563734	*AC083892.1*	G/T	G	G	G	G
rs31569041	*Itln1*	G/A	G	G	G	G
rs31537441	*Cd244*	T/C	T	C		
rs49701531	*Cd244*	G/A	G			
rs50777261	*Olfr418-ps1*	T/G	T			
ENSMUSSNP4607197	*BC094916*	G/A	G	A		G
rs45743507	*E030037K03Rik*	C/T	C	T		
ENSMUSSNP2076364	*E430029J22Rik*	C/G	C	G	C	C

Table 2.7: Physical locations of QTL and counts of candidate genes. Table summarising physical locations and counts of candidate genes after C3H/HeJ mapping and combined *Tir1* resequencing and public data annotation, which was produced in collaboration with Dr. Noyes. QTL positions were interpolated using NCBI37 from peak marker positions and 95% confidence intervals. The physical position of the D1Nds2 marker is not known, so its position was estimated from the intervals between its flanking markers. [a] Number of SNP common to the three susceptible strains of mice: A/J; BALBc/J and C3H/HeJ. [b]At Tir2-F12 we have estimated the physical 95% confidence interval around the D5mit58 peak marker and this 1.46Mbp region contained 27 genes, however the exact position of the peak is hard to identify since both D5mit58 and Dmit258 are at 41cM in the MGI map whilst they are 7Mbp apart on the physical map. Numbers of candidate genes were calculated under two hypotheses: Hypothesis 1: all four susceptible strains have the same haplotype as each other and different from C57BL/6; Hypothesis 2: All susceptible strains have a different haplotype from C57BL/6 but not necessarily the same as each other. NB. Hypothesis 1 is a special case of hypothesis 2 and all genes included under hypothesis 1 are also included under hypothesis 2. Only A/J and BALB/c are known to carry susceptibility alleles at *Tir2* and so at this locus only the correlation of C57BL/6, A/J and BALB/c was considered. [e] nsSNP loci submitted to dbSNP.

QTL	*Tir1*	*Tir2*-F6	*Tir2*-F12b	*Tir3a*-F6	*Tir3a*-F12	*Tir3b*-F6	*Tir3b*-F12	*Tir3c*-F6	*Tir3c*-F12
Chromosome	17	5	5	1	1	1	1	1	1
Peak marker	D17Mit16	D5Mit114	D5Mit58	D1Nds2a	DiMit286	D1Mit102	D1Mit102-DiMit105	D1Mit113	D1Mit107-DiMit16
95%CI (cM)	0.9	12	1	1.8	6	10	7	8	2
Median Mbp per cM	1.04	1.77	1.46	3.9	1.93	5.49	1.92	0.69	1.16
Start (Mbp)	33.27	71.02	73.45	100.54	124.71	121.63	148.15	170.96	164.3
End (Mbp)	34.2	92.3	73.91	107.57	136.19	176.56	161.44	176.51	166.6
Size (Mbp)	0.93	21.25	1.46	7.03	11.56	54.93	13.44	5.54	2.23
# Genes	43	210	27	20	127	650	113	143	35
Number of Candidate Genes (H1)	0	42	12	10	33	144	30	54	8
Number of Candidate Genes (H2)	27	74	14	10	63	355	61	122	8
454 Sequencing Data									
Common SNP (d)	194								
Common nsSNP	2								
Additional Data from Illumina Comparison									
nsSNP	0								
5'-UTR SNP	0								
Synonymous SNP	2								

Discussion

To fully characterise candidate genes at the *Tir* QTL and to enumerate all functional SNP contained therein, a number of additional analyses were conducted by colleagues within the team as presented in Goodhead *et al* (2010) [149]: The physical boundaries of the QTL had to be identified from the mapping distances (i.e. converting centimorgans to base pairs); the novel 454 SNP data had to be combined with public datasets such as from Perlegen [131] and the Mouse Genomes Project [135] and subsequently annotated for potential functional damage; and ancestral haplotypes were derived from the combined sets of SNP to assign genes to having a shared ancestral haplotype with the resistant strain and thus remove them from lists of potential candidates. Full details of these analyses are available in Appendix III: Additional Analyses.

The survival time phenotype for mapping murine QTL associated with response to *T. congolense* infection was selected in the 1990's because the large variance between strains made it more likely that there would be QTL of large enough effect to be identifiable. This prediction proved correct [126], however survival is likely to have a remote and complex relationship with the underlying quantitative trait genes (QTG). Given that trypanosomiasis is a systemic blood stream infection and the remote relationship between survival and the underlying QTG it is almost impossible to prioritise candidate genes on the basis of known functions. Previous work has included the measurement of parasitaemia, anaemia and fifteen clinical chemistry phenotypes in inbred and congenic mice, in order to identify correlations between survival and other traits that might be related to gene function, however no such associations have been found [150]. Therefore, in this study, we have identified the allele carried at each QTL in an additional strain (C3H/HeJ), and used this mouse strain to identify whether *Tlr4* has a role in moderating survival after infection. Additionally next generation sequencing technology was utilised to capture and resequence the entire *Tir1* region in four strains of susceptible mouse, which was compared to the resistant C57BL/6 reference. In a subsequent chapter, we have included the influence of copy number variation (CNV) on gene expression through the course of disease, which will result in a comprehensive genome-wide study of the genetic aberrations that may be responsible for moderating murine survival after *T. congolense* infection.

Our objective for this study was to identify the SNP that were most likely to have an impact on function. These were considered to be nsSNP that altered the physical properties of the protein as judged by Polyphen analysis, SNP in essential splice sites and regulatory SNP that correlated with changes in expression. It should be emphasised, however, that many types of SNP can underlie QTL, for example the QTL SNP at the *Idd5* locus appears to be a synonymous SNP that gives rise to a splice variant [151]. This SNP would not have been identified as a high priority by this pipeline. Furthermore, although we have substantially complete sequence coverage of the *Tir1* locus, at other loci we have used the Perlegen dataset, which is estimated to be about 45% complete [131]. Therefore although the candidate QTL SNP presented here are the most likely given the available data and annotation, both SNP data and annotation are incomplete and other candidates may be discovered in the future.

QTL mapping

The mapping studies showed that C3H/HeJ mice carry susceptible alleles at the *Tir1* and *Tir3* loci. No QTL were observed at the *Tir2* locus. The *Tir1* locus as defined by previous fine mapping studies is just proximal to the major histocompatibility complex (MHC) (Table 2.7), and the conversion of genetic distances to physical positions presented here shows that *Tir1* includes three classical MHC class I H2K genes. However, previous studies have found no correlation between MHC haplotype and response to infection [106], consistent with the QTL gene not being a classical MHC molecule. The mapping population was also screened for an association between *Tlr4* and survival; no association was found. This observation implies that the presence or absence of a functional *Tlr4* gene has no effect on survival, but does not preclude the pathway from *Tlr4* to *Nfkb* (nuclear factor kappa-B) from responding to infection. *Tlr4* could still participate in the regulation of anaemia and parasitaemia, which are not correlated with survival [152].

Identification of physical boundaries of QTL

Whilst the exact assignment of physical boundaries to the QTL is not possible, different locations have been reported for the *Tir2* and *Tir3a-c* QTL in the F6 and the F12 AIL generations [128, 129]. Furthermore, mice congenic for the C57BL/6 *Tir* alleles on an

A/J background agreed with the F6 location for *Tir2*, but with the F12 location for *Tir3a* [150]. The estimates of numbers of candidate genes, based on position within the predicted boundaries of the QTL alone, were as follows: *Tir1* contained 43 genes; *Tir2* between 12 to 42 genes and *Tir3* from 275 to 813, depending upon whether the F6 or the F12 study is used (Table 2.7).

SNP in *Glo1* copy number variant region proximal to *Tir1*

A two- to three-fold increase in read coverage compared to the overall mean coverage was detected at the proximal end of the targeted resequencing region. This region correlated with a copy number variant region at the *Glo1* locus on chromosome 17 for A/J, C3H/HeJ and 129P3/J [146]. CNV may have an impact upon SNP calling, as repetitive sequences across a small area cannot be assembled easily by short-read technologies. In this manner, erroneous SNP may be detected that are, in fact, segmental copies with small differences. dbSNP build 128 contains 3,204 entries within the database that are associated with this region, and indeed, our resequencing detected 373 SNP loci within this region that are likely miscalled due to CNV.

Whilst next-generation sequencing technologies probably exacerbate this problem due to the difficulties with mapping short reads, the extremely deep read coverage that they offer represents another method by which localised copy-number variants can be detected, potentially at single-nucleotide resolution. Indeed, recent studies have coined studies of this nature CNV-sequencing (CNV-seq), and have used this method to detect copy number variants in humans [153] and yeast [154].

Identification of functional nsSNP

Resequencing of the QTL region on the Roche 454 platform to 15X coverage discovered 3,618 novel SNP loci that were deposited in dbSNP. Comparison with a resequencing project on the Illumina platform at the Wellcome Trust Sanger Institute to 22X coverage [135] showed 99.98% consistency in SNP calls even when no minimum coverage criterion was applied for calling a SNP (Appendix III: Additional Analyses). Both datasets contained large numbers of SNP called as heterozygotes with alternative allele frequencies between 25-80%. These loci from both data sets were

associated with significantly higher sequence coverage in our data indicating that the majority were likely to be due to mapping artefacts probably caused by copy number variants. The 454 data contained only 71% of the SNP discovered by the higher coverage Illumina data but both methods discovered the same set of nsSNP. The 454 data discovered an additional 3% of SNP that were not in the Illumina data. Utilising all SNP from the 454, Perlegen and Illumina data sets, three probably damaging nsSNP were identified in genes at the peak of the *Tir1* QTL that correlated perfectly with phenotype (Table 2.4). Two nsSNP were in *Pram1*, with the *Pram1*$^{537L/P}$ polymorphism being scored as probably damaging by Polyphen. The *Pram1*$^{103R/K}$ polymorphism was classed as benign by Polyphen, but lies within a proline rich domain (PRINTS: PR01217) that is involved in binding the "SH3 domain of hematopoietic progenitor kinase 1 (HPK-1)-interacting protein of 55 kDa (HIP-55)". This region is known to stimulate the activity of HPK-1 and c-Jun N-terminal kinase (JNK)" [155]. C57BL/6 appears to have the derived allele of *Pram1*$^{537L/P}$ since A/J, BALB/c and C3H/HeJ had the same allele as Hominidae and dogs.

Pram1 is almost exclusively expressed in myeloid cells [156] and specifically in granulocytes in terminal stages of differentiation [157] where it is induced by retinoic acid. It was thought that *Pram1* might be a negative regulator of neutrophil differentiation since it is repressed in acute myeloid leukaemia. The deletion of *Pram1*, however, has no effect on neutrophil differentiation and maturation but does disrupt reactive oxygen intermediate production and degranulation by neutrophils [158]. This may affect the early pro-inflammatory response to infection, or signalling downstream of TNFa, which has itself been shown to be differentially expressed between susceptible and resistant mice [159]. PRAM1 appears to have a key role in the inflammatory response, whose differential expression is involved in inflammatory responses such as asthma exaggeration [160], and may be associated with pathology as it has been shown to be differentially regulated in cases of Dengue-associated haemorrhagic fever and Dengue Shock Syndrome [161]. It also shares structural homology with ADAP adaptor proteins such as SLAP130 (SLP-76-associated protein, 130 kDa), which are essential signalling molecules in integrin activation. Integrins have an essential role leukocyte adhesion and subsequent inflammation and the ability to fight infections [162]. Since C57BL/6 tend to have a more inflammatory phenotype, it is possible that the

polymorphisms discovered lead to a gain of function with stronger binding to HIP55 leading to faster and more persistent ROI induction and a more inflammatory state.

The other probably damaging SNP at *Tir1* were *CR974462* and *Rgl2*. There is no annotation for CR974462. *Rgl2 (Rif)* is a small GTPase that is most highly expressed in macrophages and B cells and appears to be involved in *Ras* mediated signalling [163]. The *Rgl2*$^{147H\rightarrow Y}$ polymorphism could affect the *Ras* pathway that plays a key role in leukocyte activation and is therefore a plausible candidate gene. The Fas death domain-associated protein (*Daxx*) gene, which has been previously reported to contain a deletion of a single aspartate residue in susceptible mice [164], is also under the peak of *Tir1*. *Daxx* is within the MAPK pathway, which was found to respond to *T. congolense* infection in microarray data. A new Polyphen analysis of the aspartate deletion in *Daxx* indicates that this polymorphism will be benign in effect. The aspartate deletion is within a run of 11 aspartate residues and a region where 35/41 residues are acidic [164]. Therefore this polymorphism is probably less significant than the probably damaging ones reported here.

Regulatory polymorphisms could also cause the phenotypic difference: one SNP (rs33223038) was identified in Ensembl as being in a regulatory region upstream of *Pde9a* but although this SNP correlated with differential expression it did not correlate with survival differences between susceptible and resistant mouse breeds.

Haplotype block analysis

This strategy has been previously used to show a strong association between upstream haplotype differences and high confidence ($p<0.005$) differences in gene expression [165] and also short listed genes under QTL for differences in response to *Heligmosomoides bakeri* infection [166]. Candidate gene numbers were reduced by assigning genes to haplotype blocks under two hypotheses: That a haplotype block in a given region is derived from the same ancestor in all susceptible strains of mice tested, which is different to the resistant strain (hypothesis 1); or that, for a given region, that the haplotype blocks in susceptible mice are all different from the resistant strain, but not necessarily all derived from the same ancestor. By so doing, the numbers were reduced by about 76% (Hypothesis 1) and 45% (Hypothesis 2) from the 1193 genes that

were under the 95% confidence intervals of the QTL. There were 283 genes where A/J, BALB/c and C3H/HeJ had the same haplotype different from C57BL/6 and 651 genes where C57BL/6 differed from the other three. The large number of genes that had haplotypes that correlated with phenotype is mainly because: 1) C3H/HeJ, A/J and BALB/c are more similar to each other than to any other strain based on analysis of 673 genome wide SNP in 55 strains [167]; 2) we used the stringent criterion that a gene was included if any haplotype block between the two neighbouring genes correlated with phenotype; 3) The high positive predictive power of the method means that whilst it is probably very reliable for excluding loci where susceptible strains share a haplotype block with the C57BL/6 resistant strains, it assigns too many haplotype blocks to different alleles.

Conclusions and Further Work

The aim of this study was to utilise the available genomic technologies in order to substantially reduce the number of candidate genes within the murine *Tir* QTL. Utilising next-generation sequencing technology, and annotation of public databases of variation within the mouse genome, we have increased the density of SNP within this region and improved haplotype maps so as to assign candidate genes to a resistant or susceptible ancestral mouse line. In this manner, genes that did not have an ancestry that correlated with phenotype could be removed from the short-list of candidate genes. Similarly, all non-synonymous SNP that may be having an impact upon the function of genes within the *Tir* QTL could be detected. 1,632 common SNP were identified, including a non-synonymous SNP in *Pram1* (PML-RAR alpha-regulated adaptor molecule 1), which was the most plausible candidate QTL gene in *Tir1*. It should now be practical to test the function of the candidate genes identified and the associated causative polymorphisms to determine their role in response to infection with *T. congolense*.

QTL involved with resistance to other parasitic diseases overlap with the *Tir* QTL, raising the possibility that polymorphisms discovered here may be involved in the response to other parasites. *Leishmania* resistance 1 (*Lmr1*) [168], *Plasmodium chabaudi* resistance QTL 3 (*Char3*) [169] and *Heligmosomoides bakeri* nematode resistance 2 (*Hbnr2*)

[170] all overlap with *Tir1*. Similarly, the *Tir3c* QTL overlaps with a QTL for murine resistance to *Plasmodium berghei*-driven experimental cerebral malaria (*Berr1*) [171].

Thirteen genes around the peak of *Tir1* show conserved order and sequence homology to a ~311Kbp region of BTA7 (15,412,179:15,723,462bp) where there is a QTL in cattle that regulates the level of parasitaemia in cattle infections with *T. congolense* [104]. This region includes *Pram1,* which has a probably damaging mutation that correlates with phenotype in mice and was the most plausible candidate gene in *Tir1* and is therefore a candidate QTL gene in cattle as well. However since trypanotolerance QTL cover approximately 15% of the bovine genome it would be expected that at least one of the five murine QTL would coincide with a bovine QTL by chance (p=0.56).

By combining these SNP with publicly available data, annotating the subsequent lists and identifying genes within shared haplotypes amongst susceptible lines of mice, we have demonstrated how large QTL regions can be reduced to tractable short lists of candidate genes for functional analysis. Nevertheless, for this analysis to be truly 'comprehensive', genetic aberrations other than SNP must be studied. Copy number variation is now widely regarded as a very important source of genetic variation: Redon *et al* reported that over 12% of the human genome is affected by CNV; with no large stretches remaining unaffected [172]. A similar study in laboratory mice has revealed variations between experimental lab strains [146]. The potential impact of CNV upon candidate genes within the *Tir* QTL will be covered in the next chapter.

Chapter Three

The influence of copy number variation on candidate gene expression at *Trypanosoma Infection Response* QTL in mice

Abstract

Copy number variants (CNV) have been shown to constitute large proportions of mammalian genomes: Greater than 12% of the human genome shows variable copy number. CNV can alter gene expression by increasing or decreasing the number of coding sequences of the gene, or by affecting their regulatory elements, and have been shown to significantly overlap with QTL in a range of traits. We have used the microarray-based Complete Genomic Hybridisation (aCGH) to identify aberrant copy number shared between susceptible mouse breeds genome-wide. These have been correlated to previous gene expression assays that identified genes responding to infection. Genes with variations in copy number that are at the *Tir* loci and are known to respond to *T. congolense* infection may indicate plausible candidate genes. Whilst no CNV could be detected at *Tir1*, *Tir2* or*Tir3a-b*, a significant CNV was detected at *Tir3c*: A two- to four-fold reduction in C57BL/6 copy number relative to A/J, BALB/c and 129P3/J overlapped with *Cd244*: a surface antigen that binds CD48 on lymphocytes and is involved in NK:NK cell and NK:T cell interactions leading to NK- and T-cell proliferation.

We have shown in a previous chapter that a nearby gene, *Cd48*, has a non-synonymous SNP, and since CD48 and CD244 directly interact, it is possible that the QTL is a consequence of the combined effect of the probably damaging nsSNP in *Cd48* and the CNV in *Cd244*. It may be possible to subsequently test this by inserting an additional copy of *Cd244* into the C57BL/6 background, so that it had a similar gene dosage to the susceptible strains.

Introduction

In the previous chapter, we have sequenced one of the QTL regions, *Trypanosoma infection response 1* (*Tir1*), in four strains of mice to identify novel SNP and candidate genes that may be responsible for regulating survival in mice after infection with *Trypanosoma congolense* strain IL1180. These data was correlated with genome-wide resequencing projects that have been made publicly available. SNP and small insertions / deletions, however, do not account for all of the polymorphisms that can affect candidate genes at a QTL.

Copy number variants (CNV) range in size from >1Kbp and <2Mbp. They are predicted to have an impact upon the expression of genes in a number of ways: As the affected regions are often segmentally duplicated, the expression of the genes contained therein may be proportionally affected, (e.g. If there are multiple copies of a gene, then expression may be increased). Likewise, if regulatory elements (either positive or negative) are duplicated or deleted, then expression could be similarly affected. CNV have been predicted to constitute greater than 12% of the human genome [172]. CNV have also been identified in the mouse genome, which vary between inbred lines: Cutler *et al* identified 2,096 CNV across 42 inbred mouse strains [173-175]. CNV have significant overlap with a number of quantitative traits in mice: Cho and colleagues demonstrated that 12/21 QTL significantly overlapped with CNV for the seven traits studied, including QTL involved with immunity on mouse chromosomes 13 and 17 ($p<0.006$) [176].

Array-based comparative genomic hybridisation (aCGH) is a high-throughput assay used to quickly and accurately compare the relative fluorescence of experimental DNA probes against a target sequence. Targets can be specific regions of interest or, due to the increased density of probes on a single microarray, entire reference genomes.

The Agilent 244A platform has demonstrated the highest sensitivity of the oligonucleotide-based CGH platforms: The platform can detect differences using a single probe when using dye-flip replicates (repeating the experiment using the opposite dye and normalising the resulting data) [177]. The Agilent 244A mouse array contains probes for approximately 235,000 coding sequences with an overall median probe

spacing of approximately 7.8Kbp (Agilent Technologies website). As the distributed software (Agilent CGH analytics) requires at least three probes to display differences in probe fluorescence in order to identify a CNV, and the irregular probe distribution across the mouse genome (probes are biased towards coding regions) this results in the smallest CNV that the platform can detect being ~36Kbp [177]. In this manner, some CNV (>1Kbp and <36Kbp) may not be detected, however given the predicted size of the *Tir* QTL is between 1 – 21Mbp (depending upon the QTL in question, and the cross in which it was identified; Chapter 2; Table 2.7), the QTL that has been mapped to the highest resolution (*Tir1*) should contain at least 13 probes.

We have used array comparative genomic hybridisation (aCGH) to identify CNV in QTL regions that correlate with survival in the four mouse strains. We have also correlated CNV with existing gene expression data from three of the mouse strains [152] to identify CNV that putatively cause expression differences.

Materials and Methods

CNV identification

Array CGH was performed using the Agilent Mouse Genome CGH Microarray 244A platform by Catriona Rennie [178]. Genomic DNA was obtained from Jackson Laboratories (JAX) for the reference mouse strain C57BL/6 (JAX mouse stock number 000664), and for three test strains: BALB/cJ (#000651); 129P3/J (#000690) and A/J (#000646). Dye-flip replicates were carried out, and the data normalised as previously described [178].

Overlapping "aberrations" (significant differences in $\log_2$ fluorescence signal ratio) were grouped into CNVR (t-test analysis, $P \leq 0.05$, Overlap 0.9) by the Agilent CGH analytics software (v 4.0), using the ADM-2 algorithm (threshold 6.0), centralization (threshold 6.0, bin size 1) and Fuzzy Zero [179]. CGH array data have been submitted to the NCBI Gene Expression Omnibus database (GEO) [GEO: GSE9669].

Measurement of gene expression

Gene expression data were obtained for A/J, BALB/c and C57BL/6 mice before infection and at four time points post infection on Affymetrix 450_2 microarrays from a previous dataset [152]. All microarray data has been deposited at ArrayExpress under the accession number E-MEXP-1190. The expression data and plots like those presented here are also available for all genes on the microarrays from the authors' website [180].

Results

To assess the impact of copy number variation regions (CNVR) upon the expression of genes that may influence response to *T. congolense* infection we performed array-based comparative genomic hybridisation (aCGH) on the complete genome of three mouse strains: 129P3, A/J and BALB/c, relative to C57BL/6.

Signals were obtained for 235,389 60-mer oligonucleotide probes across the whole-genome array, which equates to one probe every 11.5Kbp, assuming equal spacing of probes equally along the length of the entire 2.7Gbp mouse genome (NCBI build 37; Mouse Genome Informatics, Jackson Laboratories). Assuming the necessity for three probes to show significant changes in log2 fluorescence in order for the CNV to be detected by the Agilent software, the minimum sized CNV detectable under these conditions is approximately 34.5Kbp, in line with previous estimates [177].

The expression of genes within CNVR in A/J, BALB/c and C57BL/6 mice over the course of infection was evaluated using a previously described dataset [152]. In this manner, CNV that alter gene expression in all susceptible and/or the resistant mouse breed, and for which expression is modulated throughout infection, can be highlighted as good candidates for being a QTL gene.

A CNV at *Tir3c* affects the expression of *Cd244* in susceptible breeds of mice relative to C57BL/6

One significant CNVR was detected close to the peak of *Tir3c* in the F6 population (D1Mit113: 173,734,611bp). A two to four-fold reduction in C57BL/6 copy number

relative to A/J, BALB/c and 129P3/J encompassed, or overlapped with, the coding sequences of *Itln1* (intelectin 1), *Cd244* (Natural Killer Cell Receptor 2B4), and *AC083892.19-1* and may affect the nearby *Ly9* (lymphocyte antigen 9) (173,441,746-173,499,029bp; 11 probes; p=0.0003; Figure 3.1A). There were expression differences in *Cd244* (Figure 2A), but not *Itln1* or *Ly9* [180], over the course of infection between resistant C57BL/6 and susceptible A/J and BALB/c. *AC083892.19-1* was not on the expression microarray. This CNV region has also been previously reported by Graubert *et al* [146] who showed that an additional susceptible strain, C3H/HeJ, carries the same variant as A/J and BALB/c.

Other genome-wide CNV

No common CNVR were detected within *Tir1* or *Tir2*. The CNVR that was previously reported to be the cause of differential expression of Glyoxalase 1 (*Glo1*) [181], and is 2.8Mbp from the peak of *Tir1*, was detected as a two to fourfold reduction in copy number for C57BL/6 and BALB/c relative to A/J and 129P3 (Chr17: 30,176,153bp – 30,650,413bp; 68 probes; p<0.001; Figure 3.1B). Since the CNVR did not correlate with phenotype, this polymorphism is unlikely to contribute to the difference in response to infection.

Genome-wide, one hundred and twenty-nine CNVR involving three or more probes were common to A/J, BALBc/J and 129P3/J. These encompassed a total of 317 genes, and ranged in size from 400bp to 6.4Mbp, although 96% were smaller than 1Mbp. Twelve CNVR, containing the complete coding sequences of genes and that had corresponding differences in gene expression, were common to all susceptible breeds of mice tested. A list of the genome-wide CNVR is shown in Table 3.1.

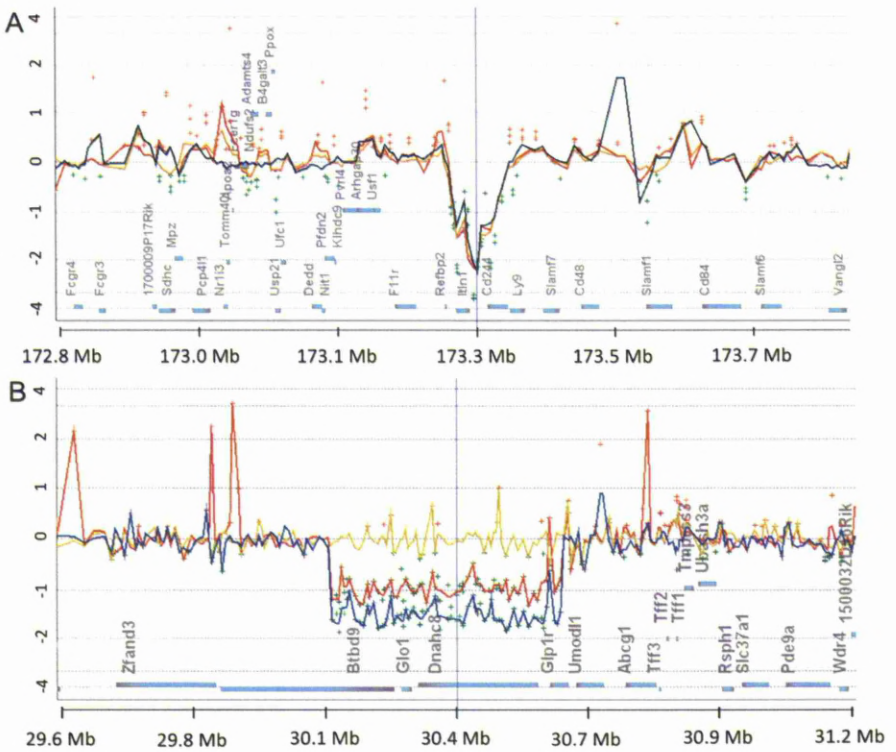


Figure 3.1: CNV plots from Agilent DNA Analytics software. **A:** Reduced copy numbers in C57BL/6 of *Itlnb* and *Cd244* near *Tir3c* relative to two susceptible breeds of mice (Chr 1: 172,831,532-173,931,532bp). **B:** CNV data at the proximal end of *Tir1* showing a deletion of *Glo1* and *Dnahc8* in C57BL/6 and BALB/c relative to A/J and 129P3. (Chr 17: 29,854,972bp–30,954,972bp). Probes are plotted at their genomic position relative to their respective $\log_2$ fluorescence intensity ratios (Y-axis) along with genes on the x-axis (filled blue rectangles). Green dots are negative ratios and red dots positive ratios (threshold 0.5). Lines are a moving average over a 10Kbp window for A/J (blue); 129P3 (red) and BALBc (yellow). Genomic positions are based on mouse build mm8 (NCBI36). In this manner, positive averages indicate an increase in copy number in C57BL/6 and negative values indicate a reduction in copy number in C57BL/6, relative to the test strain.

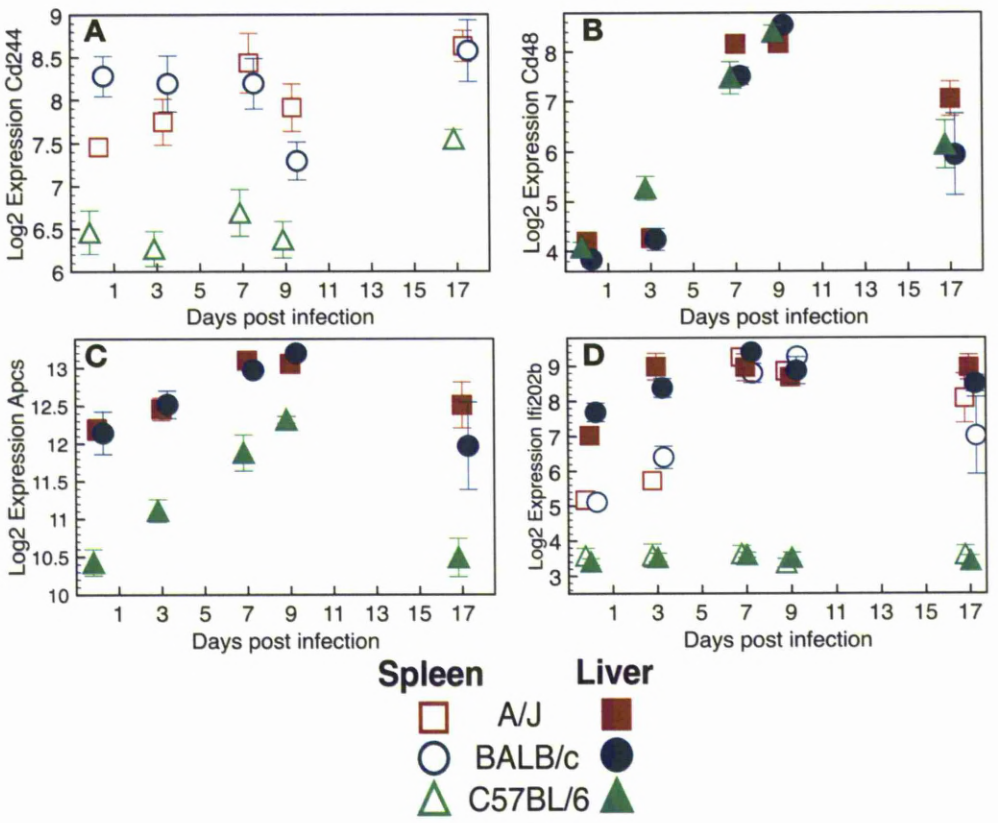


Figure 3.2: Expression of A/J OlaHsdnd (A/J), BALB/cJ OlaHsdce (BALB/c) and C57BL/6JOlaHSD (C57BL/6) mouse genes in the *Tir3c* locus at five time points in the course of *T. congolense* strain IL1180 infection (0 days; 3 days; 5days; 9 days; 17 days). Graphs include a small x-axis offset to improve spatial clarity. **A** *Cd244* in the spleen, **B** *Cd48* in the liver, **C** *Apcs* in the liver **D** *Ifi202b* in liver and spleen. *Cd244* expression was low in liver in all strains until Day 7 when it rose above background and C57BL/6 had slightly lower levels than A/J or BALB/c (data not shown).

Chromosome	CNV Start (bp)	CNVR End (bp)	# CNV Probes	Common Aberration P value	Score	CNV Location (Relative to gene)	Gene Symbol	Expression difference between breeds? (Tissue)	Expression responds to infection?
17 (*Tir1*)[6]	30,586,088	31,060,347	68	$<1.4x10^{-237}$**	-	Gene within CNVR	*Glo1*	Yes (L)	Yes
						Gene within CNVR	*Dnahc8*	Yes (S)	No
						Gene within CNVR	*Gpr1*	No	No
						Gene within CNVR	*AC165951.3-1*	n/a	
						Gene within CNVR	*AC125544.4*	n/a	
						Border	*Btbd9*	Yes	No
1 (*Tir3c*)	173,441,746	173,499,029	11	0.00029	-1104	Border	*Cd244*	Yes (L, S)	Yes
						Gene within CNVR	*AC083892.19-1*	n/a	
						Gene within CNVR	*Itlnb*	n/a	
3	142,269,450	142,286,796	4	0.002	-15.84	Border	*Gbp1*	Yes	Yes
4	62,157,766	62,182,695	8	0.005	-58.13	Gene within CNVR	*Alad*	Yes (L, S)	Yes (L)
						Gene within CNVR	*Hdhd3*	n/a	
4	111,725,815	113,560,896	106	0.004	11.7	Gene within CNVR	*9530098N22Rik*	n/a	
						Gene within CNVR	*A430090E18Rik*	n/a	
						Gene within CNVR	*A030013N09Rik*	n/a	
						Gene within CNVR	*Skint6*	n/a	
6	129,600,838	129,618,743	4	0.014	5.83	Border	*Klrc1*	No	No
						Border	*Klrc2*	No	No
6	129,690,146	129,733,410	8	0.025		Gene Within CNVR	*Gm156*	n/a	
						Gene Within CNVR	*Klri2*	n/a	
6	129,936,848	130,172,513	26	0.03	9.57	Complex	*Klra*	Yes (S)[5]	Yes[5]

Table 3.1: A list of significant CNVR in C57BL/6 (resistant) relative to A/J, BALB/c and 129P3 (susceptible) mice. (continued)

7	111,427,300	111,514,865	7	0.008	16.53	Border	*Trim34*	No	Yes
7	111,644,545	111,694,781	7	0.003	12.36	Border	*AI451617*	n/a	
						Gene within CNVR	*EG625321*	n/a	
14	69877096	70083115	32	0.002	32	Border	*Loxl2*	No	No
						Border	*Slc25a37*	n/a	
						Gene Within CNVR	*D930020E02Rik*	n/a	
						Gene Within CNVR	*Entpd4*	n/a	
X	166,413,387	166,422,251	3	6.81×10^{-5}	85.69	CNVR within gene	*Mid1*	Yes (S)	No
						CNVR within gene	*G530011O06Rik*	n/a	

Table 3.1: A list of significant CNVR in C57BL/6 (resistant) relative to A/J, BALB/c and 129P3 (susceptible) mice. Negative scores indicate deletions in C57BL/6 and positive scores amplifications, respectively. Genes predicted to be involved with chemosensory pathways have been removed due to having large local variations in copy number [182]. Aberrations were grouped into copy number variant regions (CNVR) using the Agilent CGH Analytics Software "Common Aberration test" (Overlap: 0.9; $p<0.01$) using the ADM-2 algorithm (Threshold: 6.0) with Centralization (Threshold: 6; Bin Size: 1) and Fuzzy Zero [179]. CNVR positions were converted to NCBI37 using Liftover (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Expression data for liver (L) and spleen (S) tissue was not available (n/a) for all genes. CNVR are listed by locus and ordered by chromosome, except for those at or near *Tir* loci, which are presented at the top of the list (associated *Tir* locus is presented in brackets). "Expression responds to infection" column indicates whether gene expression was observed to change across four time points post infection relative to prior to infection on Affymetrix 450_2 microarrays [138].

§: Amplification only occurs in 129/J and A/J mice.
****: P value too low for software to determine precise p value for A/J (reported as 'zero'). P value quoted for *Glo1* is for 129/J only.**
#: CNVR varies between breeds and only reported by algorithm in BALBc and A/J mice, although smaller amplification can be seen in 129/J (data not shown). P value and score reported is for CNVR based on A/J and BALB/c only.
¶: *Klra2*, *Klra3* and *Klra7* show reduced expression in C57BL/6 mice compared to susceptible breeds. All except *Klra5* and *Klra6* show differences in expression through infection. Expression data only available for *Klra 1-3; 5-9; 12,13,16* and *17*.

Discussion

CNV have previously been shown to be a major cause of quantative trait differences [176]. We used Agilent whole mouse genome aCGH arrays to identify CNV between C57BL/6 mice and A/J, BALB/c and 129P3 mice. The aCGH data highlighted a CNVR containing three genes close to the peak of the *Tir3c* QTL: *Cd244*; *Ly9* and *Itln1* (Figure 3.1A). A nearby gene *Cd48*, also had a probably damaging nsSNP (Chapter 2). *Cd244*, *Cd48* and *Ly9* are important genes involved in the production and regulation of IFNg by NK and T cells. CD244 binds CD48 on lymphocytes and is involved in NK:NK cell and NK:T cell interactions leading to NK and T cell proliferation [183], which are important mechanisms in innate resistance to protozoan infection [184, 185].

Spleen expression of *Cd244* differed between strains with the resistant C57BL/6 mice having the lowest expression consistent with the low copy number of *Cd244* in C57BL/6. *Cd48* expression increased 16-fold in liver after infection with *T. congolense,* but this occurred in all strains tested (Figure 3.2). Since CD48 and CD244 directly interact, it is possible that the QTL is a consequence of the combined effect of the probably damaging nsSNP in *Cd48* and the CNV in *Cd244*. Differences in expression could not be seen in *Itlnb* or *Ly9*.

Moving from candidate genes in mice to potentially useful candidate genes in cattle is difficult, as it is unlikely that a candidate gene for a mouse QTL would also be the QTG in cattle. More likely is that studies such as these highlight important pathways which may control response to the disease and may form the basis for tolerance, which in turn may help accelerate the discovery of candidate gene targets in cattle. A study into differences in gene expression between C57BL/6 and A/J mice in response to trypanosome infection highlighted the role of specific cytokines in the host-response to infection [186]. Recently, a similar approach of overlaying QTL mapping, DNA sequencing and gene-expression in bovine trypanosomiasis has identified two QTG candidates, TICAM1 and ARHGAP15: TICAM1 is a Toll-like receptor adaptor molecule, involved with the innate immune response; ARHGAP15 is involved with regulating natural killer (NK) cell response after infection, and as such, the NK response may be a shared pathway involved with both murine and cattle trypanotolerance [187].

Conclusions and Further Work

Whilst the large number of genes in *Tir3c* that have CNV, nsSNP or haplotypes that correlate with phenotype may make it difficult to identify the QTL gene at this locus, the CNV at *Cd244* was the most substantial DNA polymorphism in the region making *Cd244* a strong candidate QTL gene. It is also possible that the QTL is not a consequence of a single polymorphism but the combined effect of multiple polymorphisms in an extended haplotype, however Inserting an additional copy of *Cd244* into the C57BL/6 background, so that it had a similar gene dosage to the susceptible strains, could test the effect of this CNV on the response to infection.

Two other loci, outside of the *Tir* QTL may also have genes that respond to infection and have altered gene expression that may be due to CNV: *Alad* on chromosome 4 and the *Klra* genes on chromosome 6, which may warrant further investigation.

The killer cell lectin-like receptor genes *(Klr)* genes in particular show remarkable variability in copy number between mouse breeds. Previous studies have shown this in a wider range on mice [174], and have suggested that these CNV may affect the ability of the mouse breed to respond to infection. Indeed, work on the BALB/c mouse suggested that there might be additional members of the *Klr* family present that have not been probed by aCGH arrays and may be detected by further sequencing [188]. Furthermore, whilst the other genome-wide CNV do not overlap with trypanosomiasis QTL, they may be responsible for QTL involving resistance to other disease, and notably, survival after infection with the closely related *L. major*. QTL have been detected on chromosomes 6 and X that affect *Leishmania* lesion size [189]. In this manner, murine copy number studies may reveal further overlaps between CNV and disease QTL.

Whilst microarray-based studies have clearly paved the way for copy number research, it is clear that these technologies are being surpassed by faster, cheaper, higher-throughput, and more sensitive techniques: microarray-based gene expression assays are being replaced by transcriptome sequencing [190]; Chromatin immuno-precipitation (ChIP-chip) by next-generation ChIP-sequencing (ChIP-Seq) [191] and aCGH by CNV-seq, which uses the extremely high read depth generated by next-

generation sequencing technologies to identify duplicated regions [153]. Indeed, in chapter two, we have utilised next-generation sequencing to elucidate SNP that may be driving resistance to disease. It may be that more current methods will have the resolution and reproducibility to detect a wider range of polymorphisms that may alter the copy number of other genes or regulatory elements that may be driving differences in resistance to a range of diseases that it is not possible to detect with the techniques used in this study. Nevertheless, using established high-throughput microarray technologies, and combining them with gene expression assays, we have discovered CNV that correspond with genes that show a change in gene expression during *T. congolense* infection. In this manner, the CNV may be the polymorphism driving the QTL gene at the *Tir3c* locus: specifically affecting the expression of *Cd244* throughout infection.

By systematically combining aCGH and gene expression data from this chapter with next-generation DNA capture and sequencing and SNP annotation from the previous chapter, we have comprehensively analysed the genetic aberrations that underlie the *Trypanosoma infection response* QTL in mice and have generated a short list of polymorphisms in candidate QTL genes that can be functionally tested in cattle.

Chapter Four

Phenotypic and genetic analysis of *T. b. rhodesiense* field isolates reveals differences in virulence in mice that correlates with human disease

Abstract

The human-infective parasite *Trypanosoma brucei rhodesiense* generally causes an acute form of "sleeping sickness" across Eastern Africa, compared to *T. b. gambiense* infections in Western Africa, which are more chronic. The 1988 Ugandan *T. b. rhodesiense* outbreak constituted infections by parasites with different 'zymodemes': parasites that have variations in the electrophoretic mobility of a series of enzymes. The two predominant zymodemes, *Busoga* 17 (B17) and *Zambesi* 310 (Z310), each displayed differences in their clinical manifestation: Z310 infections were more chronic, and B17 more acute. Differences in survival phenotype could be replicated in experimental infections in mice: CD-1, BALB/c and 129/sv mice infected with Z310 survived for a significantly shorter length of time than those infected with B17.

In order to investigate whether *Tir1* regulates survival in *T. b. rhodesiense* infections in a similar manner to *T. congolense* (Chapter Two), mice congenic for the C57BL/6 allele (Tir1CC) at *Tir1* were infected with Z310 and B17 zymodeme *T. b. rhodesiense* parasites. Tir1CC mice did not show any significant difference in survival to A/J controls (Tir1AA), after infection with either Z310 or B17 zymodeme parasites. As *T. b. rhodesiense* infections are generally more acute than *T. congolense* infections, it may be that *Tir1* regulates long-term survival in mice after survival of the initial peak of parasitemia. Differences in survival were observed between zymodemes of the infecting parasite: Both TIR1CC and TIR1AA mice had a significantly shorter mean survival time when infected with B17 (~10.7 days) than those infected with Z310 (~15.6 days), in line with previous observations of human infections.

Cluster analysis of the microsatellite genotypes of 31 *T. b. rhodesiense* isolates that represented nine different zymodemes could not distinguish between Z310 and B17 parasite populations. STRUCTURE identified three population clusters, including a single cluster of Z366 parasites from a single 1993 outbreak in south-Eastern Uganda, a mixed population including Z310 and B17 isolates, and a single Z377 outlying individual. This suggests that either multiple genes control virulence, that there is gene flow between similar parasite populations, or that the microsatellite genotyping was insufficient to distinguish between different parasite populations. Further genetic analysis, utilising next generation whole-genome sequencing may be necessary to elucidate the loci responsible for the different virulence phenotypes between *T. b. rhodesiense* field isolates.

Introduction

Human African trypanosomiasis, or "Sleeping Sickness", is a vector-borne, parasitic disease caused by two subspecies of *Trypanosoma*: *T. brucei gambiense* and *T. brucei rhodesiense*. HAT patients exhibit early stage symptoms of fever and malaise through to later stage symptoms of confusion, reversal of sleep patterns and coma if the infected patient remains untreated. There have been many major epidemics in East Africa since 1896, with the latest epidemic in the Busoga region of Uganda running from 1971 to the present (Reviewed [192]).

Differential virulence phenotypes in *Trypanosoma brucei*

The clinical profiles of *T. brucei* infections vary depending on the subspecies of the infecting parasite: *T. b. gambiense* is classically defined as producing a chronic infection whereas *T. b. rhodesiense* infections tend to be very acute, with progression to late-stage disease often between 4-6 weeks [193], and 80% of deaths are within 6 months of the initial infection [194]. Differences in virulence phenotypes have not only been observed between subspecies but also been observed between isolates from the same subspecies. For instance, differences in pathogenicity have been observed in *T. b. rhodesiense* isolates from different Ugandan outbreaks [85] and for different isolates of *T. b. brucei* (strains TREU927/4 and STIB247) [195]. In the latter case, QTL underlying the observed differences in virulence in mice between *T. b. brucei* isolates have been established from artificial crosses of TREU927 and STIB247 [196].

Reports exist of differences in the severity of disease correlating with differences in parasite genotype [94, 197], which is not solely linked to parasitemia, as studies in *T. b. gambiense* have suggested that parasite isolates that generate elevated parasitemia do not necessarily result in increased pathogenicity [198]. Differences in *T. rhodesiense* morphology (linked to isolates from different disease foci) have been shown to correlate with their rate of growth in rodents [87], however it is likely that the effects of a combination of both the parasite and the host genotypes have an effect on the severity of the symptoms, and the speed of progression from early- to late-stage disease [94].

T. b. rhodesiense infections from the 1989–1993 Ugandan outbreak follow different clinical profiles

Three recent studies have observed similar differences in clinical presentation from allopatric samples (samples from different foci) from the Soroti and Tororo districts of Uganda, and from Malawi [85, 197, 199]. A study of 275 patients showed that isolates from Soroti presented with severe symptoms, a more rapid progression to stage 2 disease, and an earlier onset of neurological dysfunction. A greater percentage of Tororo patients had been infected for longer periods of time, and as such were more commonly associated with severe neuropathology. Samples from Malawi have been associated with a more chronic disease onset, with mild anaemia as the only common symptoms, and were rarely associated with chancre formation.

Whilst these observations have been made between spatially distinct foci in Uganda [85], few have shown similar results from sympatric isolates. One such study focussed on 42 isolates from Busoga, from an outbreak in Uganda from 1989-1993 [200]. Patients reporting to treatment centres suffering from *T. b. rhodesiense* sleeping sickness presented with two sets of symptoms: Patients, often from central villages, had short clinical histories with early symptoms such as fever and a chancre at the site of the tsetse bite; Those cases that had already progressed to late-stage disease had severe neurological symptoms. Other patients, who presented later due to a lack of initial symptoms and a more chronic disease onset, had symptoms that were more HIV/AIDS-like, including a lack of co-ordination and general malaise. These patients were often from areas close to the River Nile and Lake Victoria

Characterisation of *T. b. rhodesiense* by isoenzyme electrophoresis

At least two studies have attempted to classify *T. b. rhodesiense* according to similarities in their multilocus enzyme electrophoresis (MLEE) patterns. Initially zymodemes were grouped into two 'strain groups': *Busoga* and *Zambesi*, based on the regions from which they predominantly originated [88]. This is, however, less well represented in a more recent study using a reduced set of ten enzymes [201]. Whilst *Zambesi* zymodemes remained relatively unchanged, the *Busoga* group was split into two further groups: *Busoga* and the more *T. b. gambiense* - like *Bouaflé* group. Nevertheless, *Busoga* and *Zambesi* strain groups still share up to 75% similarity according to the latter study.

The clinical histories of Ugandan patients from the 1989-1993 Ugandan outbreak correlated with the zymodeme of the infecting parasite [200]: *Busoga* zymodeme infections caused acute infections in 93% of the tested cases, with 92% presenting with chancres. By contrast, *Zambesi* zymodeme infections occurred most frequently in those patients presenting with late-stage disease (Yates corrected Chi-Squared; $p = 0.001$) and without a chancre (Yates corrected Chi-Squared; $p < 0.01$), with the exception of the *Zambesi* 366 (Z366) zymodeme. Z366 infections were from a single, four-year outbreak in the Bugiri region that caused more *Busoga*-like symptoms.

Table 4.1: Enzyme banding patterns for six different zymodemes sampled in this study (from Stevens *et al* (1992)). Numbers represent a pattern of bands on a thin-layer starch gel, and as such, each pattern of numbers represents a 'barcode' from which a new zymodeme can be assigned. In this manner, B359 can be differentiated from B17 at the "NHD" (pattern 3 to pattern 1) and "SODb" (Pattern 8 to Pattern 9) loci. Enzyme abbreviations are as follows: NHI = Nucleoside hydrolase (utilising inosine); NHD = Nucleoside hydrolase (utilising deoxyinosine); TDH = Threonine dehydrogenase; ICD = Isocitrate dehydrogenase; MDH = Malate dehydrogenase; PGM = Phosphoglucomutase; ASAT = Aspartate aminotransferase; ALAT = Alanine aminotransferase; SOD = Superoxide dismutase. SOD production was deemed to be controlled by two genes and patterns attributed to either SOD_A (anodic group) or SOD_B (cathodic group). Details were not available for zymodemes Z375, Z377 or B376.

Zymodeme	NHI	NHD	TDH	ICD	MDH	PGM	ASAT	ALAT	SOD_A	SOD_B
B17	1	1	1	3	1	3	1	2	1	9
B359	1	3	1	3	1	3	1	2	1	8
Z309	1	3	1	1	1	1	1	2	1	7
Z310	1	3	1	1	1	1	1	2	1	8
Z311	1	3	1	1	1	1	1	2	1	9
Z366	1	3	1	1	1	3	1	2	1	9

Host response to experimental *T. b. rhodesiense* infection

Studies in experimental rodents have shown differences in morphology and growth rate between strains of *T. rhodesiense* (now reclassified as *T. b. rhodesiense*) from Botswana [202]. It was subsequently found that those strains that were from former human epidemic areas, and as such were deemed to be more infective in humans, developed more slowly in the rodent model. Likewise, those parasites that were from endemic regions, and deemed to cause a more chronic disease, appeared to grow quicker in number [203]. Similar behaviour was observed in samples from a more recent Ugandan outbreak, with clear differences in histopathology between zymodemes [200].

Different strains of laboratory mice exhibit different levels of resistance to infection with *T. b. rhodesiense* [204]. BALB/c mice survive for mean duration of approximately 20 days post-infection, whereas C57BL/6 mice survive for between 40-60 days. Similarly, C57BL/6 mice are relatively resistant to *T. congolense* infection, and A/J, BALB/c and 129/J mice are relatively susceptible [105-107]. This is similar to the situation seen in previous chapters with infections with *T. congolense*, where resistance had been linked to three quantitative trait loci (QTL), *Tir1*, *2* and *3a-c* (for *Trypanosoma Infection Response*), with *Tir1* having the largest effect upon survival [126]. The effects of these QTL on murine survival after *T. b. rhodesiense* infection have not hitherto been established. Resistance to *T. b. rhodesiense* infection in rodents has been shown to be both IFNG [205] and sex-dependent [206], albeit not X-linked [207].

Genetic variability of *T. b. rhodesiense*

Despite the phenotypic differences, little is known about the genetic variability within the *T. b. rhodesiense* subspecies; *T. brucei* genome sequences currently only exist for *T. b. brucei* strain TREU927/4 [42], and more recently, Type 1 *T. b. gambiense* [90]. Molecular characterisation has suggested that the *T. b. brucei* population is more heterogeneous than *T. b. rhodesiense*. Alongside this, similar studies have revealed that *T. b. rhodesiense* isolates from Uganda are more closely related to *T. b. brucei* isolates from the same geographical focus than *T. b. rhodesiense* isolates from Zambia [76]. Taken together, these data give rise to the hypothesis that *T. b. rhodesiense* is a host-range variant of the more genetically diverse *T. b. brucei*, and different epidemic foci arise from

T. b. brucei variants that have gained the SRA gene through mating followed by the selection for human resistance and subsequent expansion.

Aims and Objectives

The three *Tir* QTL have been established as regulating murine survival after infection with *T. congolense*, but have not been shown to affect mice infected with *T. b. rhodesiense*. In order to establish whether this is the case, we can take advantage of several mouse resources that are available to study the effect of a given locus on a complex phenotype, including the advanced intercross lines (AIL) described in chapter 2. Figure 4.1 shows a schematic of two such lines: consomic (4.1A) and congenic lines (4.2B). A consomic mouse strain is an inbred strain with one of its chromosomes replaced by the homologous chromosome of another inbred strain. Congenic mice are similar, except that they differ from a particular inbred strain at a single locus as a result of backcrossing whilst selecting for a particular allele at that given locus [208]. By using these mice, the difficulties of studying phenotypes that might be affected by loci outside of a QTL of interest can be reduced, albeit can be less reliable if two QTL are linked and interact. Mice congenic for multiple QTL can be used if this is the case, and have been used previously to study blood pressure QTL in rats [209].

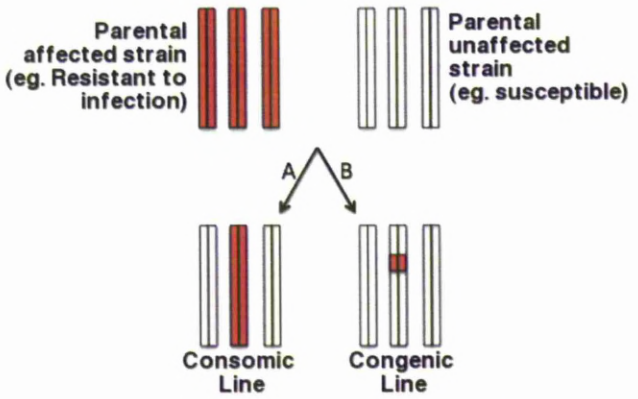


Figure 4.1: Diagram describing chromosome substitution strains and congenic strains of mice in relation to their parental strains. (A) A theoretical chromosome substitution strain with three chromosomes where one chromosome has been replaced with the chromosome of a resistant parent (or vice versa); (B) A similar theoretical strain that is congenic for a QTL on the same chromosome, which is created by genotyping offspring at the locus of interest and subsequently backcrossing to the unaffected line for several generations [208]. Strains such as these are important for the mapping and analysis of complex phenotypes.

A/J mice congenic for the resistant (C57BL/6) alleles at *Tir1* have been experimentally infected with one zymodeme from each strain group, and survival time monitored. In so doing, differences in pathology between different zymodemes can also be studied. These data have been compared to previous data for CD-1 mice, including parasitemia and cytokine response [210].

Additionally, in order to identify loci in the parasite genome that might be contributing to differences in virulence, we have analysed the multilocus genotypes of 31 isolates representing nine known zymodemes. Shared alleles between multiple isolates from the same zymodeme that differ between differentially virulent parasites may reveal those loci that underlie the observed differences in phenotype.

Materials and Methods

Trypanosome Stocks

31 of the *T. brucei rhodesiense* isolates used in this study (Table 4.2) have been previously described [200]. Zymodeme profiles discussed are according to Stevens and Godfrey [201].

Isolates stored as stabiliated blood were passaged through mice as follows: CD-1 mice (Charles River, Kent, UK) were infected intraperitoneally with 0.2mL stabilated blood. After 3 days post-infection, levels of parasitemia were assessed by tail bleed and those mice found to have a high level of parasitemia (>30 parasites per field; wet film) were immediately sacrificed and ex-sanguinated by cardiac puncture into 4mM EDTA as an anti-coagulant. DNA was extracted from mouse blood using a DNEasy Blood & Tissue Kit (Qiagen) as per the manufacturer's instructions.

In order to increase the amount of DNA available for PCR and subsequent storage, isolates that had previously been stored as stabilated procyclic cultures were additionally amplified using ϕ29-based Whole Genome Amplification (Illustra GenomiPhi V2 DNA Amplification Kit, GE Healthcare). As microsatellite allelic dropout has been reported for whole-genome amplified material [211], amplifications were performed in triplicate and reactions were pooled prior to subsequent use.

Table 4.2: *T. b. rhodesiense* isolates used in this study, including details of zymodeme, original storage conditions, and year of collection. Sample 32 was collected from a British tourist visiting Zambia in 2010 and is discussed in later chapters. All other isolates were collected as previously described [200]. Zymodeme profiles discussed are according to Stevens and Godfrey (1992).

Isolate Number	Zymodeme	Stage of Infection at time of collection	Year of Collection	Storage Conditions
1	Z375	Late	1993	Blood
2	B17	Early	1991	Blood
3	Z375	Early	1993	Blood
4	Z366	Early	1993	Blood
5	Z310	Late	1992	Blood
6	Z310	Late	1992	Procyclic
7	B17	Early	1991	Blood
8	Z366	Early	1993	Blood
9	Z309	Late	1993	Procyclic
10	B359	Late	1992	Procyclic
11	Z366	Early	1993	Procyclic
12	B17	Early	1990	Blood
13	Z366	Early	1993	Procyclic
14	Z366	Early	1993	Procyclic
15	Z366	Early	1993	Blood
16	B17	Early	1993	Blood
17	B17	Early	1991	Blood
18	Z375	Late	1993	Blood
19	Z366	Early	1993	Procyclic
20	Z311	Late	1991	Blood
21	B17	Early	1991	Blood
22	Z377	Late	1991	Blood
23	Z310	Early	1990	Blood
24	unknown	Early	1993	Blood
25	Z366	Early	1993	Blood
26	Z310	Late	1990	Blood
27	B359	unknown	1991	Blood
28	B17	Early	1991	Blood
29	Z375	Early	1993	Procyclic
30	Z310	Early	1990	Blood
31	B376	unknown	1991	Blood
32	unknown	Early	2010	Blood

Multilocus Microsatellite Genotyping

Full details of all primers used in this study are available in Appendix IV (Table A4.1), ten of which have been described elsewhere [85, 212] Microsatellite repeats on chromosome 8 were identified on the *Trypanosoma brucei brucei* TREU927 v.4 genome sequence [42, 61] utilising a Perl script as previously described [213]. PCR primers surrounding these sequences were designed using PRIMER3 [214].

PCRs were performed using: PCR buffer (45mM Tris-HCl, pH 8.8; 11mM $(NH_4)2SO_4$; 4.5mM $MgCl_2$; 6.7mM 2-mercaptoethanol; 4.4μM EDTA; 113μg/ml BSA; 1mM of each of 4 deoxyribonucleotide triphosphates), 1μM of each oligonucleotide primer and 0.5U of Taq polymerase (Thermo) was used per 10μL reaction; Alternatively, Reddymix (Thermo) was used for some PCRs. In both cases, 1μL of template DNA (20ng/μL) was used, except in the case of nested PCR, where 1μL of a 1/100 dilution of the first product was used in the subsequent nested reaction. The cycling conditions in every case were as follows: 95°C, 10secs; 50-55°C, 30secs (melting temperature (Tm) minus 5°C); 72°C, 10secs; 30x cycles. PCRs were resolved by ethidium bromide stained agarose-gel electrophoresis (Nusieve GTG, Cambrex, NJ) and visualised under UV-light.

Genotyping primers included a 5' fluorescent dye modification (FAM), which enabled accurate detection and sizing using a capillary-based sequencing instrument (ABI 3130 / ABI 3100; Applied Biosystems, Foster City, CA, USA) against a set of ROX-labelled proprietary size standards (GS-LIZ500; Applied Biosystems). Allele scores were generated using PeakScanner software (Applied Biosystems).

A bootstrapped dendrogram showing the relationship between the different *T. b. rhodesiense* multilocus genotypes was generated using an unweighted arithmetic average based on Jaccard's similarity index [215]. One marker that was found to be uninformative across all samples (M12C12) was removed prior to subsequent analysis. Bootstrap values were based on 100 replicates and those >70 are indicated on the dendrogram. The clustering calculator was accessed at http://www2.biology.ualberta.ca/jbrzusto. The number of population clusters was estimated using STRUCTURE [123], wherein the estimated number of populations

(K) was iteratively tested between one and ten; The most likely final value of "K" was taken as the lowest mean log likelihood score (LnP(D)) from twenty iterations. This was further tested using the Delta K analysis method [124]. The most likely number of populations was the highest value for the mean absolute rate of change between consecutive values for LnP(D) (across twenty iterations) divided by the standard deviation from the mean ($|\Delta LnP(D)|/SD$). As delta K analyses on hierarchical data structures tend towards the simplest number of populations, additional cluster analysis was performed using the BAPS package (v5.2; [125]), which determines the most likely number of clusters present in the data.

Survival in Tir1CC congenic mice

Tir1 has only been shown to mediate survival time in experimental infections with *T. congolense*. In order to investigate the effect of *Tir1* in *T. b. rhodesiense* infections, congenic mice containing the resistant (C57BL/6) allele at *Tir1*, on a susceptible A/J background (termed Tir1CC mice) and controls with the susceptible (A/J) allele (Tir1AA) were established by Susan Anderson (Roslin Institute, Edinburgh). In summary: C57BL/6 mice were crossed with A/J and at every generation after the F1 progeny, those mice that contain the C57BL/6 at *Tir1* were subsequently backcrossed to A/J for seven generations, resulting in <1% of additional C57BL/6 alleles elsewhere in the Tir1CC genome [150]. Colonies of the two congenic lines were established at the University of Liverpool and subsequently all offspring were checked for having the correct alleles by PCR and sequencing a known SNP between C57BL/6 and A/J (dbSNP ID: rs13465576) as previously described (Appendix Table A1.1.1).

24 age- and sex-matched congenic mice were infected with 10^4 B17 parasites and, similarly, 25 mice were infected with 10^4 Z310 parasites (i/p) from CD-1 donor mice. After positive parasitemia was established microscopically, mice were monitored until substantial symptoms were exhibited, at which point the mice were humanely sacrificed, and survival time noted. Differences in survival between mice of differing genotypes, sex and age at infection, and for those infected with different zymodemes of parasite were compared using Kaplan-Meier log-rank survival tests and linear regression (SPSS v.16).

Results

Experimental infections with different *T. b. rhodesiense* zymodemes in Tir1CC congenic mice suggests a complex survival phenotype between isolates that is not controlled by *Tir1*

Tir1CC mice, which are A/J mice with a C57BL/6-derived allele that confers increased survival time after infection with *T. congolense* parasites, showed no significant difference in survival after *T. b. rhodesiense* infections to their Tir1/AA controls regardless of the zymodeme of the infecting parasite (Kaplan-Meier survival test; $\chi^2=2.7$; df=1; p=0.09; Figure 4.2). Z310 infected mice, however, survived for a significantly longer period than those infected with B17 zymodeme parasites, with Z310-infected mice surviving for an average of 15.6 days versus B17-infected mice surviving for an average of just 10.7 days (Kaplan-Meier survival test; $\chi^2=16.1$; df=1; $p<0.001$). Table 4.3 shows a comparison of survival times for both congenic, and three breeds of inbred mouse. Figure 4.3 shows a survival curve, and associated boxplot for congenic mice infected with different *T. b. rhodesiense* zymodemes.

Table 4.3: Mean survival times (days ± standard error) for three common breeds of experimental inbred mouse (BALB/c; 129/sv and C57BL/6) and Tir1AA and Tir1CC (grouped under Tir1AA+CC) mice after infection with an isolate representing two different zymodemes of *T. b. rhodesiense*.

	Tir1AA+CC	BALB/c	129/sv	C57BL/6
Z310	15.6 ± 1.1 days	9.2 ± 0.4 days	15.4 ± 0.5 days	18 ± 0.7 days
B17	10.7 ± 0.3 days	16 ± 0.3 days	26.4 ± 0.7 days	29 ± 0.7 days

Whilst similar tests showed that, alone, the sex of the infected mouse did not have a significant effect upon survival, stepwise linear regression suggested a significant change in the goodness-of-fit statistic upon its inclusion (regression analysis; F-change=4.24; p=0.045). Figure 4.2b shows a box-plot of mouse survival similar to the inset box-plot in Figure 4.3, but with the results grouped by sex of the infected mouse. Whilst median survival does not appear to differ significantly, the variation in survival appears to be greater in female mice, with many more surviving for longer periods. Correspondingly, neither the age of the mouse at the point of infection, nor the mouse genotype significantly altered the regression statistic.

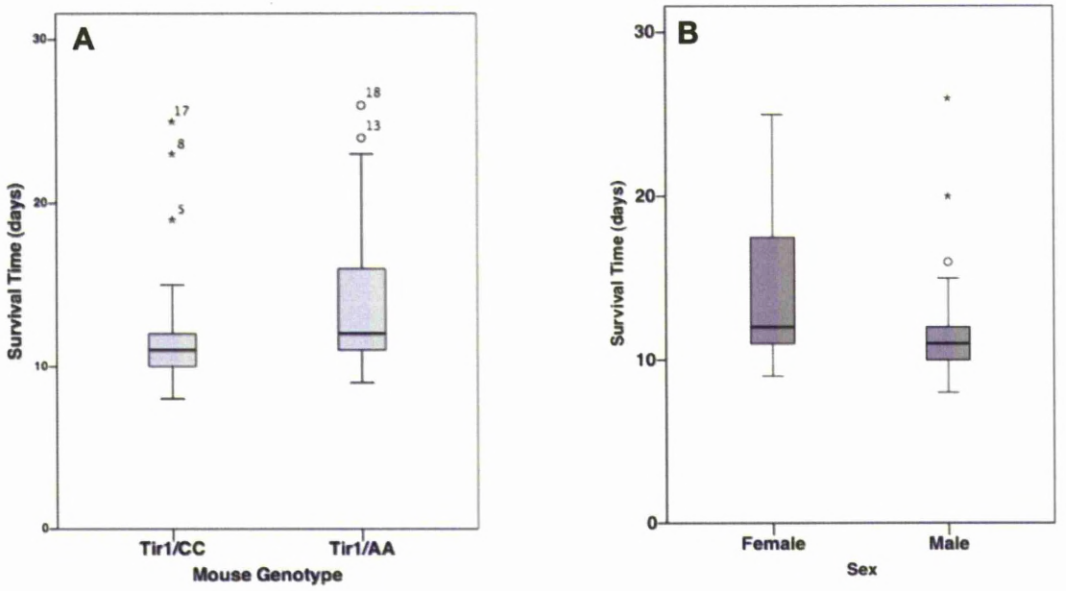


Figure 4.2: Boxplots of congenic mouse survival after infection with *T. b. rhodesiense Busoga* 17 and *Zambesi* 310 zymodemes, grouped by the genotype of infected mouse (A), or the sex of the infected mouse (B). Upper and lower limits of the box represent the upper and lower quartiles of survival, respectively. Median survival (days) is shown as the dark line towards the centre of the box. Error bars represent 95% confidence values (SPSS version 16).

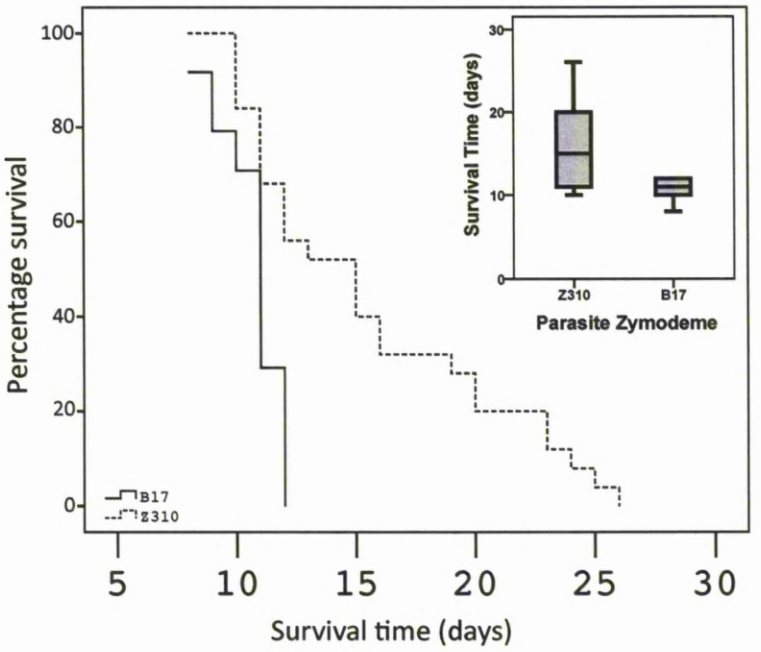


Figure 4.3: Kaplan-Meier survival curve of congenic mice (and controls) infected with Z310 and B17 zymodeme *T. b. rhodesiense* parasites. Inset boxplot represents median survival time (days) for mice after infection, grouped by zymodeme of the infecting parasite: Error bars are 95% confidence values; Upper and lower limits of boxes are the upper and lower quartiles, respectively; The dark line towards the centre represents median survival in days (SPSS version 16).

Analysis of the multilocus genotypes of 31 *T. b. rhodesiense* isolates was unable to distinguish between Z310 and B17 zymodemes

In order to examine the genetic similarity between *T. b. rhodesiense* strains originally grouped based on zymodeme, the multilocus genotypes of 31 isolates were determined by amplification at twelve microsatellite loci (Appendix IV). One locus was uninformative across the entire panel and removed from subsequent analysis. All except two isolates had been previously grouped into nine different zymodemes as previously described [200, 201].

Clustering the genotype data for the 1990-1993 samples using an unweighted pair group method with arithmetic mean (UPGMA) based on Jaccard's similarity index revealed three distinct groups of individuals (Figure 4.4): The single Z377 zymodeme isolate clustered separately, with a bootstrap value of 100 (based on 100 replicates). Two other clusters separated with a bootstrap value of 85: Firstly, a group containing Z366 isolates, together with a B376 isolate (bootstrap support of 73) and an additional isolate of unknown zymodeme. Secondly, a group containing a mixture of: Z309-Z311; Z375; B17 and B359 isolates and two isolates of unknown zymodeme. Within this cluster, all bootstrap values were less than five indicating that the branches therein were of low confidence. Population analysis using STRUCTURE (Figure 4.5A) revealed little difference between simulations of two or three populations, with similar log likelihoods between population estimates (K=2, -475.95; K=3, -476.91). Subsequent delta K analysis using the same data also suggested two populations was the most likely (Figure 4.5B). As delta K analyses tend towards the uppermost number of populations given hierarchical data structures, the data was additionally evaluated using Bayesian cluster analysis using BAPS, which suggested that three populations was the most likely by further discriminating the Z377 sample as an out-group. BAPS groups were largely identical to those shown in the UPGMA cluster analysis, except for a single isolate of unknown zymodeme (Isolate 24; Figure 4.5A). Repeating the cluster analysis without the Z366 data revealed no further substructure between Z310 and B17 populations. Full genotyping results are presented in a BAPS data format in Appendix IV (Table A.4.2).

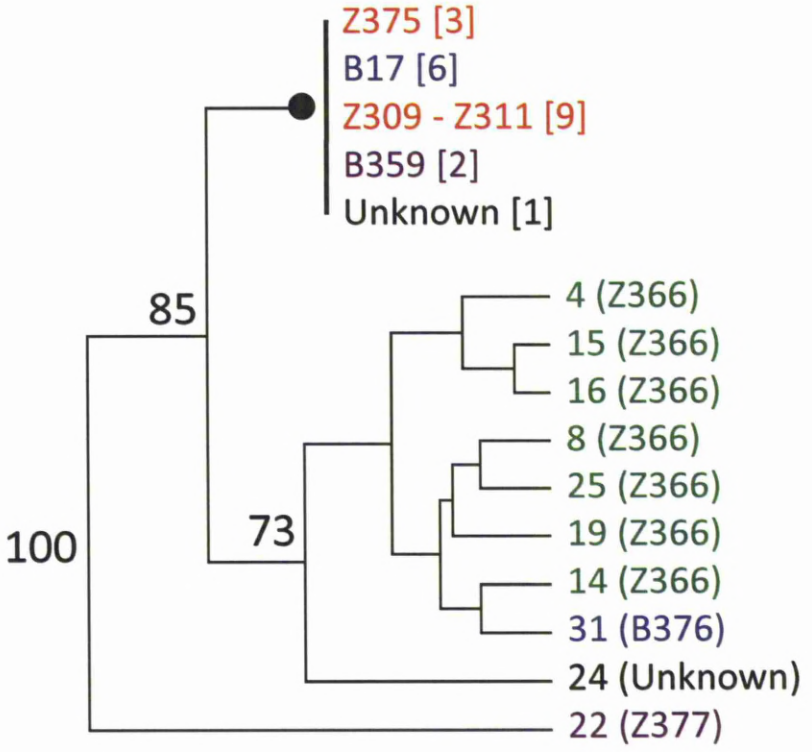
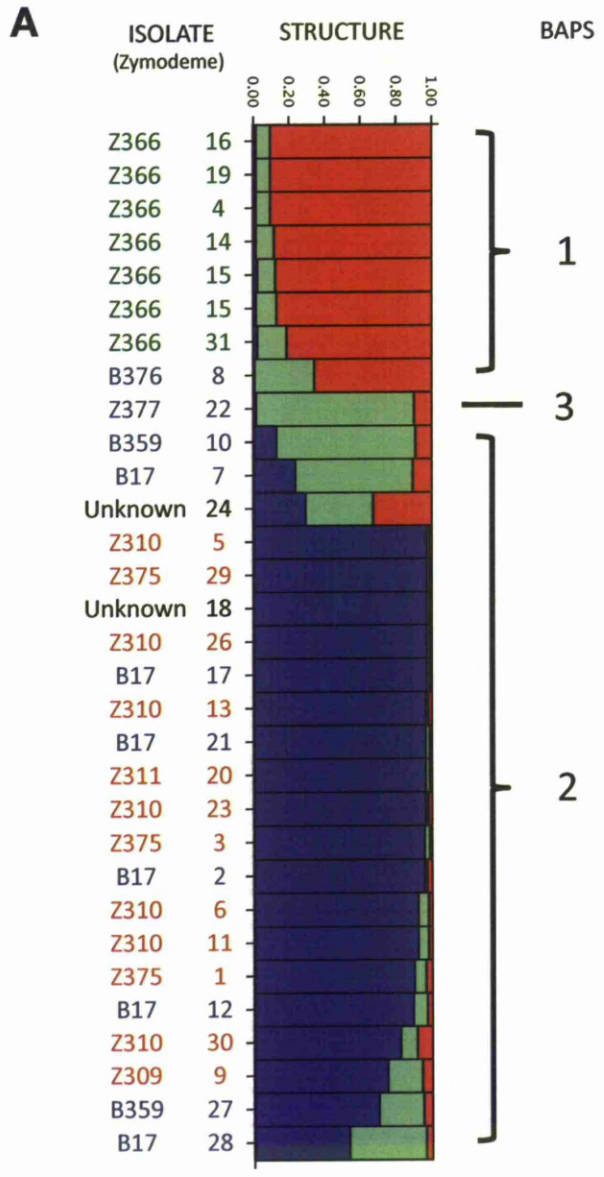


Figure 4.4: Dendrogram showing the relationship between 31 different *T. brucei rhodesiense* isolates at eleven informative microsatellite loci and their respective zymodemes (where known). Tree was generated using an unweighted arithmetic average (UPGMA) as the clustering method. Bootstrap values are based on 100 replicates and those >70 are indicated on the dendrogram. Sample numbers (Table 4.2) are displayed alongside zymodeme (if known) for nodes with high bootstrap support. Tree has been collapsed for the node representing B17, Z309-Z311, Z375 and B359 isolates due to low bootstrap support. In this case, the numbers of isolates associated with each zymodeme is shown in square brackets.



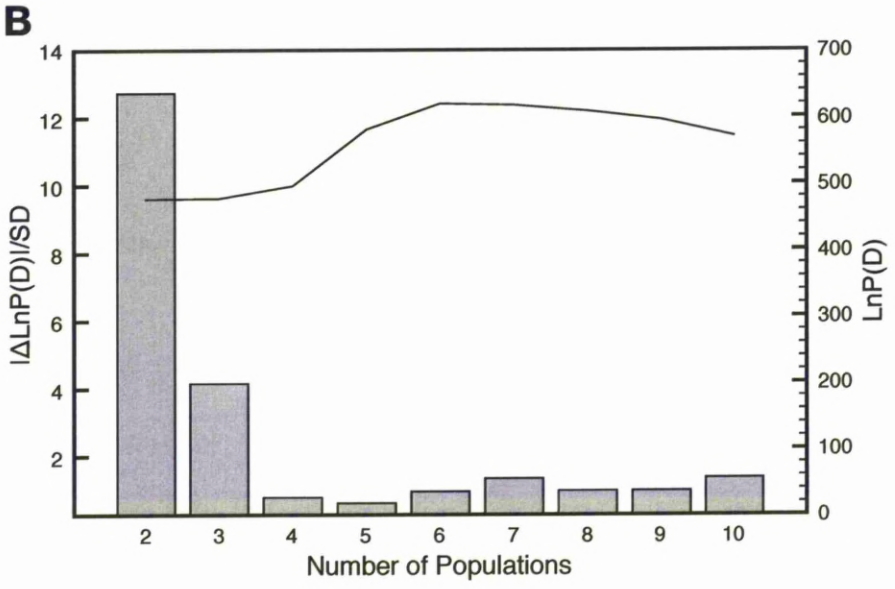


Figure 4.5A (left): Summary cluster analysis results from STRUCTURE and BAPS. Bar plot is a summary plot from STRUCTURE for K=3, indicating individuals broken into K coloured segments representing the estimated membership that each individual has within each of the K inferred clusters. Zymodeme (left) and BAPS cluster assignments (right) for each individual are described.

Figure 4.5B (above right): Delta K cluster analysis on STRUCTURE data. Line indicates absolute mean log likelihood scores for estimated population numbers. Most likely estimate for the number of populations is the value closest to zero. Bar chart indicates Delta K cluster analysis, where the most likely estimated number of populations from the data is indicated by the highest value of $|\Delta LnP(D)|/SD$ where $|\Delta LnP(D)|$ is the absolute rate of change between consecutive log probabilities for the estimated number of populations, and SD is the standard deviation from the mean.

Discussion

T. brucei rhodesiense infections produce both highly acute infections, rapidly progressing to late-stage disease, and more chronic infections lasting many months – often indistinguishable from *T. b. gambiense* infections. Two zymodemes from the Ugandan outbreak that began in 1989 – *Busoga* 17 (B17) and *Zambesi* 310 (Z310) are distinguishable by four isoenzymes, but tend to cause very different courses of infection in man. Z310 causes a more chronic infection and patients were often unaware of being infected due to a lack of a chancre at the site of a tsetse bite. Patients infected with these parasites often presented at clinics showing late-stage disease. B17 patients often presented earlier in the course of infection, especially as chancres were often present and patients learned to associate these with *T. b. rhodesiense* infections. Those patients that had been observed with late-stage B17 infections had progressed to this stage rapidly, with severe symptoms [200]. Whilst similar observations have been made for isolates from different foci of disease [197], this represents the only case of different isolates from the same focus showing differences in clinical manifestation.

T. b. rhodesiense virulence in humans correlates with survival in susceptible A/J mice

Despite their genetic similarity, differences in survival are consistently and reproducibly seen between Z310 and B17 parasite infections. No difference in overall survival (combined Z310/B17 infections) was detected between Tir1CC and Tir1AA congenic mice. Both Tir1CC and Tir1AA were more susceptible to the B17 isolate (Sample 11, Table 4.2) than the Z310 isolate (sample (Sample 26; Table 4.2) in line with human infections. For other wildtype mice, Z310-infected inbred strains of mice consistently had to be humanely sacrificed earlier than B17-infected mice due to substantial symptoms. Despite higher mortality in Z310, B17-infections still appear to have a higher level of parasitemia at the initial peak (Figure 4.3). Whilst the sex of the mouse did not have a significant effect on survival alone, including sex data in a stepwise linear regression analysis significantly changed the goodness of fit statistic; Female mice survive significantly longer than their male counterparts, agreeing with previous studies [206].

Multilocus microsatellite genotyping was unable to resolve populations of Z310 and B17 zymodemes

Z310 and B17 isolates are difficult to separate based on the markers used in this study alone. UPGMA, STRUCTURE and BAPS cluster analysis each organised both of these zymodemes into a single group, alongside Z309, Z311, Z375 and B359 zymodemes. Eleven out of twelve microsatellite loci were informative for at least one of the samples tested, although two loci (5L5; 401/1) only distinguished the Z377 outlying individual. Cluster analysis on these markers was able to distinguish a separate population containing all Z366 isolates, and a third group containing a single Z377 isolate. Z366 isolates correspond to a 4-year outbreak in the Bugiri region around 1993. Patients suffering from Z366 infections presented to treatment clinics with early-stage disease exhibiting a B17-like infection, often with a fever and a chancre at the site of a recent tsetse bite, and showed more intermediate symptoms in mice [200]. Indeed, isoenzyme analysis shows that Z366 is more closely related to B17 than Z310 (Table 4.1): Z366 differs from B17 only by the mobility of two enzymes as compared to four between Z310 and B17. The inability to distinguish between differentially virulent subtypes suggests that virulence is a multigenic trait and/or that there may be some gene flow occurring between similar parasite populations.

Despite MLMT being arguably a more powerful molecular technique to infer population structure, the data presented here suggests that MLMT (or the currently used loci) may not be the most suitable method for studying the population genetics of *T. brucei ssp.* as microsatellites designed at informative isoenzyme loci were unable to distinguish between parasites that cause severe- and less-severe disease (i.e.. Z310 and B17 parasites), relying on microsatellite data alone to characterise virulence is probably insufficient. Preferential amplification of a single microsatellite allele at a given heterozygous locus has been previously described [216], which may account for an inability to separate the strains. Interestingly, the microsatellite data presented (Appendix IV: Table A4.2) show that the Z366 isolates were distinguished by the presence of a unique heterozygous allele to that zymodeme at five loci (11/13; 407/1; Ch8_001; Ch8_002; 2/21). As such, further genetic analysis of isoenzyme loci may reveal differences that could not be detected on electropherograms.

Four isoenzymes are informative for distinguishing between *T. b. rhodesiense* zymodemes Z310 and B17 (Table 4.1), and only one, isocitrate dehydrogenase (ICD), appears to distinguish fully between *Busoga* from *Zambesi* groups, wherein a heterozygous band identifies *Busoga* strain groups, in the place of a homozygous band for *Zambesi*.

Using isoenzyme data to classify a species has been shown to be inconsistent with multilocus microsatellite genotyping (MLMT), for instance in the case of the classification of East African Visceral Leishmaniasis (VL). Originally, three species were identified by MLEE: *Leishmania donovani*, *L. infantum* and *L. archibaldi*, albeit this classification was based on polymorphisms in a single enzyme – Glutamate oxaloacetate transaminase (GOT) [217]. Subsequently, neither clusters of *L. archibaldi* nor *L. infantum* were detectable using microsatellite genotyping, suggesting that a single species – *L. donovani*, is responsible for VL in Sudan [218]. Whilst this is an extreme case, this emphasises the fact that it is perhaps not possible to directly compare the population structures suggested by MLEE and MLMT.

Conclusions and Further Work

Despite inbred laboratory mice exhibiting resistance to *T. b. rhodesiense* infection that differs between breeds, as measured by survival time after infection, this effect has been shown to not be under the influence of *Tir1*, as is observed in *T. congolense* infections (Chapters Two and Three).

Genetically similar field isolates of *T. b. rhodesiense*, as characterised by MLEE have exhibited different clinical profiles in man. Experimental infections in A/J mice congenic for the C57BL/6 *Tir1* alleles with two *T. b. rhodesiense* zymodemes (B17 and Z310), suggest that isolates that cause rapid and severe symptoms in man (zymodeme B17) cause hastened mortality in mice. Similarly, mice infected with samples from patients exhibiting a more chronic disease onset (zymodeme Z310) survive for a longer period of time.

Virulence in mice and rats had previously been shown to be inversely correlated to the phenotype observed in man: strains that cause more severe disease in man (are more

pathogenic) are less pathogenic in mice, as exhibited by longer survival and less severe symptoms [87]. Data have also suggested that virulence may be linked to cytokine levels and parasite numbers at the early peaks of parasitemia (Appendix V). In this manner it may be the case that for mice that are able to overcome this initial phase, mouse mortality is inversely correlated to humans, however very susceptible mice (such as A/J) are unable to survive this initial period.

Genotyping at eleven informative loci did not resolve populations of more pathogenic (B17) *T. b. rhodesiense* isolates from those that are less pathogenic in man (Z310). This implies that virulence is multigenic or that there is gene flow between parasite populations of different zymodemes. Whilst the host genotype has a major part to play in the overall phenotype of survival after infection and/or virulence, a more powerful study, utilising next generation sequencing technology, may be useful in studying the genetic diversity between *T. b. rhodesiense* field isolates. This will be covered in the next chapter.

Chapter Five

Epidemic *T. b. rhodesiense* strains have signatures of introgression with West-African trypanosomes that associates with altered virulence phenotypes

Abstract

Genomic analysis of East African *T. b. rhodesiense* and West African *T. b. gambiense*, has suggested that recombination is occurring between them. SNP genotyping of 32 *T. b. rhodesiense* isolates showed that differences in clinical phenotypes were associated with differences in alleles on chromosome 8. Genomic sequence of two isolates showed that chromosome 8 was heterozygous for alleles of West African origin in the more virulent strain, suggesting that recombination may be associated with parasite virulence. Combining SNP data with the observed patterns of heterozygosity has identified candidate genes that may underlie the observed differences in virulence. These parasite strains are from an outbreak that began in 1989 in Uganda, where both subspecies are found but thought to be in discrete geographical locations; however our data suggest that recombination has occurred at least once and that the human-infective subspecies of *T. brucei* are not genetically isolated. Our data have major implications for the control of the parasite, the spread of drug resistance and understanding the variation in virulence and the emergence of human infectivity. Further genetic analysis of *T. b. brucei* populations from Western, Central and Eastern Africa may be necessary to ascertain whether recombination is occurring directly between human-infective subspecies, or in the underlying animal-infective population.

Introduction

Hoare (1972) originally proposed the current three sub-species model of African trypanosomes on the basis of human-infectivity (i.e. both the ability to infect man, and the severity of the disease caused) and geographical location [3]; however problems with the classical assignment of the sub-species within the *T. brucei* sub-species complex are being revealed: Molecular analyses now suggest that the east African *T. b. rhodesiense* is a host-range variant of the more widespread *T. b. brucei* [219]; Similarly, west African human-infective trypanosomes have been split into two isoforms: Group 1 *T. b. gambiense* are clonal [220], more prevalent, less virulent in experimental rodents [95] and lack the serum resistance associated (SRA) gene, with serum resistance mediated by an invariant TLF-1 resistance mechanism [48]; Group 2 *T. b. gambiense* is more like *T. b. brucei* [99], more infective to experimental rodents [221] and whilst they also lack SRA, can lose serum resistance after serial passage [95].

The case that *T. b. gambiense* causes chronic disease, whilst *T. b. rhodesiense* causes acute disease is also under scrutiny. There is now evidence for both acute [222] and asymptomatic *T. b. gambiense* infections in Cote d'Ivoire [223]. Differential acuteness and severity also exists in *T. b. rhodesiense* infections throughout South-Eastern Africa, from asymptomatic carriers in Botswana [224], mild disease in Zambia [225] and Malawi [85], through to severe and acute disease in Uganda [200].

Genetics underlying virulence

Whether pathogenicity (i.e. the ability to cause disease) is a function of parasite or host genotype remains to be fully elucidated. Of particular interest, therefore, are the *T. b. rhodesiense* samples collected by J. Wendi Bailey between 1988 – 1993 in Southern Uganda [200], as these sympatric isolates show differences in disease severity within the same focus, removing the confounding factor of geographical location as seen in outbreaks described by Maclean (2007) [85]. The observed differences in virulence in man have been correlated to zymodeme strain groups: *Busoga* infections tended to be more acute with severe symptoms; *Zambesi* infections were relatively chronic, with patients often not recalling a chancre and presenting with late-stage disease. Earlier studies have reported that human disease correlates with differences in host survival in

experimental rodents and corresponds to differences in parasite morphology and on the location from which the sample originated; those parasites from previously epidemic areas developed more slowly in the rodent model, and vice versa [203].

The population structure of *T. b. rhodesiense* has been suggested to be either clonal [76], or panmictic with a few subtypes undergoing local and rapid clonal expansion [77]. Genetic analysis of Ugandan isolates has added weight to the argument of an underlying clonal population structure [226]. The recent addition of a combination of microsatellite genotyping and DNA sequencing of 142 *T. brucei* isolates from across Africa has suggested that recombination between *T. b. rhodesiense* and East African *T. b. brucei* is relatively common and further suggests that Type 1 *T. b. gambiense* is distinct from the other sub-species and Type 2 *T. b. gambiense* is more closely related to both *T. b. brucei* and *T. b. rhodesiense* [122].

Reviewing the dynamics of the trypanosome life-cycle

The complex life-cycle of trypanosomes involves multiple morphologically similar forms within both the mammalian host and the tsetse vector. Initially, the parasites rapidly proliferate as long-slender forms (slender), as which numbers generally double approximately every three hours *in vivo*. These parasites express, and continually change between forms of, variant surface glycoprotein (VSG), a surface coat by which they evade the adaptive immune response. This process also allows the parasites to maintain a chronic infection within the mammalian bloodstream as numbers generally dwindle as antibodies begin to recognise the currently expressed form of VSG whilst smaller numbers expressing a newer form continue to proliferate. This process of antigen switching results in the 'waves' of parasitemia that are characteristic of these infections (Chapter One: Figure 1.4).

Towards the peaks of parasitemia, subsets of these long-slender forms differentiate into an alternative bloodstream form – morphologically short and stumpy (stumpy) – that is intermediate to the procyclic, insect form. Stumpy forms exhibit a number of adaptations that ready the parasite for survival within the tsetse midgut, such as pH changes and a sharp decrease in glucose levels [227]. Notably, stumpy forms express EP (glu–pro repeat) procyclin, an alternative surface coat to VSG; the combination of

expressing procyclin and shedding unnecessary VSG into the host bloodstream is thought to increase the immune response against stumpy parasites as compared to the slender forms [39]. As the differentiation to the stumpy form is irreversible, a balance between different forms is important for maximising the likelihood of transmission [228]. Differentiation from slender to stumpy bloodstream forms is mediated by cell-density sensing [229], and the release of a hitherto unidentified stumpy induction factor (SIF). SIF is known to be of low molecular weight (<500Da) that is thought to signal via the cyclic AMP pathway [230].

Some experimental monomorphic clones, which are unable to differentiate to stumpy forms, have been shown to be more virulent in C3H mice due to the consumption of up to 40% of the mouse carbohydrate intake, as compared to carbohydrate consumption being less than 30% in pleomorphic clones [228]. This leads to a hypothesis that if Z310 parasites differentiate less readily in the mouse model, then a greater number of the slender form will be present in the bloodstream. This would impart a greater burden upon the mouse metabolism and could lead to increased mortality. Similarly, one could speculate that for those mice that are able to survive the initial peak of parasitemia, B17 infections would contain proportionally more stumpy forms, which would have a lower burden on the mouse metabolism and are potentially more easily targeted by the immune system, and as such, parasite numbers could be more easily controlled [231].

Aims and Objectives

Utilising next-generation sequencing technologies, we can now rapidly sequence representative isolates of differentially virulent *T. b. rhodesiense* and compare these sequences to the reference Kenyan *T. b. brucei* (TREU927/4), and to recently sequenced Type 1 (DAL972; [90]) and Type 2 (STIB386) *T. b. gambiense*, and other *T. b. rhodesiense* isolates (Unpublished data). In so doing, homology between chronic or virulent isolates may reveal genetic loci or candidate genes that underlie the observed differences in virulence. The addition of whole genome sequence data to recent microsatellite analyses [122] will shed light on the relationship between East African *T. b. brucei* and *T. b. rhodesiense* and other *T. brucei* isolates.

Proteins that are more abundant in slender forms may represent important factors involved in parasite growth, replication and metabolism, or may be involved in slender to stumpy differentiation. Similarly, as stumpy parasites are thought to be more immunogenic, proteins that are expressed to a greater degree in stumpy forms may be important virulence factors. Of interest, therefore, is the study by Jensen *et al* (2009), who used microarrays to analyse the relative mRNA abundance between stumpy and slender forms of *T. b. brucei* TREU927/4 [232]. Genes that are preferentially expressed at different stages of the life cycle that contain nsSNP between Z310 and B17 may represent interesting candidate genes for further study.

Materials and Methods

SOLID sequencing of Z310 and B17 isolates

Two isolates of *T. b. rhodesiense*, representing one each of zymodemes *Zambesi* (Z) 310 (Sample 26: Chapter 4; Table 4.2) and *Busoga* (B) 17 (Sample 11) were cultured as described previously [233], and summarised as follows: SDM-79 culture medium was kindly donated by Annette MacLeod, University of Glasgow. SDM-79 was supplemented with sterile foetal bovine serum to a concentration of 10% (v/v) and streptomycin (10mg/ml). Parasites were maintained in SDM-79 culture at 27°C in increasing quantities (2-5mL) until sufficient parasitaemia was established for DNA extraction and sequencing (~50ng genomic DNA).

DNA from cultured parasites was extracted using a Blood and Cell Culture DNA Kit (Qiagen, UK). Sequencing libraries were prepared and amplified by emulsion PCR according to the manufacturer's protocols (Life Technologies, USA). Whole genome sequencing was performed on a single slide using the ABI SOLID Analyser version 3 (Life Technologies, Foster City, USA). The resulting colour-space sequences were mapped to the *T. b. brucei* TREU927/4 v.4 genome sequence [42, 61]. Sequencing reads and associated coverage were visualised using the IGV browser (Broad Institute of MIT, USA). SNP were extracted using the BIOSCOPE pipeline (Life Technologies, USA) and deposited into a MySQL database using a bespoke Perl script (Appendix IX: Additional data file 3), wherein those associated with low coverage (<5X) were subsequently removed. The resulting filtered SNP were compared to generate lists of SNP shared by each isolate, and for unique homozygous and heterozygous SNP for each zymodeme.

SNP validation

33 SNP loci were validated using the PCR-based cleaved amplified polymorphic sequence (CAPS) method [234]. By choosing SNP that either create or destroy a restriction site for a given enzyme, SNP can be validated by comparing bands on an agarose gel to those predicted *in silico* in the absence of the SNP. Candidate loci were identified using the SNP2CAPS perl script [235] and Primer3 [214]. PCR products

were generated as per those for the microsatellite genotyping, and the resulting products were digested using the BSTNI enzyme (NEB). Restriction patterns were compared on a 2% agarose gel. For those restriction patterns that were unclear, additional validation was performed by directly sequencing the PCR products on an ABI-3130XL capillary sequencer using BigDye v3.1 chemistry (Applied Biosystems) after the excess nucleotides and primers were digested using a mixture of Exonuclease I (Thermo) and Shrimp Alkaline Phosphatase (Thermo) as per standard protocols [236]. All loci and primers are available in Appendix VI (Table A6.1).

Confirmation of Isocitrate dehydrogenase alleles

Two heterozygous, non-synonymous SNP that were 3bp apart within the chromosome 8 copy of isocitrate dehydrogenase (Tb927.8.3690) were confirmed by dideoxynucleotide sequencing as previously described. The theoretical isoelectric point and molecular weight for the Z310 and B17 copies of Tb927.8.3690 were predicted using the ExPASy compute pI/mw online tool (http://expasy.org/tools/pi_tool.html). Motifs within amino acid sequences were predicted using PROSITE [237].

Candidate gene identification

The numbers of heterozygous non-synonymous SNP (nsSNP) between the sequenced Z310 and B17 isolates were totalled by gene. A list of genes containing heterozygous nsSNP were compared to those identified to be differentially expressed between slender and stumpy forms of the parasite by Jensen *et al* (2009) [232] in a MySQL database, and shared genes identified. Additionally, as kinases represent potential drivers of differentiation from slender to stumpy parasites [238], Gene Ontology (GO) terms were downloaded into a MySQL database for all genes from GeneDB using the AmiGO tool (http://www.genedb.org/cgi-bin/amigo/go.cgi) and those genes that contain nsSNP, are differentially expressed between stumpy and slender forms and contain "kinase" GO terms were identified.

Selection of SNP loci for KASPAR genotyping

50 non-synonymous SNP loci (25 homozygous and 25 heterozygous) were selected from a MySQL database of all SNP between Z310 and B17 for subsequent typing of the

remaining isolates. Loci were chosen to represent all eleven megabase chromosomes, and all were predicted to have an impact upon protein structure by way of a low (<0) BLOSUM50 score. BLOSUM scores are a prediction of the likelihood of a given substitution having an effect on function based on the frequency that similar substitutions are observed in a reference dataset [140]. For instance, two similarly charged residues or, alternatively, two polar residues, are more likely to be substituted for one another. In this manner, BLOSUM scores range from -4 to +4, with negative scores predicting a less frequent substitution, and positive scores being more frequent.

KASPAR genotyping

A 100bp window surrounding the SNP was extracted from a consensus sequence for the Z310 and B17 genomes using a bespoke perl script (Appendix IX: Additional data file 4) and submitted to KBiosciences (KBiosciences Ltd, Hoddesdon, UK) for SNP genotyping using their proprietary KASPAR platform (http://www.kbioscience.co.uk/). Loci are presented in Appendix VI (Table A6.2).

Of the 50 loci selected, 31 non-synonymous SNP loci between Z310 and B17 *T. b. rhodesiense* were successfully genotyped by KASPAR SNP genotyping according to the manufacturer's protocols (Appendix VII: Figure 7.1.2). SNP were genotyped across 31 *T. b. rhodesiense* samples collected from Uganda between 1989 – 1993 [86], and a single sample from a 2010 patient from Zambia (Chapter 4, Table 4.2). 28 of the Ugandan isolates had been successfully typed by MLEE at the University of Bristol as part of the original study.

SNP data for the publicly available *T. brucei ssp.* genomes were analysed alongside the SNP genotyped by KASPAR. The number of populations was estimated using the STRUCTURE package as previously described (Chapter Four). A SPLITSTREE EqualAngle NeighbourNet phylogenetic network [239] was generated for all SNP using the default software settings.

Publicly Available Sequence data

Publicly available genome sequences for *T. brucei ssp.* were downloaded from the Sequence Read Archive (trace.ncbi.nlm.nih.gov) for: *T. brucei brucei* (TREU927/4; Accession number: ERX009953) and for the progeny of an artificial cross between *T. brucei brucei* STIB247 and TREU927 (ERX008996); a 1960 Ugandan *T. brucei rhodesiense* (EATRO3; ERX007603); a 1977 Kenyan *T. brucei rhodesiense* (EATRO2340; ERX007601) and for the progeny of an artificial cross between Type 2 *T. brucei gambiense* STIB386 and *T. brucei brucei* STIB247 (ERX000726). Similarly, sequencing reads for: *T. brucei brucei* (STIB247); a Type 1 *T. b. gambiense* isolate (DAL972) [90] and a Type 2 *T. brucei gambiense* from the Ivory Coast (STIB386) were downloaded directly from the Wellcome Trust Sanger Institute (WTSI) FTP website (ftp://ftp.sanger.ac.uk/pub/pathogens/Trypanosoma/brucei/T.b.gambiense_sequences/). All publicly available sequence data were generated on an Illumina Genetic Analyser (GA), except for DAL972, which was sequenced using dideoxynucleotide (Sanger) sequencing [90]. As Sanger sequence read-lengths exceed the maximum read-length permissible by the BOWTIE aligner, in order to align all data using the same alignment software, Sanger reads were artificially split into 50bp reads using a bespoke Perl script and treated as per next generation sequencing data (Appendix IX: Additional data file 5). As the original SOLID sequencing reads were aligned using BIOSCOPE, a comparison was performed to ascertain whether new mapping algorithms affected the SNP data and subsequent analysis. A comparison of mean coverage and SNP is presented in Appendix VI, Table A7.1.

The Illumina GA and artificial 50bp Sanger sequencing reads were aligned to the *T. b. brucei* TREU927/4 reference sequence using BOWTIE [240]. SNP were extracted using the PILEUP feature in the SAMtools package [241].

Genome-wide SNP analysis

123,543 genome-wide SNP were extracted for all six *T. brucei* genomes where both a polymorphism was present in one genome, and all genomes had coverage > 5. Due to the large number of VSG elements present in the T. b. brucei TREU927/4 reference sequence, all SNP within VSG coding sequences were removed, leaving 118,161 SNP. Differential non-synonymous SNP between Z310 and B17 were compared in a pairwise

fashion to both *T. b. brucei* (TREU927) and *T. b. gambiense* (Type 1, DAL972). SNP loci were colour coded (green = homozygous SNP; blue = heterozygous SNP) and plotted against genomic position to create a plot the introgression of alleles into *T. b. rhodesiense* from West African *T. b. gambiense*. A similar plot was created for a comparison of TREU927 and Type 2 *T. b. gambiense*, and was found to show similar patterns of introgression and is therefore not presented.

A bootstrapped (based on 1000 replicates) Jukes-Cantor Neighbour Joining (NJ) tree was created using SPLITSTREE. A third tree was constructed showing the distances between strains based on the 9,443 SNP loci on chromosome 8. Non-synonymous SNP were then extracted and split into individual chromosomes using a bespoke perl script similar to that described previously (Appendix IX: Additional data file 2). Additional SPLITSTREE NJ trees for each individual chromosome are presented in Appendix VII, including a comparison of the SNP predicted by both the BIOSCOPE and BOWTIE mapping algorithms (Figures A7.2.1 to A7.2.11).

Results

ABI SOLID sequencing reveals patterns of homozygous and heterozygous SNP between zymodemes

Alignment of the SOLID sequencing data for B17 and Z310 using BIOSCOPE revealed a total of 203,049 Z310 and 209,415 B17 raw SNP relative to the *T. b. brucei* TREU927/4 reference sequence. Removing shared SNP between the two samples revealed putative SNP loci between the samples as shown in Table 5.1. Plotting numbers of homozygous and heterozygous non-synonymous SNP between the sequenced *T. b. rhodesiense* strains and both *T. b. brucei* and *T. b. gambiense* (Type 1), respectively, revealed patterns of recombination across chromosomal regions (Figure 5.2).

Table 5.1: ABI SOLID sequencing results. Raw SNP between two individual isolates representing each of the *T. b. rhodesiense* Z310 and B17 zymodemes after mapping to the *T. b. brucei* TREU927/4 reference sequence using BIOSCOPE.

Zymodeme	Homozygous Non-synonymous	Heterozygous Non-synonymous	Total
Z310 (vs B17)	2,013	8,470	22,803
B17 (vs Z310)	2,464	9,061	29,169
Shared	55,178	32,429	180,246

The chromosome 8 copy of isocitrate dehydrogenase is responsible for differences in MLEE patterns between *Zambesi* and *Busoga* zymodeme strain groups

Isocitrate dehydrogenase (ICD) was identified as the only isoenzyme that differentiates between *Busoga* and *Zambesi* zymodeme strain groups of *T. b. rhodesiense* [89]. ABI SOLID sequencing revealed the presence of three heterozygous, non-synonymous SNP within the chromosome 8 copy of the ICD gene (Figure 5.1). Two SNP, that occurred within a 3bp window (Tb927.8.3690; Genomic coordinates: 21,622,832-4bp) were tested by capillary-based dideoxynucleotide sequencing and were found to occur in all four of the B17 isolates tested, and in none of the four Z310 isolates (Figure 5.2). Assuming that the SNP are linked, these changes confer a change in isoelectric point

(pI) and molecular weight of the protein sufficient for a change in mobility on a multilocus enzyme electrophoresis (thin-layer starch) gel [88] (Table 5.2). No non-synonymous SNP were detected in the chromosome 11 orthologue of ICD.

Figure 5.1: Predicted amino acid sequence of two sections of the chromosome 8 copy of the isocitrate dehydrogenase gene. Three heterozygous SNP were predicted in the Z310 sequence by SOLID sequencing and confirmed in four each of B17 and Z310 isolates by Sanger sequencing. Positions are relative to the initiation methionine. Polymorphisms are at amino acid positions 29 (F>L); 30 (D>G) and 435 (I>V).

Table 5.2: Predicted molecular weight and isoelectric point of isocitrate dehydrogenase (Tb927.8.3690) for B17 and Z310 SOLID sequenced isolates. Assuming that the SNP are linked, and as the SNP are heterozygous, the Z310 has another copy identical to the B17 copy.

Predictions:	**Isoelectric point**	**Molecular weight (kDa)**
Z310 (Heterozygous)	8.10	48.57
B17	7.66	48.67

SNP validation and refinement of SNP-calling criteria

33 genome-wide SNP identified between the Z310 and B17 SOLID sequenced isolates were chosen for CAPS-based validation. Of the 33 loci, 22 were confirmed and 12 unconfirmed. 5 of the loci that remained unconfirmed were predicted to be within the coding sequences of VSG elements, which may have caused problems with alignment due to the number contained within the genome. As such it was deemed that SNP at VSG sites could not be reliably predicted, and were removed from subsequent analyses. Results are presented in Appendix Table A6.1.

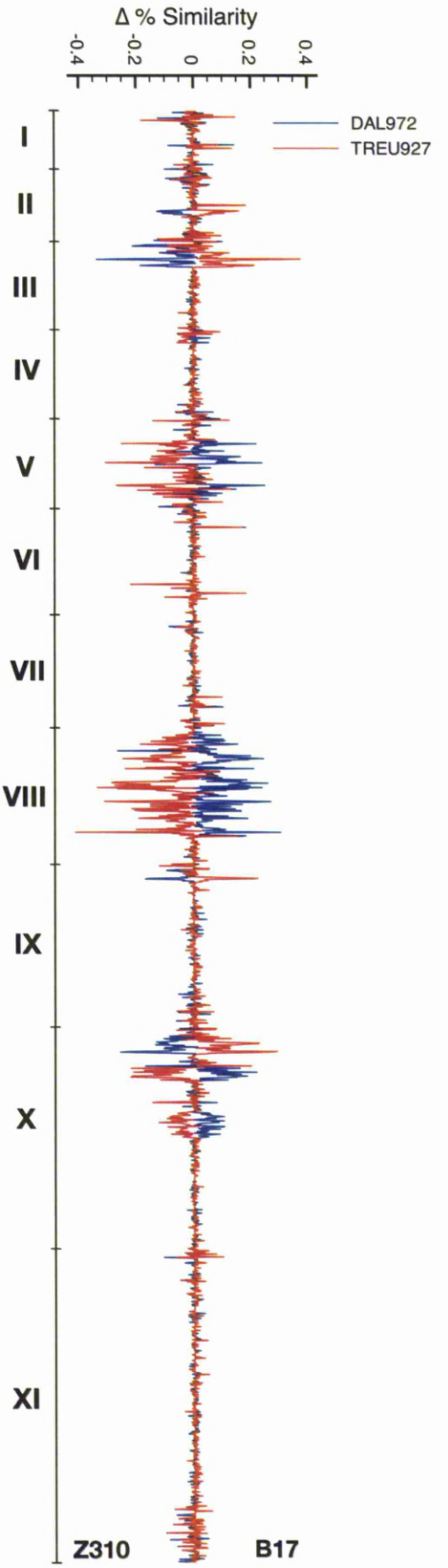


Figure 5.2: Difference in percentage similarity of two *T. b. rhodesiense* strains (zymodeme B17 and Z310) to *T. b. brucei* (TREU927; red) and Type 1 *T. b. gambiense* (DAL972; blue).

Percentage similarity to TREU927 and DAL972 was calculated for moving 10Kbp windows across B17 and Z310. The values of Z310 were subtracted from B17, giving "Delta percentage similarity" (Y-axis).

Negative values indicate Z310 is more similar to the given reference; Positive values indicate B17 is more similar to the given reference. Chromosomes are indicated by Roman numerals. In so doing, B17 chromosome 8 can be seen to be more similar to *T. b. gambiense* (blue), than Z310, which is correspondingly more similar to *T. b. brucei* (red). Other regions of introgression can be seen on chromosomes 3, 5 and 10.

Patterns of recombination between T. b. brucei, T. b. gambiense and T. b. rhodesiense differ between chromosomes

Comparing alleles from *T. b. rhodesiense* to both *T. b. brucei* TREU927 and to Type 1 *T. b. gambiense* (DAL972) showed patterns of homology and heterozygosity. ~73% of B17 chromosome 8 is heterozygous for alleles of both *T. b. brucei* and *T. b. gambiense* (Figure 5.2); Z310 is homozygous and similar to *T. b. brucei*. Overall sequence comparison of B17 and Z310 shows that the genomes of the sequenced Z310 and B17 isoaltes are >99.8% identical, however at SNP loci between the two, the B17 isolate is 23% more similar to Type 1 *T. b. gambiense* than Z310.

Other chromosomes show similar patterns of haplotypes of either *T. b. brucei* or *T. b. gambiense* origin: Z310 chromosome 5 appears to be heterozygous, in part, between *T. b. gambiense* and *T. b. brucei*. B17 is more similar to *T. b. brucei* at the same loci. Chromosomes 2, 3, 5, 8 and 10 showed such distinct heterozygosity patterns (Figure 5.3).

Examining SNP data on a SPLITSTREE phylogenetic network revealed patterns of historical recombination, represented by interweaving at the centre of the tree (Figure 5.4A). Z310 and B17 populations were close together, distinct from *T. b. brucei* and *T. b. gambiense*. Z366 isolates were tightly clustered, distinct from other *T. b. rhodesiense* zymodemes and situated between *T. b. brucei* and *T. b. gambiense*.

Expanding these data to shared SNP loci between all six next-generation sequenced *T. brucei ssp.* strains agrees with the historical classification of the species. A Jukes-Cantor Neighbour Joining (NJ) tree of the 118,161 genome-wide SNP loci shows *T. b. gambiense* clusters separately to both *T. b. brucei* and *T. b. rhodesiense* (Figure 5.4B). Different phylogenetic relationships can be seen on splitting these data into individual chromosomes: (Figure 5.4C) shows a similar Jukes-Cantor NJ tree wherein Z310 chromosome 8 is more closely related to the represented *T. b. brucei* strains, whereas B17 chromosome 8 is located between *T. b. brucei* and *T. b. gambiense*.

Subsequent comparison of the newer mapping algorithm – Bowtie [240], showed concordance with the BIOSCOPE SNP data that was in line with the difference in number of sequence reads successfully aligned (Appendix Table A7.1). Subsequent

analyses were not significantly affected by the differences in overall SNP number and as such the original BIOSCOPE results were used for Z310 and B17 sequences: SPLITSTREE NJ trees for each individual chromosome are presented in Appendix VII, including a comparison of the SNP predicted by both the BIOSCOPE and BOWTIE mapping algorithms showing little difference between the results from the different mapping algorithms (Appendix Figures A7.2.1 to A7.2.11).

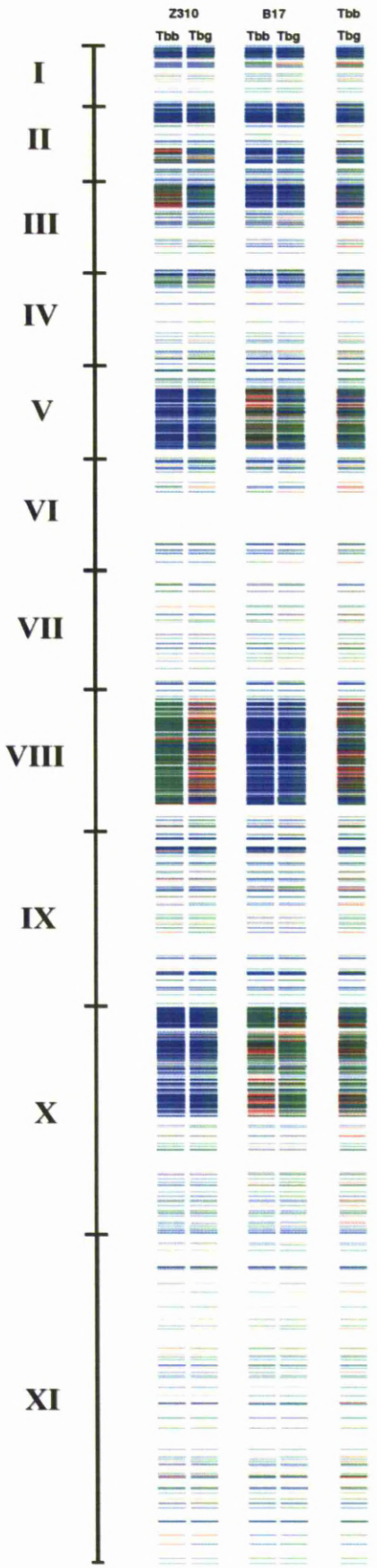


Figure 5.3: Introgression plot of *T. b. brucei* and *T. b. gambiense* (Type 1) alleles into two *T. b. rhodesiense* genomes (Z310 and B17). 123,543 genome-wide SNP were identified between six *T. b. rhodesiense* genome sequence data as described in the main text (Genome-wide SNP analysis).

Lines represent shared alleles between TREU927 (Tbb; left) or DAL972 (Tbg; right) with the two SOLID sequenced *T. b. rhodesiense* genomes, Z310 and B17. Introgressed alleles are represented as follows: Two shared alleles (i.e. homozygous; green); One shared allele (i.e. heterozygous; blue); No shared alleles (red). Polymorphisms are plotted alongside a comparison between TREU927 and DAL972 in the same manner. Chromosomes are represented in Roman numerals I – XI.

In this manner, Z310 chromosome 8 can be seen to be homozygously similar to *Tbb* and not *Tbg*, whereas B17 shares one allele each with *Tbb* and *Tbg* across much of the chromosome.

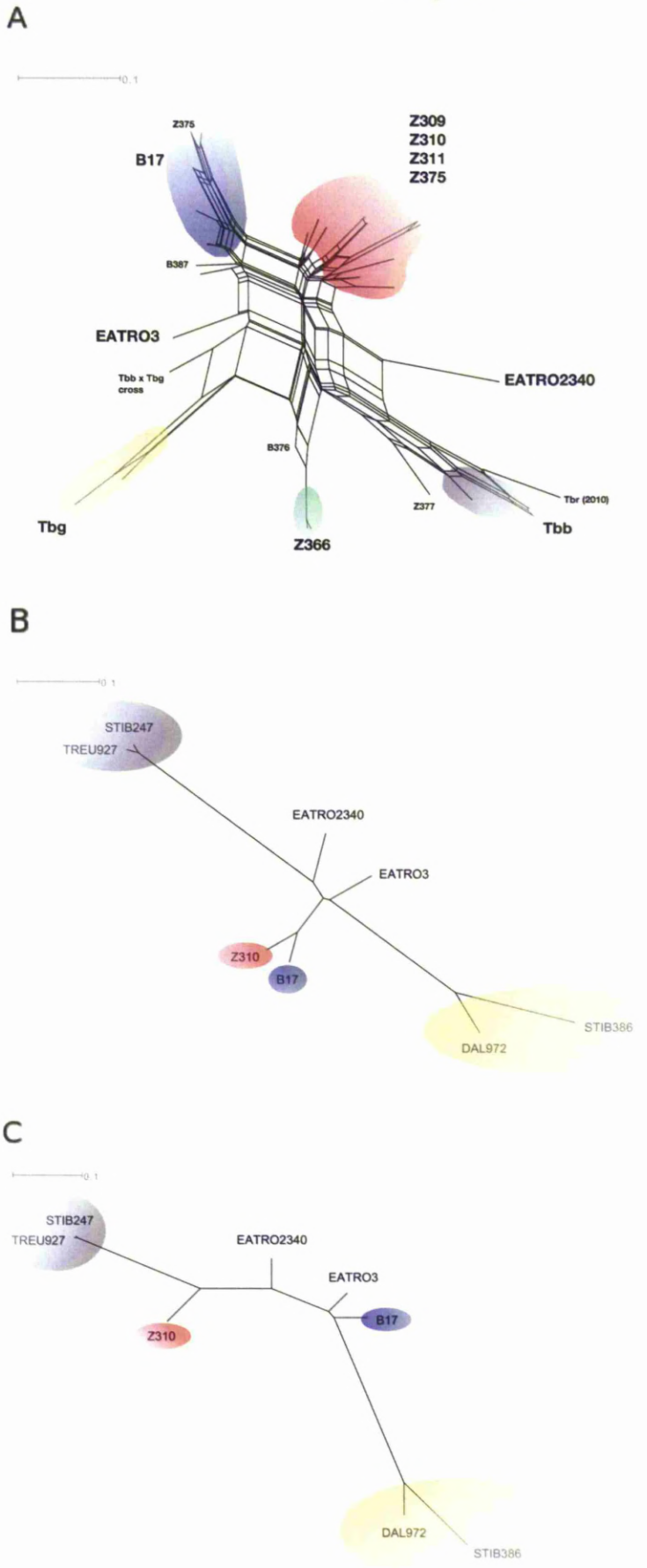


Figure 5.4: SPLITSTREE phylogenetic networks and trees:

A: EqualAngle phylogenetic network for KASPAR genotyped SNP loci and next-generation sequencing data at the same loci for 40 *T. brucei* isolates. Zymodemes and/or subspecies have been highlighted. Tbb = *T. b. brucei* strains TREU927, STIB247 and the TREU927 x STIB247 cross; |Tbg = *T. b. gambiense* Type 1 (DAL972) and Type 2 (STIB386); Tbr (2010) = Zambian *T. b. rhodesiense* isolate (Sample 32; Table 4.2); "Tbb x Tbg cross" = STIB247 x STIB386 cross.

B: EqualAngle Jukes-Cantor Neighbour-Joining tree of next-generation whole-genome sequencing data. Bootstrapped tree (based on 1000 replicates) represents 118,161 genome-wide SNP filtered for SNP within VSG coding sequences. All bootstraps > 90%.

C: As B, but limited to 9,443 SNP loci on chromosome 8. All bootstraps > 98%.

Colours symbolise groups of similar sub-species / zymodeme:
Blue: *T. b. rhodesiense* B17
Red: *T. b. rhodesiense* Z310
Green: *T. b. rhodesiense* Z366
Grey: *T. b. brucei*
Yellow: *T. b. gambiense*

Genes affected by differences in heterozygosity between B17 and Z310 zymodemes

Heterozygous non-synonymous SNP (nsSNP) between the sequenced B17 and Z310 isolates were totalled by chromosome (Figure 5.5) Chromosome 8 was the most affected (B17: 507 genes contained heterozygous nsSNP; Z310: 21 genes). (Figures 5.1 and 5.2). Similar differences are observed for chromosome 3, 5 and 10 in line with having the largest regions of differential heterozygosity between zymodemes.

Genes that contained heterozygous nsSNP were compared to genes that are differentially expressed between slender and stumpy forms of the parasite [232]. Those genes identified by Jensen *et al* (2009) that were both among the most differentially expressed between slender and stumpy forms, and that contain nsSNP between zymodemes, are shown in Tables 5.3A-D.

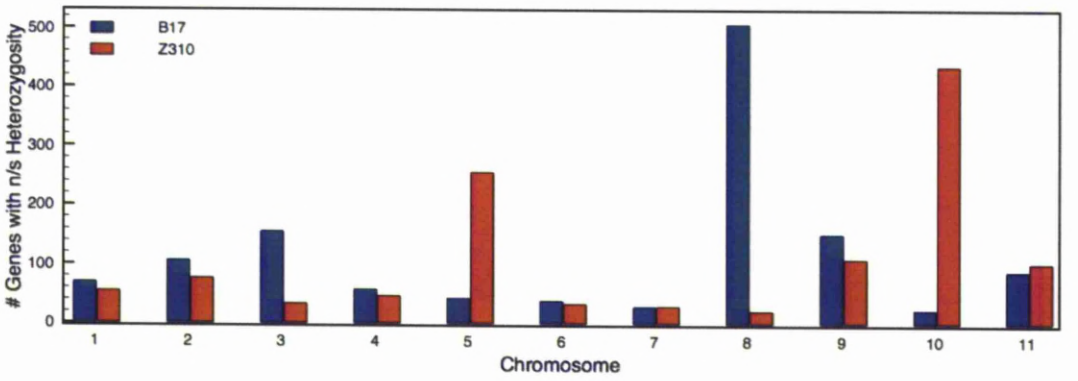


Figure 5.5: Genes on each chromosome in regions that are heterozygous in one zymodeme, and not in the other. Counts are the number of differential heterozygous, non-synonymous SNP between the sequenced B17 and Z310 isolates. Shared SNP loci vs. *T. b. brucei* TREU927 have been removed. Blue = B17; Red = Z310.

Table 5.3: Genes with non-synonymous heterozygous SNP that are amongst the most differentially expressed between slender and stumpy bloodstream isoforms of the parasite. Genes are ordered by the number of nsSNP [232].

Table 5.3A Z310: Slender

Chromosome	Gene	Function	SNP count
10	Tb10.70.7320	Hypothetical, conserved	18
5	Tb927.5.4010	Hypothetical	14
10	Tb10.70.4030	Hypothetical, conserved	11
10	Tb10.70.7280	Hypothetical, conserved	9
10	Tb10.70.1720	Dynein; Molecular motor activity	4
5	Tb927.5.1410	64kDa invariant surface glycoprotein	4
5	Tb927.5.1430	64kDa invariant surface glycoprotein	4
10	Tb10.70.4020	Hypothetical	2

Table 5.3B Z310: Stumpy

Chromosome	Gene	GO Term	SNP count
5	Tb927.5.4620	ESAG protein	11
9	Tb09.160.5430	ESAG9 associated protein	2
10	Tb10.70.2840	Hypothetical	2
6	Tb927.6.1520	Transporter activity	2
9	Tb09.142.0370	Hypothetical, conserved	1
9	Tb09.160.5400	ESAG9 associated protein	1

Table 5.3C B17: Slender

Chromosome	Gene	GO Term	SNP count
5	Tb927.5.1410	64kDa invariant surface glycoprotein	2
5	Tb927.5.1430	64kDa invariant surface glycoprotein	2
3	Tb927.3.1910	Hypothetical, conserved	1
3	Tb927.3.930	Dynein; Molecular motor activity	1

Table 5.3D B17: Stumpy

Chromosome	Gene	GO Term	SNP count
8	Tb927.8.1130	Calcium binding protein phosphatase	14
5	Tb927.5.4620	ESAG protein	3
8	Tb927.8.6930	Serine/threonine protein kinase NrkA	3
9	Tb09.142.0380	Hypothetical, conserved	1
9	Tb09.160.5400	ESAG9 associated protein	1
1	Tb927.1.5220	ESAG9 associated protein	1
8	Tb927.8.6170	Transketolase	1

SNP genotyping reveals distinct populations of Z310 and B17 isolates

STRUCTURE analysis of thirty-one SNP loci for 31 Ugandan *T. b. rhodesiense* isolates and a single Zambian isolate from 2010 (Chapter Four, Table 4.2), combined with next-generation sequencing data for six *T. brucei ssp.* isolates (plus two artificial crosses) suggested that 5 populations were most likely present in the data, as demonstrated by the highest value of LnP(D) [123] and by delta K analysis [124] (Figure 5.6). BAPS analysis of the same dataset suggested a different structure, with six populations (p=0.84) predicted to be the most likely, by predicting that *T. b. rhodesiense* EATRO2340 is an outlying isolate. A bar-plot for the STRUCTURE analysis using K=5 populations is shown in Figure 5.7, showing proportion of membership of each individual to the overall populations. The figure displays three populations of Z366, B17 and Z309-311 isolates, with two further populations of *T. b. brucei* with *T. b. rhodesiense* zymodeme Z377 and a mixed population of *T. b. gambiense*, *T. b. rhodesiense* B17 and B387, and *T. b. rhodesiense* EATRO3 and EATRO2340.

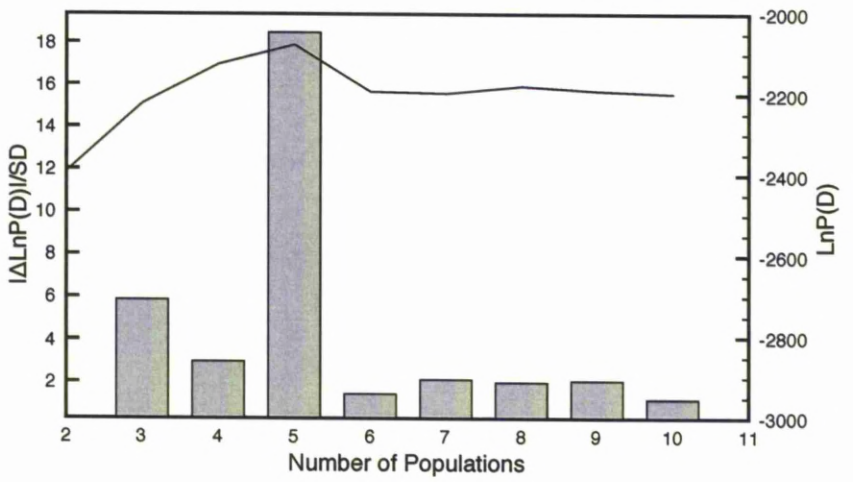


Figure 5.6: STRUCTURE analysis results for KASPAR SNP genotyping. Primary Y-axis (bar chart) represents Delta K analysis [124], where the most likely estimated number of populations from the data is indicated by the highest value of |ΔLnP(D)|/SD where |ΔLnP(D)| is the absolute rate of change between consecutive log probabilities for the estimated number of populations, and SD is the standard deviation from the mean for a given K. Secondary Y-axis (Line chart) is the corresponding log-likelihood that the number of popoulations estimated is the correct number. The most likely number of populations in the sample is the highest value of LnP(D) [123].

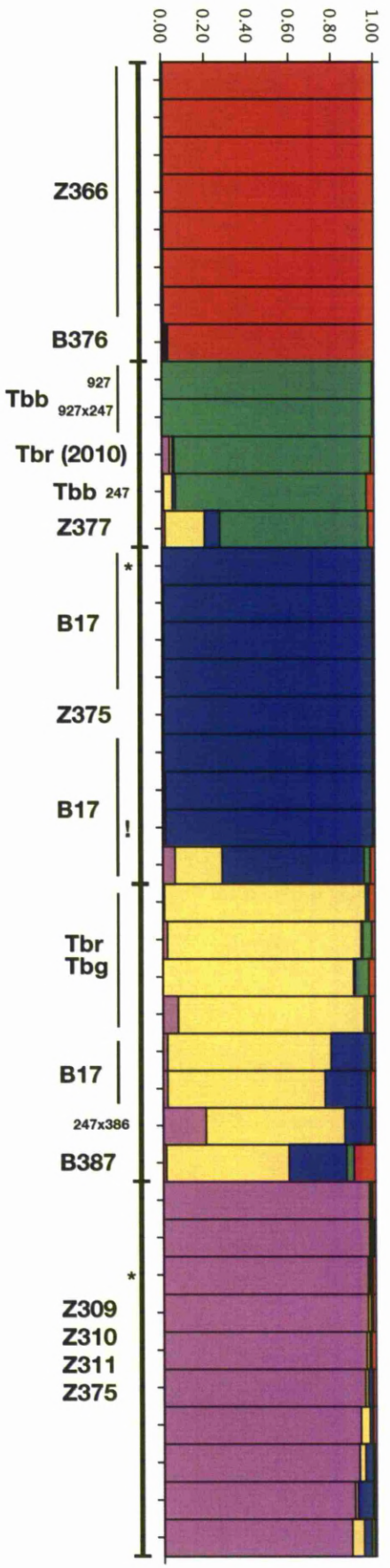


Figure 5.7: STRUCTURE bar plot for KASPAR SNP genotyping where K=5 populations, which was the most likely number of populations estimated by both LnP(D) and Delta K analysis (Figure 5.6). The bar plot represents isolates broken into five coloured segments representing the estimated membership that each individual has within each of the five inferred clusters.

Zymodemes are described for all *T. b. rhodesiense* isolates where available.

Other labels are as follows:

Tbb = *T. b. brucei* (927 = TREU927; 247 = STIB247; 927x247 = artificial cross between the two);

Tbg = *T. b. gambiense*; 247x386 = artificial cross between STIB247 and STIB386;

Tbr = *T. b. rhodesiense* EATRO3 and EATRO2340 isolates. Tbr (2010) = Zambian 2010 isolate;

! = Isolate labelled as Z310 but suspected to be B17. * = ABI SOLID sequenced strains on which the SNP panel was based.

Discussion

Isocitrate dehydrogenase as a marker for differences in virulence between *Busoga* and *Zambesi* strain group *T. b. rhodesiense* parasites

Whilst isoenzyme analysis has clearly distinguished between parasite strains, which correlated with differences in virulence [200], multilocus microsatellite genotyping at twelve sites genome-wide could not distinguish Z310 and B17 zymodemes (chapter four). Only one isoenzyme consistently differentiates between the *Zambesi* and *Busoga* strain groups: isocitrate dehydrogenase (ICD), for which GeneDB lists two genes within the *T. b. brucei* genome (v4; [42]). The two homologues are present on chromosomes 8 and 11, respectively (Ch8: Tb927.8.3690; Ch11: Tb11.03.0230), and as virulence appeared to be associated with the strain group of the given infecting parasite, it may be the case that the observed difference in virulence is linked to one, or both, of these markers.

Three non-synonymous SNP were predicted to occur within the chromosome 8 copy of ICD in the B17 isolate, relative to Z310. These SNP are predicted to make three heterozygous changes in amino acid sequence (Figure 5.2), which in turn changes the predicted charge and isoelectric point of the predicted protein (Table 5.2). Thus, these changes may be sufficient to create the observed change in the mobility of ICD on an MLEE thin-layer starch gel, as observed between *Zambesi* and *Busoga* strain groups (Dr. J. Wilson, personal communication). As no non-synonymous SNP were detected in the chromosome 11 orthologue of ICD, it is unlikely that this copy of the gene is responsible. On this evidence alone, differences in virulence between B17 and Z310 may be associated with chromosome 8, and not 11.

Other enzymes that are informative for differences between Z310 and B17 may also be associated with differences in heterozygosity around the *T. b. rhodesiense* genome: Whilst, on the evidence available, it is impossible to know exactly which genes are tested on an MLEE gel, of the six annotated genes that may code for phosphoglucomutase, which differentiates B17 from Z310, but not from Z366, only one – also on chromosome 8 (Tb927.8.980), contains a heterozygous, non-synonymous SNP. The SNP may affect an N-glycosylation site according to PROSITE [237], although POLYPHEN analysis

predicts the amino acid (AA) change to be benign [143] (AA #291; AA change: D>G; BLOSUM50: -1). Annotated genes for the other informative enzyme, nucleoside hydrolase (NHD), are present on chromosomes 3 and 7, although no non-synonymous SNP between Z310 and B17 were detected in either of these genes after SOLID sequencing.

It is important to note that, whilst ICD appears to be a marker for differences in heterozygosity at chromosome 8, which may correspond to differences in virulence, this does not account for all instances of acute disease. Acute symptoms associated with *Zambesi* strain groups were occasionally noted by Bailey (1997), and correspondingly, the recent (2010) Zambian isolate which was not predicted to have the non-synonymous SNP shared by the *Busoga* isolates, also presented as an acute disease. Whether the virulence of the disease is linked to the genotype of the host remains to be elucidated, particularly as the original (1989-1993) isolates were from the indigenous population, whereas the recent Zambian isolate was from a tourist, host effects cannot be discounted.

Shared heterozygous SNP at chromosome 8 may underlie differences in virulence between Z310 and B17 zymodemes

Expanding the dataset to genome-wide SNP loci suggests that the Z310 and B17 genomes are very similar (Figure 5.4B). The overall percentage difference between strains is less than 0.2%, which may underlie the problems with identifying the variation using microsatellite genotyping. Neighbour joining trees constructed in a chromosome-by-chromosome manner (Appendix VII: Figures A7.2.1-A7.2.11) suggests that variation between the Z310 and B17 zymodemes is restricted to differences at heterozygous loci on chromosomes 3, 5, 8 and 10. The largest difference was observed at chromosome 8, where B17, EATRO3, and EATRO2340 are each heterozygous across ~70% of the 2.5Mbp chromosome, with shared alleles from both Type 1 *T. b. gambiense*, and *T. b. brucei* TREU927 (Figures 5.3; 5.4C). Z310, by contrast, shows little introgression on chromosome 8 and is notably similar to *T. b. brucei*. A genetic cross between Type 2 *T. b. gambiense* and *T. b. brucei* lies at a point on the phylogenetic tree between *T. b. gambiense* and B17, highlighting the greater similarity between these genomes that between Type 2 *T. b. gambiense* and Z310.

Z310 shows similar patterns of heterozygosity on other chromosomes: Regions of chromosomes 3, 5 and 10 also display similar patterns of heterozygosity with *T. b. gambiense*, as shown by the extended distance of Z310 from B17 on the phylogenetic trees shown in Appendix VII (Figures A7.2.3; A7.2.5; A7.2.10, respectively). This is in contrast to, for example, chromosome 11, which shows little heterozygosity, and as such the isolates are closer together on the tree (Appendix VII: Figure A7.2.11). The KASPAR genotyping results suggest that similar variation occurs within the B17 population, and it is variation at these sites that results in B17 isolates being split between two population clusters, and highlights the importance of the shared alleles at chromosome 8 possibly underlying differences in virulence between Z310 and B17 zymodemes.

Through a combination of microsatellite genotyping and SRA and kDNA sequencing, Balmer and colleagues have identified *T. b. rhodesiense* in amongst seven (out of eleven) distinct populations of the underlying *T. b. brucei* population, corroborating the hypothesis that *T. b. rhodesiense* is not monophyletic [122]. Delta K analysis of their data identified five populations, the same number as this dataset (Figure 5.6), but the authors suggested that the hierarchical data structure hid the true number of populations (eleven). Certainly, examining our dataset using the BAPS package suggested that six populations was the most likely, albeit doing so by predicting EATRO2340 as an outlier, which was not suggested by STRUCTURE. Additional data may reveal further sub-structuring and extend the predicted number of population clusters. The microsatellite data presented by Balmer *et al* did not show the introgression of *T. b. gambiense* alleles within *T. b. rhodesiense* isolates, however, in line with our microsatellite data (Chapter Four), that was unable to distinguish Z310 from B17. Nevertheless, both datasets suggest Types 1 and 2 *T. b. gambiense* are divergent from other *T. brucei* sub-species.

Identifying candidate genes that may influence differential virulence between zymodemes

Tables 5.3A-D show the most differentially expressed genes with non-synonymous SNP between zymodemes [232]. Four differentially expressed genes with nsSNP were ESAG-related proteins and a further two are invariant surface glycoproteins, which are typically sub-telomeric, repetitive, and associated with antigenic variation. As such, any putative differences may be a function of the inherent difficulty in aligning short sequence reads to repetitive sequences. Similarly, eight genes have been annotated as 'hypothetical' and as such further conclusions about the significance of nsSNP within these genes are difficult to draw.

Two identified genes: Tb10.70.1720 (Table 5.3A) and Tb927.3.930 (Table 5.3C) encode dynein heavy chain molecules and contained four and one nsSNP respectively. The dynein complex is involved in flagellum manufacture, and as such, functional nsSNP may cause problems with motility. Indeed, motility is necessary for cell division as flagella induce the shearing forces necessary to separate daughter cells from their parents [242], and so it is, perhaps, unsurprising to find these genes as being preferentially expressed in the dividing, slender, bloodstream forms.

Two annotated genes identified as having nsSNP in B17 (relative to Z310) and being expressed to a greater extent in stumpy forms (Table 5.3D) are a transketolase (Tb927.8.6170) and a protein phosphatase (Tb927.8.1130). Transketolases are key metabolic enzymes involved in glucose metabolism via the pentose phosphate pathway and had been previously detected to be absent in the bloodstream form, versus the procyclic form [243]. The protein phosphatase on chromosome 8, Tb927.8.1130, has been recently functionally annotated with the aid of RNA-interference target sequencing and knockdowns shown to be non-lethal in bloodstream and procyclic forms [244]. Tb927.8.6930, also on chromosome 8, codes for NrkA, a serine/threonine protein kinase that displays a 1.78-fold increase in expression in stumpy forms compared to slender [232] and as such was identified as being one of the most differentially expressed genes between bloodstream forms. Furthermore, sequencing of B17 identified 3 nsSNP compared to Z310 (Table 5.3D) and as such may represent an interesting candidate gene for differential virulence displayed between zymodeme strain

groups. Protein kinases have been identified as possible drivers of slender to stumpy differentiation [238], and excreted proteins such as kinases and phosphatases represent good candidates for virulence genes, as transduction pathways will rapidly amplify the signals that they affect, and have multiple effects upon the host cell. Indeed, secreted kinases and phosphatases have been identified as potential virulence factors in bacteria [245], and polymorphic serine-threonine kinases are important effectors of disease in *Toxoplasma gondii* [246]. The interest in protein kinases and phosphatases is not limited to their potential as virulence factors as they have important roles in metabolism and cell development. Protein kinases and phostphatases play a pivotal role in controlling proliferation and differentiation in protozoa [247], including *Trypanosoma* [248], [249].

STRUCTURE analysis suggests that *T. b. rhodesiense* is not clonal and monophyletic

STRUCTURE analysis of the KASPAR SNP data (Figure 5.7) suggests that the parasite isolates can be split into five populations. Firstly, a population of Z366 (and a single B376 isolate) and a second population of Z377 with *T. b. brucei* and a recent acute Zambian infection were distinct. A third population cluster contained a single Z375 isolate with zymodemes Z309, Z310 and Z311. B17 isolates were split between the remaining two population clusters: One distinct population (also including a *Zambesi* 375 isolate) and another population, more closely related to *T. b. gambiense*, the *Busoga* 387 isolate and two *T. b. rhodesiense* isolates from earlier outbreaks (EATRO3 and EATRO2340, from Uganda, 1960 and Kenya, 1977, respectively). The population containing Z310 isolates appears to be distinguished from the two populations containing B17 due to heterozygous SNP loci at chromosome 8. Differences at similar heterozygous loci on chromosomes 2, 3, 5 and 10 separate the two B17 populations (Appendix VII: Figure A7.1.1).

Given that the SNP loci assayed by KASPAR genotyping were chosen to represent differences between B17 and Z310 isolates, it is perhaps not surprising that the populations of the two zymodemes can be split in this manner. Nevertheless, SNP genotyping distinguishes more divergent zymodemes such as Z366, and suggests that variation within the *Busoga* strain-group is greater than within *Zambesi*, as highlighted by

the presence of two population clusters containing B17 isolates, and a single population of Z309-311.

Are *T. b. rhodesiense* and *T. b. gambiense* sympatric?

Our data suggests that the different HAT subspecies have been sympatric. Given that the host range of *T. b. gambiense* is mainly restricted to man, it is likely that any genetic exchange would occur in Tsetse co-infected with *T. b. rhodesiense* from either livestock or wild-animals, and from North-Western Ugandans or refugee Sudanese, with chronic *T. b. gambiense* disease. A 2005 study focussing on microsatellite data for in humans and livestock suggested that Ugandan foci remained discrete and existed only 150Km apart [250]. The prospect of the HAT diseases existing sympatrically raises the prospect of new, virulent, foci of disease, and complicates the process of monitoring and the treatment of infections. The prospect of spreading drug resistance throughout the *T. brucei* population is also a possibility.

Conclusions

Analysis of four *T. b. rhodesiense* genomes has revealed clear differences and recombination between isolates. Phylogenetic networks constructed on the basis of genome-wide SNP show Ugandan B17 and Z310 are similar and the Ugandan EATRO3 and Kenyan EATRO2340 are dissimilar to them, and to each other. A focussed analysis on 32 *T. b. rhodesiense* strains at 31 non-synonymous SNP loci (Figure 5.4A) suggests recombination has occurred between strains, but with clear lineages of specific zymodemes. Previous analyses have suggested that the population structure of *T. b. rhodesiense* is either panmictic [77], with rapid clonal expansion of virulent epidemic sub-types; or clonal [76], and our data favours the former panmictic hypothesis.

Our data suggests that the genomes of *T. brucei ssp.* parasites are recombining regularly, resulting in differential virulence and pathogenicity in the field. Furthermore, hotspots of recombination, such as chromosomes 5, 8 and 10 suggest that these loci underlie the differences in phenotype; By way of contrast, chromosome 11 appears to be very conserved between species, in line with the *T. b. brucei* (TREU927) genetic map (recombination unit: 95.64 kb/cM) [251] and the *T. b. gambiense* (STIB386) genetic map

(up to 170kb/cM) [252], which both suggested that little recombination was taking place.

Whilst our data suggest that recombination between *T. b. gambiense* and *T. b. rhodesiense* has occurred, it remains to be elucidated whether this has occurred directly. It is possible that a lack of data from *T. b. brucei* populations in Western Africa may result in our data being solely a function of geographical isolation, i.e. has identified the presence of recombination between East-African and West-African *T. b. brucei* strains, which have each, in turn, crossed with other sub-species in their locality. In order to further reveal the underlying process of the spread of *T. b. gambiense* alleles through the *T. b. rhodesiense* population, it is important to further study the underlying *T. b. brucei* population: As *T. b. brucei* TREU927 is east African in origin, comparing our data to other *T. b. brucei* isolates from central and Western Africa may help resolve this issue.

It is still not known whether it the presence of *T. b. gambiense* – like alleles in B17 directly leads to increased virulence, or whether the loss of these alleles (particularly across chromosome 8), leads to decreased virulence in Z310 due to it being more like *T. b. brucei*. Understanding the mechanism by which these differences arise may help solve this problem: Studies elsewhere on cultured forms of trypanosomes have suggested that a 'loss of heterozygosity' can lead to benefits to growth in culture [252]. The ability to perform crosses in the laboratory setting may provide the opportunity to resolve this issue: Trypanosomes have been shown to self-fertilise [253], and so the selection of progeny from a 'selfed' B17-line with a homozygous chromosome 8 (for alleles from both of *T. b. gambiense* and *T. b. brucei*), and subsequent phenotyping for differences in virulence may show whether this was the case: A corresponding loss of virulence corresponding to a *T. b. brucei* – like chromosome 8, would suggest that it is indeed these alleles that contribute towards differential virulence.

Nevertheless, the presence of heterozygosity that may be correlating with differential virulence is an important observation. Furthermore, these data highlight the importance of monitoring the disease, particularly in Uganda where the two sub-species exist in close proximity to one another, particularly with respect to the possibility of the spread of drug resistance and of novel, virulent, epidemics arising

Secreted protein kinases and phosphatases as potential drivers of differences in virulence

One explanation for the observed differences in virulence between parasite zymodemes is that virulence is driven by variable rates of differentiation from slender to stumpy forms, the latter being more immunogenic than the former due to VSG shedding [254]. Differentiation from slender to stumpy bloodstream forms is mediated by cell-density sensing [229] through the release of a, so-far unidentified, stumpy induction factor (SIF). SIF is a low molecular weight protein that triggers cell-cycle arrest in G1/G0 [255] via the cAMP pathway [230]. Whilst there are many candidates for molecules that fulfil this role, protein kinases are prominent candidates for being involved in the signalling transduction pathway as many signal via cAMP and, together, can act as negative regulators of the response of B lymphocytes, T lymphocytes and macrophages; key factors in trypanosome pathogenesis [256].

Key to understanding this mechanism is elucidating whether the relative numbers of stumpy and slender parasites exist within the bloodstream between experimental infections. The recent identification of a stumpy-specific protein, PAD1 [257] would allow for the construction of fluorescently labelled antibodies to aid the detection and counting of the relative numbers of the different bloodstream forms in experimental infections by fluorescent-based sorting methods such as FACS. Assessing the relative numbers of the different forms of the parasite throughout infection would be crucial to understanding these mechanisms, and until this is resolved, has implications for studying virulence in *T. b. rhodesiense* in the mouse model. Sequencing further isolates and comparing the patterns of heterozygosity and comparing this to the relative numbers of stumpy and slender forms throughout bloodstream infection may reveal whether the observed differential heterozygosity is responsible for decreased differentiation to the stumpy form, decreased transmissibility and whether an associated increased virulence in mice is connected.

Clones that do not readily differentiate to stumpy forms in mice, do so in cattle [258], so it is possible that the same phenotype would not be seen in humans and therefore that there exists a mouse (or rodent) factor that explains the apparent inverse correlation between virulence in man, and in mice other than the very-susceptible A/J strain.

Chapter Six: Conclusions and Further Work

This project sought to utilise the recent advances in molecular genetics technologies to study of some of the genetic aspects underlying host resistance in the mouse to *T. congolense* infection and the factors underlying differential virulence between sub-types of *T. b. rhodesiense* from south-Eastern Uganda. As trypanosomes cause economically and clinically important diseases across much of sub-Saharan Africa, identifying genes that may be responsible for regulating response to infection, or that modulate the virulence of the parasite, may help reduce the burden of the disease on the population: Identifying candidate genes and the underlying pathways involved with resistance may aid scientists in breeding more tolerant cattle; Similarly, studying genes and loci involved with parasite pathogenesis represents useful data for monitoring the disease in the field, and any genes identified may represent useful therapeutic targets.

Candidate genes regulating response to *T. congolense* infection

T. congolense causes economic hardship to large numbers of farmers due to the effect of the parasite on cattle. Whilst trypanotolerant breeds of cattle exist (cattle that have an innate ability to remain productive despite infection), they are not taken up by the farmers due to their preference for larger breeds that are easier to handle and are generally more productive, relying instead on disease treatment and monitoring.

Different breeds of inbred mice exhibit a similar phenotype to cattle trypanotolerance: survival time after infection varies between breeds [105] and three major genetic loci have been shown to be responsible [126]. Comprehensive genetic analysis of these QTL has revealed candidate genes at two known QTL – *Cd244* and *Pram1* – that may suggest pathways that underlie tolerance in cattle [149]. CD244 is involved with NK cell immunity and cytokine production: it belongs to the signalling lymphocyte activation molecule (SLAM)-related receptor family, which in turn regulate a wide range of immune cell types [259]. *Pram-1* is an important molecule involved with neutrophil function and cytokine production that initiate and amplify the inflammatory response [158]. Both molecules are differentially expressed throughout the course of

infection in mice and as such represent good candidates for functional testing in mice, to ascertain whether the genes are, in fact, the QTL genes.

Confirming candidate SNP and potential pathways involved in response to infection

Improvements to identifying causal genes within QTL are constantly sought [260, 261]. The approach that we have demonstrated of systematically reducing the number of candidate genes at a QTL by combining haplotype analysis, array-CGH, gene expression and next-generation DNA capture and sequencing has been shown to reduce the number of candidate genes at a QTL to a list of genes short enough to test for function. Nevertheless, confirmation of the role of the candidate genes in response to infection may be a necessary first step: Confirmation of the role of nsSNP in *Pram1* may be difficult, as whilst knockouts for *Pram1* exist, knocking out the gene will likely have a non-specific effect on survival after infection without providing information on the specific role of the polymorphism. Furthermore, currently available *Pram1* knockouts are based on the 129/J background, which may have other susceptible alleles around *Tir1*, and as such any survival experiment would not only measure the effect of the knockout, but also of all 129 alleles in the region. Better would be the use of allele-substitution, such as utilising a C57BL/6N *Pram1* knockout (for which embryonic stem cell lines for most genes now exist), to rescue them with artificial constructs containing the A/J and C57BL/6 alleles and looking for differences in survival time after rescue. It would be easier to confirm the suspected role of CNV involving *Cd244* at *Tir3c* by artificially inserting an additional copy of *Cd244* into a C57BL/6 mouse, so that it had a similar gene dosage to the susceptible strains and observing survival time after infection.

Several of these QTL genes may also influence susceptibility to other infections. QTL involved with resistance to other parasitic diseases overlap with the *Tir* QTL: *Leishmania* resistance 1 (*Lmr1*) [168], *Plasmodium chabaudi* resistance QTL 3 (*Char3*) [169] and *Heligmosomoides bakeri* nematode resistance 2 (*Hbnr2*) [170] all overlap with *Tir1* and the *Tir3c* QTL overlaps with a QTL for murine resistance to *Plasmodium berghei*-driven experimental cerebral malaria (*Berr1*) [171]. As such the identification of candidate genes and their associated pathways in trypanosomiasis, and their subsequent confirmation may contribute to our understanding of the wider response to infectious disease.

Trypanosomiasis in cattle, humans and mice.

Clearly, it is difficult to draw parallels between the mouse model, bovine and human disease, and caution must always be exercised in extrapolating the results obtained in animal models to that of humans, particularly due to the greater complexity of the human nervous system with respect to parasite infection of the CNS. There are also differences in the response to trypanosome infection between host species – whilst both bovine and murine trypanotolerance are associated with effective parasite control, bovine tolerance is associated with decreased tissue lesions, whereas tolerance exhibited by C57BL/6 mice corresponds to an increase in associated lesions [108].

It is important to note, however, that there remains utility in using mouse models to study the effects of trypanosome infection. Whilst primary pathology in mice is not from CNS involvement, models do exist for such disease and such studies have proven useful in understanding the effects of arsenical-based treatment and the related post-treatment reactive encephalopathy (PTRE) [85], for IFNG-related passage of trypanosomes across the BBB [262] and even in displaying similarities in HAT presentation, with respect to disturbance of circadian rhythm and sleep patterns associated with late-stage disease [263]. Similarly, the utilisation of the wide-range of animal models available has resulted in the discovery of some similarities in the genetics of trypanotolerance. These have included the identification of candidate genes involved in the NK cell response (ARHGAP15 in cattle; CD244/CD48 in mice [187]) and the critical role of cytokines such as TNFa in the development of anaemia [110, 264]. Interestingly, a study has identified the role for TNFa and IL-10 SNP in HAT resistance. As such, it appears that TNFa (and IL-10) response is crucial in all three of human, mouse and cattle disease progression [265].

Variation in disease phenotype between inbred strains and in HAT

The problem with using mouse models to study HAT is confounded by two main factors, the genotype of the host, and the species of the infecting parasite. This is arguably best illustrated by the data in Table 4.3, where the strain of inbred mouse and the genotype of the infecting (*T. b. rhodesiense*) parasite had an effect on survival after infection: A/J mice (congenic for C57BL/6 alleles at *Tir1*) are more susceptible to B17, in line with human infection; BALB/c, 129/sv and C57BL/6 mice show the opposite correlation. A further confounding factor is that the sex of the mouse has been shown to have an effect on survival [206]. Death after infection in inbred mice can result from a number of routes, including systemic inflammatory response syndrome (SIRS), anaemia or CNS invasion [Reviewed [108]]. The degree to which these modes of mortality contribute to death also differs between breeds of mouse: mortality in A/J mice after *T. congolense* infection is associated with higher parasitaemia than C57BL/6, which in turn develop severe anaemia towards the latter stages of survival that is not present in A/J [266]. Death in the early stages of *T. brucei* and *T. congolense* infection (due to SIRS) in BALB/c mice is associated with increased IFNG [267]. Correspondingly, increased IFNG expression is associated with the increased prevalence of CNS involvement between spatially distinct *T. b. rhodesiense* HAT infections [85]. We have shown IFNG to be more greatly expressed in *T. b. rhodesiense* Z310 infections versus B17 (Figure A5.4b), which correlates with all breeds of inbred mouse examined except for A/J mice, and as such clearly confounds the situation in humans, wherein Z310 infections are more chronic in their presentation.

Epidemic *T. b. rhodesiense* strains have differential heterozygosity that may associate with virulence

Sympatric zymodemes of *T. b. rhodesiense* have shown differential virulence in humans and mice. In order to identify loci that may contain virulence genes that drive the observed differences, nine different zymodemes from the 1980's/early 1990's epidemic in Uganda were initially characterised by microsatellite analysis. Subsequently, the whole-genome sequences of the two predominant zymodemes from the epidemic – *Zambesi 310* and *Busoga 17*, were compared. Despite the two genomes being >99.8% identical, differential heterozygosity was identified between the two zymodemes that occurred in discrete regions of at least five chromosomes: chromosomes: 2, 3, 5, 8 and

10. Microsatellite data also suggested that Z366 isolates, which present an intermediate clinical profile to Z310 and B17 [200], were distinguished by the presence of a unique heterozygous allele at five loci across chromosomes 1, 2, 8 and 11, further suggesting that heterozygosity may have an important role in differences in virulence between outbreaks. Whole genome sequencing of additional Z310, B17, Z366 and other isolates, alongside correlation of pathogenicity in experimental mice will be necessary to elucidate which genetic loci are driving these differences.

Identifying the loci driving differences in phenotype remains a major challenge. Colleagues at the University of Glasgow have identified a loss of heterozygosity associated with chromosome 10 that confers a growth advantage in culture [252]. The same group has identified a trypanosome QTL associated with enlarged liver and spleen in mice on chromosome 3 from a synthetic cross between two *T. b. brucei* laboratory strains with different virulence phenotypes. The differential heterozygosity on chromosome 8 alone affects 507 genes (Figure 5.5); Many of these are annotated as 'hypothetical genes' and as such, improvements to the annotation of *T. b. brucei* TREU927 may yield insights into candidate pathways or genes that may affect virulence. In order to further investigate which genetic loci are responsible for the virulence phenotype, a similar synthetic cross to that produced by Cooper *et al*, between B17 and Z310, and subsequent phenotyping and genetic mapping may reveal QTL associated with differences in virulence between zymodemes. This however, represents a major effort, as phenotyping large numbers of parasite infections, and subsequently identifying the genetic loci that underlie these traits can be expensive and laborious. Despite these challenges, these data represents an important step in identifying loci that underlie differences in virulence between epidemic *T. b. rhodesiense* populations.

Whilst differences in virulence between zymodemes have been established in humans and the mouse model, they have not been established in tsetse. If, indeed, rates of differentiation are responsible for the virulence phenotype, experimental infections in tsetse may reveal differences in infection rates as those parasites that have predominantly greater numbers of stumpy forms should be more infective to flies.

Recent studies have been made into expression differences between stumpy and slender forms of *T. b. brucei* TREU927 [226]. This has allowed for the hypothesis that

differences in the rate of differentiation from slender to stumpy forms being responsible for differences in virulence between zymodeme strain groups to be further studied. Whilst speculative, by comparing those genes affected by non-synonymous polymorphisms with those showing differential expression between stumpy and slender forms, lists of potential candidate genes have been generated (Table 5.3). It would be useful to assess the relative numbers of slender and stumpy forms throughout mouse infection with *T. b. rhodesiense*. Assessing infections in a range of inbred mouse breeds with several zymodemes can now be facilitated by using fluorescence-based sorting using an antibody against a stumpy-specific protein, such as PAD1 [257].

Examining the differences in expression between zymodemes of *T. b. rhodesiense* could be performed utilising the recent advances in expression studies on next-generation sequencing platforms. Holzmuller and colleagues have demonstrated that differences between the secretomes of genetically homologous strains of *T. b. gambiense* can be correlated to differences in virulence [198]. It may be possible to employ techniques such as allele-specific expression (ASE) assays [268], or RNA-seq to rapidly assess expression differences in a wide number of *T. b. rhodesiense* isolates. RNA-seq has previously been used to identify the expression profile of the laboratory adapted Ugandan *T. b. rhodesiense* strain YTat1.1 [269].

References

1. Welburn, S.C., et al., *Sleeping sickness: a tale of two diseases.* Trends Parasitol, 2001. **17**(1): p. 19-24.
2. Simarro, P.P., et al., *The Atlas of human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases.* Int J Health Geogr, 2010. **9**: p. 57.
3. Hoare, C., *The Trypanosomes of Mammals.* 1972: Blackwell Scientific Publications, Oxford.
4. Herwaldt, B.L., *Leishmaniasis.* Lancet, 1999. **354**(9185): p. 1191-9.
5. Haag, J., C. O'HUigin, and P. Overath, *The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria.* Mol Biochem Parasitol, 1998. **91**(1): p. 37-49.
6. Barrett, M., et al., *The trypanosomiases,* in *The Lancet.* 2003. p. 1469-1480.
7. Prata, A., *Clinical and epidemiological aspects of Chagas disease.* Lancet Infect Dis, 2001. **1**(2): p. 92-100.
8. WHO. *WHO report on global surveillance of epidemic-prone infectious diseases - African trypanosomiasis.* 2001 (Accessed 2008); Available from: www.who.int/emc-documents/surveillance/docs/whocdscsrisr2001.html/African_Trypanosomiasis/A_Trypanosomiasis.htm.
9. Fevre, E.M., et al., *Human African trypanosomiasis: Epidemiology and control.* Adv Parasitol, 2006. **61**: p. 167-221.
10. Jones, T.W. and A.M. Davila, *Trypanosoma vivax--out of Africa.* Trends Parasitol, 2001. **17**(2): p. 99-101.
11. Brun, R., H. Hecker, and Z.R. Lun, *Trypanosoma evansi and T. equiperdum: distribution, biology, treatment and phylogenetic relationship (a review).* Vet Parasitol, 1998. **79**(2): p. 95-107.
12. Ford, J., *The role of the trypanosomiases in African ecology.* 1970: Clarendon Press, London.
13. Agu, W.E., *Comparative study of the susceptibility to infection with Trypanosoma simiae of Glossina morsitans and G. tachinoides.* Acta Trop, 1984. **41**(2): p. 131-4.
14. Murray, M. and A.R. Gray, *The current situation on animal trypanosomiasis in Africa.* Preventive Veterinary Medicine, 1984. **2**(1-4): p. 23-30.
15. Luckins, A.G., *Trypanosoma evansi in Asia.* Parasitol Today, 1988. **4**(5): p. 137-42.
16. Wellde, B.T., et al., *Experimental infection of cattle with Trypanosoma brucei rhodesiense.* Ann Trop Med Parasitol, 1989. **83 Suppl 1**: p. 133-50.
17. Bengaly, Z., et al., *Comparative pathogenicity of three genetically distinct types of Trypanosoma congolense in cattle: clinical observations and haematological changes.* Vet Parasitol, 2002. **108**(1): p. 1-19.
18. Gardiner, P.R., *Recent studies of the biology of Trypanosoma vivax.* Adv Parasitol, 1989. **28**: p. 229-317.
19. Welburn, S.C., et al., *Crisis, what crisis? Control of Rhodesian sleeping sickness.* Trends Parasitol, 2006. **22**(3): p. 123-8.

20. Kristjanson, P.M., et al., *Measuring the costs of African animal trypanosomiasis, the potential benefits of control and returns to research.* Agricultural Systems, 1999. **59**: p. 79–98.
21. Committee on African Trypanosomiasis., *Scientific working group on African trypanosomiasis (sleeping sickness) WHO/Tropical Disease Research Unit.* 2001.
22. Mathieu-Daude, F., et al., *Genetic diversity and population structure of Trypanosoma brucei: clonality versus sexuality,* in *Molecular & Biochemical Parasitology.* 1995.
23. Barrett, M.P., *The rise and fall of sleeping sickness.* Lancet, 2006. **367**(9520): p. 1377-8.
24. Fevre, E.M., et al., *A burgeoning epidemic of sleeping sickness in Uganda.* Lancet, 2005. **366**(9487): p. 745-7.
25. Tshimungu, K., et al., *Re-emergence of human African trypanosomiasis in Kinshasa, Democratic Republic of Congo (DRC).* Med Mal Infect, 2010. **40**(8): p. 462-7.
26. Moore, D.A., et al., *African trypanosomiasis in travelers returning to the United Kingdom.* Emerg Infect Dis, 2002. **8**(1): p. 74-6.
27. Matemba, L.E., et al., *Quantifying the burden of rhodesiense sleeping sickness in Urambo District, Tanzania.* PLoS Negl Trop Dis, 2010. **4**(11): p. e868.
28. Stich, A., P.M. Abel, and S. Krishna, *Human African trypanosomiasis.* BMJ, 2002. **325**(7357): p. 203-6.
29. Cattand, P., *The scourge of human African trypanosomiasis.* Afr Health, 1995. **17**(5): p. 9-11.
30. Sternberg, J.M., *Human African trypanosomiasis: clinical presentation and immune response.* Parasite Immunol, 2004. **26**(11-12): p. 469-76.
31. Steverding, D., *The history of African trypanosomiasis.* Parasit Vectors, 2008. **1**(1): p. 3.
32. Hide, G., et al., *The origins, dynamics and generation of Trypanosoma brucei rhodesiense epidemics in East Africa.* Parasitol Today, 1996. **12**(2): p. 50-5.
33. Onyango, R.J., K. Van Hoeve, and P. De Raadt, *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, 1966. **60**(2): p. 175-82.
34. Odiit, M., et al., *Spatial and temporal risk factors for the early detection of Trypanosoma brucei rhodesiense sleeping sickness patients in Tororo and Busia districts, Uganda.* Trans R Soc Trop Med Hyg, 2004. **98**(10): p. 569-76.
35. Drennan, M.B., et al., *The induction of a type 1 immune response following a Trypanosoma brucei infection is MyD88 dependent.* J Immunol, 2005. **175**(4): p. 2501-9.
36. Olsson, T., et al., *CD8 is critically involved in lymphocyte activation by a T. brucei brucei-released molecule.* Cell, 1993. **72**(5): p. 715-27.
37. Magez, S., et al., *Tumor necrosis factor alpha is a key mediator in the regulation of experimental Trypanosoma brucei infections.* Infect Immun, 1999. **67**(6): p. 3128-32.
38. Magez, S., et al., *Specific uptake of tumor necrosis factor-alpha is involved in growth control of Trypanosoma brucei.* J Cell Biol, 1997. **137**(3): p. 715-27.
39. Engstler, M., et al., *Kinetics of endocytosis and recycling of the GPI-anchored variant surface glycoprotein in Trypanosoma brucei.* J Cell Sci, 2004. **117**(Pt 7): p. 1105-15.

40. Barry, J.D. and R. McCulloch, *Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite.* Adv Parasitol, 2001. **49**: p. 1-70.
41. Kamper, S.M. and A.F. Barbet, *Surface epitope variation via mosaic gene formation is potential key to long-term survival of Trypanosoma brucei.* Mol Biochem Parasitol, 1992. **53**(1-2): p. 33-44.
42. Berriman, M., et al., *The genome of the African trypanosome Trypanosoma brucei.* Science, 2005. **309**(5733): p. 416-22.
43. Raper, J., *The main lytic factor of Trypanosoma brucei brucei in normal human serum is not high density lipoprotein,* in *Journal of Experimental Medicine.* 1996. p. 1023-1029.
44. Hager, K.M., et al., *Endocytosis of a cytotoxic human high density lipoprotein results in disruption of acidic intracellular vesicles and subsequent killing of African trypanosomes.* J Cell Biol, 1994. **126**(1): p. 155-67.
45. Raper, J., et al., *Characterization of a novel trypanosome lytic factor from human serum.* Infect Immun, 1999. **67**(4): p. 1910-6.
46. Smith, A.B. and S.L. Hajduk, *Identification of haptoglobin as a natural inhibitor of trypanocidal activity in human serum.* Proc Natl Acad Sci U S A, 1995. **92**(22): p. 10262-6.
47. Vanhollebeke, B. and E. Pays, *The trypanolytic factor of human serum: many ways to enter the parasite, a single way to kill.* Mol Microbiol, 2010. **76**(4): p. 806-14.
48. Kieft, R., et al., *Mechanism of Trypanosoma brucei gambiense (group 1) resistance to human trypanosome lytic factor.* Proc Natl Acad Sci U S A, 2010. **107**(37): p. 16137-41.
49. De Greef, C., et al., *A gene expressed only in serum-resistant variants of Trypanosoma brucei rhodesiense.* Mol Biochem Parasitol, 1989. **36**(2): p. 169-76.
50. Vanhamme, L., et al., *Apolipoprotein L-I is the trypanosome lytic factor of human serum.* Nature, 2003. **422**(6927): p. 83-7.
51. Oli, M.W., et al., *Serum resistance-associated protein blocks lysosomal targeting of trypanosome lytic factor in Trypanosoma brucei.* Eukaryot Cell, 2006. **5**(1): p. 132-9.
52. Lugli, E.B., et al., *Characterization of primate trypanosome lytic factors.* Mol Biochem Parasitol, 2004. **138**(1): p. 9-20.
53. Bouteille, B., et al., *Treatment perspectives for human African trypanosomiasis.* Fundamental & Clinical Pharmacology, 2003. **17**(2): p. 171-181.
54. Pepin, J. and F. Milord, *The treatment of human African trypanosomiasis.* Adv Parasitol, 1994. **33**: p. 1-47.
55. Balasegaram, M., et al., *Melarsoprol versus eflornithine for treating late-stage Gambian trypanosomiasis in the Republic of the Congo,* in *Bull World Health Organ.* 2006. p. 783-91.
56. Ross, R. and D. Thomson, *A Case of Sleeping Sickness showing Regular Periodical Increase of the Parasites Disclosed.* Br Med J, 1910. **1**(2582): p. 1544-1545.
57. Kabayo, J.P., *Aiming to eliminate tsetse from Africa.* Trends Parasitol, 2002. **18**(11): p. 473-5.
58. Vreysen, M.J.B., et al., *Glossina austeni (Diptera : Glossinidae) eradicated on the Island of Unguja, Zanzibar, using the sterile insect technique.* Journal of Economic Entomology, 2000. **93**(1): p. 123-135.

59. Simarro, P.P., J. Jannin, and P. Cattand, *Eliminating human African trypanosomiasis: where do we stand and what comes next?* PLoS Med, 2008. **5**(2): p. e55.
60. Melville, S.E., et al., *The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers.* Mol Biochem Parasitol, 1998. **94**(2): p. 155-73.
61. Hertz-Fowler, C., et al., *GeneDB: a resource for prokaryotic and eukaryotic organisms.* Nucleic Acids Res, 2004. **32**(Database issue): p. D339-43.
62. Taylor, J.E. and G. Rudenko, *Switching trypanosome coats: what's in the wardrobe?* Trends Genet, 2006. **22**(11): p. 614-20.
63. Melville, S.E., C.S. Gerrard, and J.M. Blackwell, *Multiple causes of size variation in the diploid megabase chromosomes of African tyrpanosomes.* Chromosome Res, 1999. 7(3): p. 191-203.
64. Callejas, S., et al., *Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length.* Genome Res, 2006. **16**(9): p. 1109-18.
65. Barry, J.D., et al., *Why are parasite contingency genes often associated with telomeres?* Int J Parasitol, 2003. **33**(1): p. 29-45.
66. Hall, N., et al., *The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism.* Nucleic Acids Res, 2003. **31**(16): p. 4864-73.
67. Myler, P.J., et al., *Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2902-6.
68. Liang, X.H., et al., *trans and cis splicing in trypanosomatids: mechanism, factors, and regulation.* Eukaryot Cell, 2003. **2**(5): p. 830-40.
69. Wickstead, B., K. Ersfeld, and K. Gull, *The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes.* Genome Res, 2004. **14**(6): p. 1014-24.
70. Van der Ploeg, L.H., et al., *Improved separation of chromosome-sized DNA from Trypanosoma brucei, stock 427-60.* Nucleic Acids Res, 1989. **17**(8): p. 3217-27.
71. Liu, B., et al., *Fellowship of the rings: the replication of kinetoplast DNA.* Trends Parasitol, 2005. **21**(8): p. 363-9.
72. Estevez, A.M. and L. Simpson, *Uridine insertion/deletion RNA editing in trypanosome mitochondria--a review.* Gene, 1999. **240**(2): p. 247-60.
73. Ford, J. and E. Blaser, *Some aspects of cattle raising under prophylactic treatment against trypanosomiasis on the Mkwaja Ranch, Tanzania.* Acta Trop, 1971. **28**(2): p. 69-79.
74. Turner, D.A. and R. Brightwell, *An evaluation of a sequential aerial spraying operation against Glossina pallidipes Austen (Diptera: Glossinidae) in the Lambwe Valley of Kenya: aspects of post-spray recovery and evidence of natural population regulation.* Bulletin of Entomological Research, 1986. **76**(02): p. 331-349.
75. Brun, R. and O. Balmer, *New developments in human African trypanosomiasis.* Curr Opin Infect Dis, 2006. **19**(5): p. 415-20.
76. MacLeod, A., et al., *Minisatellite marker analysis of Trypanosoma brucei: reconciliation of clonal, panmictic, and epidemic population genetic structures.* Proc Natl Acad Sci U S A, 2000. **97**(24): p. 13442-7.

77. Maynard Smith, J., et al., *How clonal are bacteria?* Proc Natl Acad Sci U S A, 1993. **90**(10): p. 4384-8.
78. Jenni, L., et al., *Hybrid formation between African trypanosomes during cyclical transmission.* Nature, 1986. **322**(6075): p. 173-5.
79. Gibson, W., et al., *Analysis of a cross between green and red fluorescent trypanosomes.* Biochem Soc Trans, 2006. **34**(Pt 4): p. 557-9.
80. Gibson, W. and J. Stevens, *Genetic exchange in the trypanosomatidae.* Adv Parasitol, 1999. **43**: p. 1-46.
81. Turner, C.M., et al., *Human infectivity trait in Trypanosoma brucei: stability, heritability and relationship to sra expression.* Parasitology, 2004. **129**(Pt 4): p. 445-54.
82. Xong, H.V., et al., *A VSG expression site-associated gene confers resistance to human serum in Trypanosoma rhodesiense.* Cell, 1998. **95**(6): p. 839-46.
83. Agbo, E.C., et al., *Measure of molecular diversity within the Trypanosoma brucei subspecies Trypanosoma brucei brucei and Trypanosoma brucei gambiense as revealed by genotypic characterization.* Exp Parasitol, 2001. **99**(3): p. 123-31.
84. Hide, G., et al., *Trypanosoma brucei: identification of trypanosomes with genotypic similarity to human infective isolates in tsetse isolated from a region free of human sleeping sickness.* Exp Parasitol, 2000. **96**(2): p. 67-74.
85. Maclean, L., et al., *Spatially and genetically distinct African Trypanosome virulence variants defined by host interferon-gamma response.* J Infect Dis, 2007. **196**(11): p. 1620-8.
86. Bailey, J.W., *The Diagnosis of Human African Trypanosomiasis.* PhD Thesis - University of Liverpool, 1995.
87. Ormerod, W., *A comparative study of growth and morphology of strains of Trypanosoma rhodesiense.*, in *Experimental Parasitology.* 1963. p. 374.
88. Godfrey, D.G., et al., *The distribution, relationships and identification of enzymic variants within the subgenus Trypanozoon,* in *Adv Parasitol.* 1990. p. 1-74.
89. Stevens, J.R. and D.G. Godfrey, *Numerical taxonomy of Trypanozoon based on polymorphisms in a reduced range of enzymes.* Parasitology, 1992. **104 Pt 1**: p. 75-86.
90. Jackson, A.P., et al., *The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis.* PLoS Negl Trop Dis, 2010. **4**(4): p. e658.
91. Gibson, W.C. and J.K. Gashumba, *Isoenzyme characterization of some Trypanozoon stocks from a recent trypanosomiasis epidemic in Uganda.* Trans R Soc Trop Med Hyg, 1983. **77**(1): p. 114-8.
92. Enyaru, J.C., et al., *Isoenzyme comparison of Trypanozoon isolates from two sleeping sickness areas of south-eastern Uganda.* Acta Trop, 1993. **55**(3): p. 97-115.
93. Simo, G., et al., *Population genetic structure of Central African Trypanosoma brucei gambiense isolates using microsatellite DNA markers.* Infect Genet Evol, 2010. **10**(1): p. 68-76.
94. MacLean, L., et al., *Severity of human african trypanosomiasis in East Africa is associated with geographic location, parasite genotype, and host inflammatory cytokine response profile.* Infect Immun, 2004. **72**(12): p. 7040-4.
95. Gibson, W.C., *Will the real Trypanosoma b. gambiense please stand up.* Parasitol Today, 1986. **2**(9): p. 255-7.

96. Koerner, T., P. De Raadt, and I. Maudlin, *The 1901 uganda sleeping sickness epidemic revisited: a case of mistaken identity?* Parasitol Today, 1995. **11**(8): p. 303-6.
97. Hide, G. and A. Tait, *Molecular epidemiology of African sleeping sickness.* Parasitology, 2009. **136**(12): p. 1491-500.
98. Gibson, W.C., *Analysis of a genetic cross between Trypanosoma brucei rhodesiense and T. b. brucei.* Parasitology, 1989. **99 Pt 3**: p. 391-402.
99. MacLeod, A., C.M. Turner, and A. Tait, *The detection of geographical substructuring of Trypanosoma brucei populations by the analysis of minisatellite polymorphisms.* Parasitology, 2001. **123**(Pt 5): p. 475-82.
100. Hill, E.W., et al., *Understanding bovine trypanosomiasis and trypanotolerance: the promise of functional genomics,* in *Vet Immunol Immunopathol.* 2005. p. 247-58.
101. Murray, M. and S.J. Black, *African trypanosomiasis in cattle: working with nature's solution.* Vet Parasitol, 1985. **18**(2): p. 167-82.
102. Agyemang, K., et al., *Milk Production Characteristics and Productivity of N'Dama Cattle Kept Under Village Management in The Gambia.* Journal of Dairy Science, 1991. **74**(5): p. 1599-1608.
103. International Laboratory for Research on Animal Disease (ILRAD). *Livestock production in tsetse affected areas of Africa : proceedings of a meeting held in Nairobi, Kenya from the 23rd to 27th November 1987.* 1988, Nairobi: ILCA/ILRAD. 473p.
104. Hanotte, O., et al., *Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7443-8.
105. Morrison, W. and M. Murray, *Trypanosoma congolense: inheritance of susceptibility to infection in inbred strains of mice.* Experimental Parasitology, 1979. **48**: p. 364-74.
106. Morrison, W., et al., *Susceptibility of inbred strains of mice to Trypanosoma congolense: correlation with changes in spleen lymphocyte populations.* Clinical and Experimental Immunology, 1978. **32**(1): p. 25-40.
107. Murray, M., W.I. Morrison, and D.D. Whitelaw, *Host susceptibility to African trypanosomiasis: trypanotolerance.* Adv Parasitol, 1982. **21**: p. 1-68.
108. Antoine-Moussiaux, N., S. Magez, and D. Desmecht, *Contributions of experimental mouse models to the understanding of African trypanosomiasis.* Trends Parasitol, 2008. **24**(9): p. 411-8.
109. Stijlemans, B., et al., *A glycosylphosphatidylinositol-based treatment alleviates trypanosomiasis-associated immunopathology.* J Immunol, 2007. **179**(6): p. 4003-14.
110. Naessens, J., *Bovine trypanotolerance: A natural ability to prevent severe anaemia and haemophagocytic syndrome?* Int J Parasitol, 2006. **36**(5): p. 521-8.
111. Flint, J., et al., *Strategies for mapping and cloning quantitative trait genes in rodents.* Nature Reviews Genetics, 2005. **6**(4): p. 271-286.
112. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
113. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 560-4.
114. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

115. The International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
116. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* J Exp Biol, 2007. **210**(Pt 9): p. 1518-25.
117. Siva, N., *1000 Genomes project.* Nat Biotechnol, 2008. **26**(3): p. 256.
118. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.
119. Thomas, R.K., et al., *Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing.* Nat Med, 2006. **12**(7): p. 852-5.
120. Albert, T., et al., *Direct selection of human genomic loci by microarray hybridization.* Nat Meth, 2007. 4(11): p. 903-905.
121. Godfrey, D.G., et al., *Enzyme polymorphism and the identity of Trypanosoma brucei gambiense.* Parasitology, 1987. **94 (Pt 2)**: p. 337-47.
122. Balmer, O., et al., *Phylogeography and Taxonomy of Trypanosoma brucei.* PLoS Negl Trop Dis, 2011. 5(2): p. e961.
123. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.
124. Evanno, G., S. Regnaut, and J. Goudet, *Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.* Mol Ecol, 2005. **14**(8): p. 2611-20.
125. Corander, J., P. Waldmann, and M.J. Sillanpaa, *Bayesian analysis of genetic differentiation between populations.* Genetics, 2003. **163**(1): p. 367-74.
126. Kemp, S.J., et al., *Localization of genes controlling resistance to trypanosomiasis in mice.* Nature Genetics, 1997. **16**(2): p. 194-196.
127. Darvasi, A. and M. Soller, *Advanced intercross lines, an experimental population for fine genetic mapping.* Genetics, 1995. **141**(3): p. 1199-207.
128. Iraqi, F., et al., *Fine mapping of trypanosomiasis resistance loci in murine advanced intercross lines.* Mammalian Genome, 2000. **11**(8): p. 645-648.
129. Nganga, J.K., M. Soller, and F.A. Iraqi, *High resolution mapping of trypanosomosis resistance loci Tir2 and Tir3 using F12 advanced intercross lines with major locus Tir1 fixed for the susceptible allele.* BMC Genomics, 2010. **11**: p. 394.
130. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
131. Frazer, K.A., et al., *A sequence-based variation map of 8.27 million SNPs in inbred mouse strains.* Nature, 2007. **448**(7157): p. 1050-3.
132. Wade, C.M., et al., *The mosaic structure of variation in the laboratory mouse genome.* Nature, 2002. **420**(6915): p. 574-8.
133. Park, Y.G., et al., *Multiple cross and inbred strain haplotype mapping of complex-trait candidate genes.* Genome Res, 2003. **13**(1): p. 118-21.
134. Wang, X., et al., *Haplotype analysis in multiple crosses to identify a QTL gene.* Genome Res, 2004. **14**(9): p. 1767-72.
135. Sudbery, I., et al., *Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels.* Genome Biol, 2009. **10**(10): p. R112.
136. Bonfield, J.K., K. Smith, and R. Staden, *A new DNA sequence assembly program.* Nucleic Acids Res, 1995. **23**(24): p. 4992-9.

137. Darvasi, A. and M. Soller, *Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus.* TAG Theoretical and Applied Genetics, 1992.
138. Poltorak, A., et al., *Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene.* Science, 1998. **282**(5396): p. 2085-8.
139. Kramer, A., *The structure and function of proteins involved in mammalian pre-mRNA splicing.* Annu Rev Biochem, 1996. **65**: p. 367-409.
140. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
141. Crawford, G.E., et al., *DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays.* Nat Methods, 2006. **3**(7): p. 503-9.
142. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.
143. Ramensky, V., P. Bork, and S. Sunyaev, *Human non-synonymous SNPs: server and survey.* Nucleic Acids Res, 2002. **30**(17): p. 3894-900.
144. Oliveira, A.C., et al., *Impaired innate immunity in Tlr4(-/-) mice but preserved CD8+ T cell responses against Trypanosoma cruzi in Tlr4-, Tlr2-, Tlr9- or Myd88-deficient mice.* PLoS Pathog. **6**(4): p. e1000870.
145. Alafiatayo, R.A., et al., *Endotoxins and the pathogenesis of Trypanosoma brucei brucei infection in mice.* Parasitology, 1993. **107 (Pt 1)**: p. 49-53.
146. Graubert, T., et al., *A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome.* PLoS Genet, 2007. **3**(1): p. e3.
147. Mural, R.J., et al., *A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome.* Science, 2002. **296**(5573): p. 1661-71.
148. Yang, H., et al., *On the subspecific origin of the laboratory mouse.* Nat Genet, 2007. **39**(9): p. 1100-7.
149. Goodhead, I., et al., *A Comprehensive Genetic Analysis of Candidate Genes Regulating Response to Trypanosoma congolense Infection in Mice.* PLoS Negl Trop Dis, 2010. **4**(11): p. e880.
150. Rathkolb, B., et al., *Clinical chemistry of congenic mice with quantitative trait loci for predicted responses to Trypanosoma congolense infection.* Infection and Immunity, 2009. 77(9): p. 3948-57.
151. Araki, M., et al., *Genetic evidence that the differential expression of the ligand-independent isoform of CTLA-4 is the molecular basis of the Idd5.1 type 1 diabetes region in nonobese diabetic mice.* The Journal of Immunology, 2009. **183**(8): p. 5146-57.
152. Noyes, H.A., et al., *Mechanisms controlling anaemia in Trypanosoma congolense infected mice.* PLoS One, 2009. **4**(4): p. e5170.
153. Xie, C. and M.T. Tammi, *CNV-seq, a new method to detect copy number variation using high-throughput sequencing.* BMC Bioinformatics, 2009. **10**: p. 80.
154. Nishant, K.T., et al., *The Baker's Yeast Diploid Genome Is Remarkably Stable in Vegetative Growth and Meiosis.* PLoS Genet. **6**(9): p. e1001109.
155. Denis, F.M., et al., *PRAM-1 potentiates arsenic trioxide-induced JNK activation.* J Biol Chem, 2005. **280**(10): p. 9043-8.
156. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proceedings of the National Academy of Sciences, 2004. **101**(16): p. 6062-7.

157. Moog-Lutz, C., et al., *PRAM-1 is a novel adaptor protein regulated by retinoic acid (RA) and promyelocytic leukemia (PML)-RA receptor alpha in acute promyelocytic leukemia cells.* J Biol Chem, 2001. **276**(25): p. 22375-81.
158. Clemens, R.A., et al., *PRAM-1 is required for optimal integrin-dependent neutrophil function.* Mol Cell Biol, 2004. **24**(24): p. 10923-32.
159. Uzonna, J.E., et al., *Experimental murine Trypanosoma congolense infections. I. Administration of anti-IFN-gamma antibodies alters trypanosome-susceptible mice to a resistant-like phenotype.* J Immunol, 1998. **161**(10): p. 5507-15.
160. Aoki, T., et al., *Expression profiling of genes related to asthma exacerbations.* Clin Exp Allergy, 2009. **39**(2): p. 213-21.
161. Loke, P., et al., *Gene expression patterns of dengue virus-infected children from nicaragua reveal a distinct signature of increased metabolism.* PLoS Negl Trop Dis. **4**(6): p. e710.
162. Gahmberg, C.G., et al., *Leukocyte integrins and inflammation.* Cell Mol Life Sci, 1998. **54**(6): p. 549-55.
163. Post, G.R., et al., *Guanine nucleotide exchange factor-like factor (Rlf) induces gene expression and potentiates alpha 1-adrenergic receptor-induced transcriptional responses in neonatal rat ventricular myocytes.* Journal of Biological Chemistry, 2002. **277**(18): p. 15286-92.
164. Fisher, P., et al., *A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis.* Nucleic Acids Research, 2007. **35**(16): p. 5625-33.
165. Noyes, H.A., et al., *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, 2010. **11**(1): p. 361.
166. Behnke, J.M., et al., *Quantitative trait loci for resistance to Heligmosomoides bakeri and associated immunological and pathological traits in mice: comparison of loci on chromosomes 5, 8 and 11 in F2 and F6/7 inter-cross lines of mice.* Parasitology, 2010. **137**(2): p. 311-20.
167. Tsang, S., et al., *A comprehensive SNP-based genetic analysis of inbred mouse strains.* Mamm Genome, 2005. **16**(7): p. 476-80.
168. Roberts, L.J., et al., *Resistance to Leishmania major is linked to the H2 region on chromosome 17 and to chromosome 9.* J Exp Med, 1997. **185**(9): p. 1705-10.
169. Burt, R.A., et al., *Temporal expression of an H2-linked locus in host response to mouse malaria.* Immunogenetics, 1999. **50**(5-6): p. 278-85.
170. Iraqi, F.A., et al., *Chromosomal regions controlling resistance to gastro-intestinal nematode infections in mice.* Mamm Genome, 2003. **14**(3): p. 184-91.
171. Bagot, S., et al., *Identification of two cerebral malaria resistance loci using an inbred wild-derived mouse strain.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9919-23.
172. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
173. Adams, D.J., et al., *Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains.* Nat Genet, 2005. **37**(5): p. 532-6.
174. Cutler, G., et al., *Significant gene content variation characterizes the genomes of inbred mouse strains.* Genome Res, 2007. **17**(12): p. 1743-54.

175. Li, J., et al., *Genomic segmental polymorphisms in inbred mouse strains.* Nat Genet, 2004. **36**(9): p. 952-4.
176. Cho, E.K., et al., *Array-based comparative genomic hybridization and copy number variation in cancer research.* Cytogenet Genome Res, 2006. **115**(3-4): p. 262-72.
177. Coe, B.P., et al., *Resolving the resolution of array CGH.* Genomics, 2007. **89**(5): p. 647-53.
178. Rennie, C., et al., *Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays.* BMC Genomics, 2008. **9**: p. 317.
179. Lipson, D., et al., *Efficient calculation of interval scores for DNA copy number data analysis.* J Comput Biol, 2006. **13**(2): p. 215-28.
180. *Expression of mouse genes after T. congolense infection.* Available from: http://www.genomics.liv.ac.uk/tryps/GeneExpressionViewer/ExpressionForm.V1.html.
181. Hovatta, I., et al., *Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice.* Nature, 2005. **438**(7068): p. 662-6.
182. Nozawa, M. and M. Nei, *Genomic drift and copy number variation of chemosensory receptor genes in humans and mice.* Cytogenet Genome Res, 2008. **123**(1-4): p. 263-9.
183. Assarsson, E., et al., *NK cells stimulate proliferation of T and NK cells through 2B4/CD48 interactions.* J Immunol, 2004. **173**(1): p. 174-80.
184. Scharton-Kersten, T.M. and A. Sher, *Role of natural killer cells in innate resistance to protozoan infections.* Curr Opin Immunol, 1997. **9**(1): p. 44-51.
185. Harty, J.T., A.R. Tvinnereim, and D.W. White, *CD8+ T cell effector mechanisms in resistance to infection.* Annu Rev Immunol, 2000. **18**: p. 275-308.
186. Kierstein, S., et al., *Gene expression profiling in a mouse model for African trypanosomiasis.* Genes Immun, 2006. **7**(8): p. 667-79.
187. Noyes, H., et al., *Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection.* Proc Natl Acad Sci U S A, 2011. **108**(22): p. 9304-9.
188. Anderson, S.K., et al., *Complete elucidation of a minimal class I MHC natural killer cell receptor haplotype.* Genes Immun, 2005. **6**(6): p. 481-92.
189. Beebe, A.M., et al., *Serial backcross mapping of multiple loci associated with resistance to Leishmania major in mice.* Immunity, 1997. **6**(5): p. 551-7.
190. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
191. Mardis, E.R., *ChIP-seq: welcome to the new frontier.* Nat Methods, 2007. **4**(8): p. 613-4.
192. Hide, G., *History of sleeping sickness in East Africa.* Clin Microbiol Rev, 1999. **12**(1): p. 112-25.
193. Ormerod, W.E., *The study of volutin granules in trypanosomes.* Trans R Soc Trop Med Hyg, 1961. **55**: p. 313-32.
194. Odiit, M., F. Kansiime, and J.C. Enyaru, *Duration of symptoms and case fatality of sleeping sickness caused by Trypanosoma brucei rhodesiense in Tororo, Uganda.* East Afr Med J, 1997. **74**(12): p. 792-5.
195. Morrison, L.J., et al., *Role for parasite genetic diversity in differential host responses to Trypanosoma brucei infection.* Infect Immun, 2010. **78**(3): p. 1096-108.

196. Morrison, L.J., et al., *A major genetic locus in Trypanosoma brucei is a determinant of host pathology.* PLoS Negl Trop Dis, 2009. **3**(12): p. e557.
197. MacLean, L.M., et al., *Focus-specific clinical profiles in human African Trypanosomiasis caused by Trypanosoma brucei rhodesiense.* PLoS Negl Trop Dis, 2010. **4**(12): p. e906.
198. Holzmuller, P., et al., *Virulence and pathogenicity patterns of Trypanosoma brucei gambiense field isolates in experimentally infected mouse: differences in host immune response modulation by secretome and proteomics*, in *Microbes Infect.* 2008. p. 79-86.
199. Chisi, J.E., et al., *Anaemia in human African trypanosomiasis caused by Trypanosoma brucei rhodesiense.* East Afr Med J, 2004. **81**(10): p. 505-8.
200. Smith, D.H. and J.W. Bailey, *Human African trypanosomiasis in south-eastern Uganda: clinical diversity and isoenzyme profiles.* Ann Trop Med Parasitol, 1997. **91**(7): p. 851-6.
201. Stevens, J.R., et al., *A simplified method for identifying subspecies and strain groups in Trypanozoon by isoenzymes.* Ann Trop Med Parasitol, 1992. **86**(1): p. 9-28.
202. Ormerod, W.E., *A comparative study of cytoplasmic inclusions (volutin granules) in different species of trypanosomes.* J Gen Microbiol, 1958. **19**(2): p. 271-88.
203. Ormerod, W.E., *A comparative study of growth and morphology of strains of Trypanosoma rhodesiense.* Exp Parasitol, 1963. **13**: p. 374-85.
204. Levine, R.F. and J.M. Mansfield, *Genetics of resistance to African trypanosomes: role of the H-2 locus in determining resistance to infection with Trypanosoma rhodesiense.* Infect Immun, 1981. **34**(2): p. 513-8.
205. Hertz, C., H. Filutowicz, and J. Mansfield, *Resistance to the African Trypanosomes Is IFN-γ Dependent*, in *The Journal of Immunology.* 1998.
206. Greenblatt, H.C. and D.L. Rosenstreich, *Trypanosoma rhodesiense infection in mice: sex dependence of resistance.* Infect Immun, 1984. **43**(1): p. 337-40.
207. Greenblatt, H.C., C.L. Diggs, and D.L. Rosenstreich, *Trypanosoma rhodesiense: analysis of the genetic control of resistance among mice.* Infect Immun, 1984. **44**(1): p. 107-11.
208. Markel, P., et al., *Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains.* Nat Genet, 1997. **17**(3): p. 280-4.
209. Rapp, J.P., M.R. Garrett, and A.Y. Deng, *Construction of a double congenic strain to prove an epistatic interaction on blood pressure between rat chromosomes 2 and 10.* J Clin Invest, 1998. **101**(8): p. 1591-5.
210. Beament, T., *Investigation into differences in pathogenesis of human isolates of African Trypanosomiasis in mice*, in *Thesis (Unpublished).* 2002, University of Liverpool: Liverpool.
211. Morrison, L.J., et al., *Use of multiple displacement amplification to increase the detection and genotyping of trypanosoma species samples immobilized on FTA filters.* Am J Trop Med Hyg, 2007. **76**(6): p. 1132-7.
212. Balmer, O., et al., *Characterization of di-, tri-, and tetranucleotide microsatellite markers with perfect repeats for Trypanosoma brucei and related species.* Mol Ecol Notes, 2006. **6**(2): p. 508-510.
213. Fakhar, M., et al., *An integrated pipeline for the development of novel panels of mapped microsatellite markers for Leishmania donovani complex, Leishmania braziliensis and Leishmania major.* Parasitology, 2008. **135**(5): p. 567-74.

214. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.
215. Jaccard, P., *Nouvelles recherches sur la distribution florale.* Bulletin de la Soci√®te Vaudense des Sciences Naturelles, 1908. **44**: p. 223-270.
216. Pemberton, J.M., et al., *Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies.* Mol Ecol, 1995. **4**(2): p. 249-52.
217. Kuhls, K., et al., *Multilocus microsatellite typing (MLMT) reveals genetically isolated populations between and within the main endemic regions of visceral leishmaniasis.* Microbes Infect, 2007. **9**(3): p. 334-43.
218. Jamjoom, M.B., et al., *Leishmania donovani is the only cause of visceral leishmaniasis in East Africa; previous descriptions of L. infantum and "L. archibaldi" from this region are a consequence of convergent evolution in the isoenzyme data.* Parasitology, 2004. **129**(Pt 4): p. 399-409.
219. Gibson, W., *Will the real Trypanosoma brucei rhodesiense please step forward?* Trends Parasitol, 2002. **18**(11): p. 486-90.
220. Morrison, L.J., et al., *Trypanosoma brucei gambiense Type 1 populations from human patients are clonal and display geographical genetic differentiation.* Infect Genet Evol, 2008. **8**(6): p. 847-54.
221. Gibson, W., *Resolution of the species problem in African trypanosomes.* Int J Parasitol, 2007. **37**(8-9): p. 829-38.
222. Truc, P., et al., *Confirmation of two distinct classes of zymodemes of Trypanosoma brucei infecting man and wild mammals in Cote d'Ivoire: suspected difference in pathogenicity.* Ann Trop Med Parasitol, 1997. **91**(8): p. 951-6.
223. Jamonneau, V., et al., *Characterization of Trypanosoma brucei s.l. infecting asymptomatic sleeping-sickness patients in Cote d'Ivoire: a new genetic group?* Ann Trop Med Parasitol, 2004. **98**(4): p. 329-37.
224. Apted, F.I., et al., *A comparative study of the epidemiology of endemic Rhodesian sleeping sickness in different parts of Africa.* J Trop Med Hyg, 1963. **66**: p. 1-16.
225. Hide, G., et al., *Trypanosoma brucei rhodesiense: characterisation of stocks from Zambia, Kenya, and Uganda using repetitive DNA probes.* Exp Parasitol, 1991. **72**(4): p. 430-9.
226. Stevens, J.R. and S.C. Welburn, *Genetic processes within an epidemic of sleeping sickness in Uganda.* Parasitol Res, 1993. **79**(5): p. 421-7.
227. van Grinsven, K.W., et al., *Adaptations in the glucose metabolism of procyclic Trypanosoma brucei isolates from tsetse flies and during differentiation of bloodstream forms.* Eukaryot Cell, 2009. **8**(8): p. 1307-11.
228. Seed, J. and M. Wenck, *Role of the long slender to short stumpy transition in the life cycle of the african trypanosomes*, in *Kinetoplastid Biol Dis.* 2003. p. 3.
229. Reuner, B., et al., *Cell density triggers slender to stumpy differentiation of Trypanosoma brucei bloodstream forms in culture.* Mol Biochem Parasitol, 1997. **90**(1): p. 269-80.
230. Vassella, E., et al., *Differentiation of African trypanosomes is controlled by a density sensing mechanism which signals cell cycle arrest via the cAMP pathway.* J Cell Sci, 1997. **110 (Pt 21)**: p. 2661-71.
231. Sendashonga, C.N. and S.J. Black, *Analysis of B cell and T cell proliferative responses induced by monomorphic and pleomorphic Trypanosoma brucei parasites in mice.* Parasite Immunol, 1986. **8**(5): p. 443-53.

232. Jensen, B.C., et al., *Widespread variation in transcript abundance within and across developmental stages of Trypanosoma brucei.* BMC Genomics, 2009. **10**: p. 482.
233. Brun, R. and Schonenberger, *Cultivation and in vitro cloning or procyclic culture forms of Trypanosoma brucei in a semi-defined medium. Short communication.* Acta Trop, 1979. **36**(3): p. 289-92.
234. Konieczny, A. and F.M. Ausubel, *A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers.* Plant J, 1993. **4**(2): p. 403-10.
235. Thiel, T., et al., *SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development.* Nucleic Acids Res, 2004. **32**(1): p. e5.
236. Sambrook, J. and D.W. Russell, *Molecular cloning : a laboratory manual.* 3rd ed. 2001, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.
237. Sigrist, C.J., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D161-6.
238. Domenicali Pfister, D., et al., *A Mitogen-Activated Protein Kinase Controls Differentiation of Bloodstream Forms of Trypanosoma brucei.* Eukaryotic Cell, 2006. **5**(7): p. 1126-1135.
239. Huson, D.H., *SplitsTree: analyzing and visualizing evolutionary data.* Bioinformatics, 1998. **14**(1): p. 68-73.
240. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
241. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
242. Branche, C., et al., *Conserved and specific functions of axoneme components in trypanosome motility.* J Cell Sci, 2006. **119**(Pt 16): p. 3443-55.
243. Cronin, C.N., D.P. Nolan, and H.P. Voorheis, *The enzymes of the classical pentose phosphate pathway display differential activities in procyclic and bloodstream forms of Trypanosoma brucei.* FEBS Lett, 1989. **244**(1): p. 26-30.
244. Alsford, S., et al., *High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome.* Genome Res, 2011.
245. DeVinney, R., O. Steele-Mortimer, and B.B. Finlay, *Phosphatases and kinases delivered to the host cell by bacterial pathogens.* Trends Microbiol, 2000. **8**(1): p. 29-33.
246. Saeij, J.P., et al., *Polymorphic secreted kinases are key virulence factors in toxoplasmosis.* Science, 2006. **314**(5806): p. 1780-3.
247. Parsons, M., M. Valentine, and V. Carter, *Protein kinases in divergent eukaryotes: identification of protein kinase activities regulated during trypanosome development.* Proc Natl Acad Sci U S A, 1993. **90**(7): p. 2656-60.
248. Garcia-Salcedo, J.A., et al., *A protein kinase specifically associated with proliferative forms of Trypanosoma brucei is functionally related to a yeast kinase involved in the co-ordination of cell shape and division.* Mol Microbiol, 2002. **45**(2): p. 307-19.
249. Szoor, B., et al., *Protein tyrosine phosphatase TbPTP1: A molecular switch controlling life cycle differentiation in trypanosomes.* J Cell Biol, 2006. **175**(2): p. 293-303.
250. Picozzi, K., *Sleeping sickness in Uganda: a thin line between two fatal diseases,* in *BMJ*. 2005. p. 1238-1241.

251. MacLeod, A., et al., *The genetic map and comparative analysis with the physical map of Trypanosoma brucei*, in *Nucleic Acids Research*. 2005.
252. Cooper, A., et al., *Genetic analysis of the human infective trypanosome Trypanosoma brucei gambiense: chromosomal segregation, crossing over, and the construction of a genetic map*. Genome Biol, 2008. **9**(6): p. R103.
253. Tait, A., et al., *Self-fertilisation in Trypanosoma brucei*. Mol Biochem Parasitol, 1996. **76**(1-2): p. 31-42.
254. Black, S.J., R.S. Hewett, and C.N. Sendashonga, *Trypanosoma brucei variable surface antigen is released by degenerating parasites but not by actively dividing parasites*. Parasite Immunol, 1982. **4**(4): p. 233-44.
255. Shapiro, S.Z., et al., *Analysis by flow cytometry of DNA synthesis during the life cycle of African trypanosomes*. Acta Trop, 1984. **41**(4): p. 313-23.
256. Kammer, G.M., *The adenylate cyclase-cAMP-protein kinase A pathway and regulation of the immune response*. Immunol Today, 1988. **9**(7-8): p. 222-9.
257. Dean, S., et al., *A surface transporter family conveys the trypanosome differentiation signal*. Nature, 2009. **459**(7244): p. 213-7.
258. Black, S.J., et al., *Regulation of the growth and differentiation of Trypanosoma (Trypanozoon) brucei brucei in resistant (C57Bl/6) and susceptible (C3H/He) mice*. Parasite Immunol, 1983. **5**(5): p. 465-78.
259. Veillette, A. and S. Latour, *The SLAM family of immune-cell receptors*. Curr Opin Immunol, 2003. **15**(3): p. 277-85.
260. Korstanje, R. and B. Paigen, *From QTL to gene: the harvest begins*. Nat Genet, 2002. **31**(3): p. 235-6.
261. Verdugo, R.A., et al., *Serious limitations of the QTL/microarray approach for QTL gene discovery*. BMC Biol, 2010. **8**: p. 96.
262. Masocha, W., et al., *Cerebral vessel laminins and IFN-gamma define Trypanosoma brucei brucei penetration of the blood-brain barrier*. J Clin Invest, 2004. **114**(5): p. 689-94.
263. Bentivoglio, M., et al., *Sleep and timekeeping changes, and dysregulation of the biological clock in experimental trypanosomiasis*. Bull Soc Pathol Exot, 1994. **87**(5): p. 372-5.
264. Naessens, J., et al., *TNF-alpha mediates the development of anaemia in a murine Trypanosoma brucei rhodesiense infection, but not the anaemia associated with a murine Trypanosoma congolense infection*. Clin Exp Immunol, 2005. **139**(3): p. 405-10.
265. Courtin, D., et al., *Comparison of cytokine plasma levels in human African trypanosomiasis*. Trop Med Int Health, 2006. **11**(5): p. 647-53.
266. Nakamura, Y., et al., *Susceptibility of heat shock protein 70.1-deficient C57BL/6 J, wild-type C57BL/6 J and A/J mice to Trypanosoma congolense infection*. Parasitol Res, 2003. **90**(2): p. 171-4.
267. Shi, M., W. Pan, and H. Tabel, *Experimental African trypanosomiasis: IFN-gamma mediates early mortality*. Eur J Immunol, 2003. **33**(1): p. 108-18.
268. Main, B.J., et al., *Allele-specific expression assays using Solexa*. BMC Genomics, 2009. **10**: p. 422.
269. Kolev, N.G., et al., *The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution*. PLoS Pathog, 2010. **6**(9).
270. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res, 2005. **15**(8): p. 1034-50.

271. Reuveni, E., V.E. Ramensky, and C. Gross, *Mouse SNP Miner: an annotated database of mouse functional single nucleotide polymorphisms.* BMC Genomics, 2007. **8**: p. 24.
272. Cervino, A.C., et al., *A comprehensive mouse IBD database for the efficient localization of quantitative trait loci.* Mamm Genome, 2006. **17**(6): p. 565-74.
273. Pentreath, V.W., et al., *Endotoxin antibodies in African sleeping sickness,* in *Parasitology.* 1997. p. 361-5.
274. Darji, A., et al., *In vitro simulation of immunosuppression caused by Trypanosoma brucei: active involvement of gamma interferon and tumor necrosis factor in the pathway of suppression.* Infect Immun, 1996. **64**(6): p. 1937-43.
275. Bakhiet, M., et al., *A Trypanosoma brucei brucei-derived factor that triggers CD8+ lymphocytes to interferon-γ secretion ...,* in *Scandinavian Journal of Immunology.* 1993.
276. Darji, A., et al., *Mechanisms underlying trypanosome-elicited immunosuppression.* Ann Soc Belg Med Trop, 1992. **72 Suppl 1**: p. 27-38.

Appendices

Index of Appendix Tables and Figures

Appendix I: C3H/HeJ x C57BL/6 F2 Genotyping

Appendix Tables A1: Microsatellite and SNP genotyping primers.
List of markers and primers used to genotype F2 crosses for resistant (C57BL/6) or susceptible (C3H/HeJ) alleles at three trypanotolerance loci and at the *Tlr4* gene

Table A1.1.1: *Tir1*

Marker Name / rsID	Chr	Position (bp)	Primer Sequences (5'-3')
rs13465576	17	32688481	Primer 1: GGCTGCTTTCTGAGTCCAAG Primer 2: GAACAGGGAAAATGGCTGAA
D17mit93	17	74149996 – 74150148	Primer 1: TGTCCTTCGAGTGTTTGTGTG Primer 2: TCCCCGGTGAATGAGTTATC
D17Mit68	17	47707105 – 47707234	Primer 1: GTCCTGACATCATGCTTTGTG Primer 2: CTACCGTTTGGAAGGCTGAG
D17Mit184	17	67915804 – 67915940	Primer 1: TGCACTACCCAAACATGCAT Primer 2: ACTTCTGACAGGAAGCATCCA
D17mit155	17	84900959 – 84901098	Primer 1: TGAGAAGGTTGGGTTTATATATTTAGG Primer 2: CGATCATTTCCTTGCAACCT
D17mit117	17	50270475 – 50270596	Primer 1: AGTCCATTTATCGGGGGC Primer 2: TTTAATGGCACATCTGGCAA

Table A1.1.2: ***Tir2***

Marker Name / rsID	Chr	Position	Primer Sequences (5'-3')
rs46746692	5	75362301	Primer 1: GATCTGGGGCAGCTCTTGA Primer 2: CATTTTACAGCAGGGTATTATGG
rs46742668	5	77357797	Primer 1: ACGGTTAGCAGAGGAGGATG Primer 2: TGTTGTTGTTGTGTGTTTTGTTTT
rs47415520	5	84510178	Primer 1: CATCCTGATTGGTCATCTCC Primer 2: TTTAGGGAGGCAAAATTCCA
rs31694652	5	87889493	Primer 1: GACCTGAGGTGTCTTTTTCTTCA Primer 2: CCTCAGCTGGTTTCAGTACCA
D5mit81	5	50722564 – 50722773	Primer 1: GGGAGTTCCAGGTTCATTGA Primer 2: TGTGCATTATGGCATGTAAATG
D5mit255	5	55345578 – 55345695	Primer 1: CCCTGTGCTCTGGATTAGTTG Primer 2: TCAAGACCAGCATCAAACCA
D5mit201	5	75550712 – 75550821	Primer 1: GAGGACTCCTTCGATTTCCC Primer 2: TTCCTAAGCAGGAACTGACCA
D5mit169	5	150226324 – 150226436	Primer 1: CCAGGTCTCCAGGGTTGTAA Primer 2: CTCCTGAGGGAACGAGTCAG

Table A1.1.3: *Tir3*

Marker Name / rsID	Chr	Position	Primer Sequences (5'-3')
D1mit215	1	78202934 – 78203082	Primer 1: GGAGCAGAGTGTGAGAAGGG Primer 2: CCAGTGTGAGCCCATTCC
D1mit155	1	196255163 – 196255414	Primer 1: ATGCATGCATGCACACGT Primer 2: ACCGTGAAATGTTCACCCAT
D1mit94	1	128076848 – 128077001	Primer 1: CGACTTCCCTTGATGTCCAT Primer 2: TTTGTGTTGTGCAGTCTGTCTG
D1mit425	1	158554749 – 158554869	Primer 1: CAAAAAAACAACACATTTTACTTTCA Primer 2: ACTTTGTATTTCACATGATGTCCTG

Table A1.1.4: *Tlr4*

Marker Name / rsID	Chr	Position	Primer Sequences (5'-3')
rs3023006	4	66502140	Primer 1: GGACTGGGTGAGAAATGAGC Primer 2: GAAACTGCCATGTTTGAGCA
D4mit178	4	66843059 – 66843205	Primer 1: GCCCTGAAGGTAAATCAGTAACT Primer 2: GCTCAGGAGGTACATTGCCT

Table A1.1.5: *U4/U6*

Marker Name / rsID	Chr	Position	Primer Sequences (5'-3')
D5Mit113 (U4)	5	77684240 – 77684338	Primer 1: ACAGTATTTTCTTTTTCCAAGTGTG Primer 2: CAAAGACTCTAGGTGTGACCCC
D5Mit10 (U4)	5	104668024 – 104668218	Primer 1: CGAGAAGTTGGAAAGACCCA Primer 2: GGCACCCATGCCTCTATG
D1Mit102 (U6)	1	149096650 – 149096762	Primer 1: AAATACCAGCAAAACAATAAAGGC Primer 2: TGAATTAAAATTGCAGAGGCG
D1mit425 (U6)	1	158554749 – 158554869	Primer 1: CAAAAAAACAACACATTTTACTTTCA Primer 2: ACTTTGTATTTCACATGATGTCCTG

Appendix Table A1.2.1: Mouse Survival Times

Survival for C3H/HeJ x C57BL/6 F2 crossed mice with the most extreme survival times after infection with *T. congolense* strain IL1180. Mice 95 and 96 are the parental mice (95 = C3H/HeJ; 96 = C57BL/6).

Mouse	Sex	Survival (Days)	Mouse	Sex	Survival (Days)	Mouse	Sex	Survival (Days)	Mouse	Sex	Survival (Days)	Mouse	Sex	Survival (Days)
1	F	38	21	M	141	41	M	128	61	M	29	81	F	141
2	F	141	22	F	138	42	M	128	62	F	58	82	F	141
3	M	128	23	M	62	43	M	57	63	M	133	83	M	141
4	F	61	24	M	36	44	F	49	64	F	130	84	M	58
5	M	60	25	F	56	45	F	141	65	M	122	85	F	134
6	F	124	26	F	141	46	M	123	66	M	127	86	M	62
7	M	130	27	M	133	47	F	141	67	F	43	87	F	54
8	F	46	28	M	141	48	M	141	68	F	60	88	F	141
9	M	141	29	F	15	49	F	36	69	M	133	89	F	141
10	F	50	30	M	62	50	M	141	70	M	127	90	M	127
11	F	122	31	F	141	51	M	59	71	M	60	91	M	141
12	M	132	32	F	128	52	F	141	72	F	132	92	F	62
13	M	121	33	F	123	53	M	132	73	M	59	93	M	141
14	M	43	34	F	141	54	M	141	74	F	139	94	M	141
15	M	141	35	M	15	55	M	141	75	F	141	95	C3H/HeJ parental	
16	F	49	36	M	141	56	M	53	76	M	141	96	C57BL/6 parental	
17	F	57	37	F	141	57	M	141	77	M	125			
18	F	141	38	M	141	58	F	54	78	M	141			
19	M	122	39	F	132	59	M	141	79	M	49			
20	F	128	40	F	121	60	M	123	80	F	60			

Appendix Table A1.3.1: Genotyping data for *Tir1*

Genotyping data for C3H/HeJ x C57BL/6 F2 crossed mice with the most extreme survival times after infection with *T. congolense* strain IL1180. Microsatellite (D17mit[xxx]) loci are listed as the allele sizes detected by PeakScanner (Applied Biosystems) after PCR and electrophoresis on an ABI-3130xl capillary sequencer. SNP (rs) loci are summarised as:
1 = C3H/HeJ homozygote ; 2 = C57BL6 homozygote ; 3 = heterozygote ; 0 = null allele (no data)

	D17mit117		**D17mit155**		**D17mit184**		**D17mit68**		**D17mit93**		**rs13465576**
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype
1	124				129.22		137		171.42		1
2		119.83	155.9		129.35	133.2		131		157.03	2
3		120	155.9		129.3	133.1		131		156.83	3
4	124	120	156	139.69	129.33	133.22	139	131	171.35	156.83	3
5	125.51	121.33	155.95	139.69	129.37	133.18	139	132.59	171.35	156.79	3
6	124	119	155	139	129.39	133.2	137.85	131.48	170	156.83	3
7		119	155	139		133.19		131.46		156.79	3
8	124	119	155		129.25		137	131	170		3
9		119	155	139		132.3		131.45		156.8	2
10		120.12		139		133.18		131.25		156	2
11	124	120		139.46	129.25	133.17	137.74	131.46	171.26	156.8	3
12		120.28		139.46		133.17		131.7		156.83	2
13	124	119	155	139	129.27	133.14	137.86	131.5	171.44	156.97	3
14	125.58			139.57	129.22	133.07	136.86			156.97	1
15	125.48			139.58	129.25		136.83			156.83	1
16		120		139.57		133.31		131.59		156.9	2
17	124.3		155.86		129.35		137.73		171.26		1
18	124.39	120.27	155.86	139.58	129.37		137.84	131.55	171.32		3
19	124.4	120.29	155.97		129.29	133.13	137.84	131.56	171.36		3
20	124	119		139		133.17	137.84	131.46		156.8	3
21		119		139		133		131		156	2
22		120.29		139.58		133.15		131.56		156.73	2
23	124.38	120.26	155.8		129.28		137.84	131.44	171.28		3

	D17mit117		D17mit155		D17mit184		D17mit68		D17mit93		rs13465576
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype
24	125.55		155.85		129.31		139		171.29		1
25	125	119	155	139		133.13		131.54		156.93	2
26		120	155	139		133.15		131.56	171.46	156.97	3
27		119		139		133.29		131.45		156.79	2
28		119		139		133.2		131.47		156.79	2
29		120.26		139.58		133.26		131.54		156.8	2
30	125.55		155.9		129.34		139		171.31		1
31	125	119	155	139	129.3	133.14	137.73	131.41	170	156	3
32		119	155	139		133.17		131.55	171.23	156.73	2
33	124.34	120.21		139.58		133.15	137.73	131.42		156.76	3
34	124.4	120.29	155.86	139.58	129.35	133.17		131.46	171.37	156.8	2
35	124.34	120.21	155.75	139.58	129.37	133.22	137.73	131.42	171.23	156.69	2
36		120.19		139.58		133.26		131.45		156.8	2
37		119		139.57		133.19		131.58		156.97	2
38		120.33	156	139.69	129.37	133.14		131.65		156.93	2
39	124	120	156		129.32	133.22	137.84	131.47	171.29	156.73	3
40	124	119		139		133.3		131.66		156.83	2
41	124.34	120.23	155.85		129.38		137.74	131.47	171.34		3
42	123	119	155	139	129.34	133.17	137.84	131.46	171.23	156.83	3
43	124.4	120.29		139.58		133.15	137.84	131.46		156.8	3
44	124	119	155	139	129.92		137.73		170	156	1
45	124.27	120.26		139.58	129.33	133.16	137.73	131.44	171.26	156.8	3
46		120	155	139		133.26		131.56			2
47		119	155	139	129.41	133.25		131.46	171.26	156.8	3
48		120.29		139.58		133.25		131.46		156.8	3
49	124		155	139	129.39	133.26	137.75		171.47	156.93	1
50		120.22	156	139.68		133.27		131.44		156.93	2
51		120.18	155.85	139.57	129.4	133.2		131.54	171.32	156.79	2
52		120		139		133.29		131.47		156.72	2

	D17mit117		D17mit155		D17mit184		D17mit68		D17mit93		rs13465576
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype
53	124	119	155	139	129.24	133.17	137.73	131.43	171.27	156.8	1
54		120.14		139.58		133.17		131.53		156.73	2
55		120.29	155.8	139.58		133.26		131.46	171.23	156.73	2
56	124	120	155		129		137.73		170	156	1
57		120.26		139.58		133.26		131.44		156.8	2
58	124.33	120.2		139.58		133.16	137.73	131.41		156.69	3
59		120.26	155.85		129.41	133.25		131.44	171.28		2
60	124.26	120.14		139.47		133.12	137.73	131.43		156.8	3
61	124	119	155		129	133	137.86	131.56	170	156	3
62	124	119	155	139	129.34	133.23	137.85	131.5	171.42	156.93	3
63	124.3	120.19	155.9	139.58	129.34	133.17	137.73	131.45	171.21	156.83	3
64	124	120	155.8	139.46	129.36	133.17	137.73	131.43	171.21	156.73	2
65		120.2		139.57		133.24		131.41		156.8	2
66		120.22	155.86	139.58	129.34	133.17		131.47	171.27	156.69	3
67	125.48		156	139	129.41	133.25	138.9		171.21	156.69	1
68	124.23		155.8		129.34		137.73		171.22		1
69	124.24	120.21		139	129.29	133.13	137.73	131.42	171.21	156.8	3
70	124.34	120.21	155	139	129.4	133.35		131.42	171.28	156.73	2
71	124.24	120.21	155.75	139.58	129.37	133.22		131.63	171.23	156.69	2
72		120.21	155	139		133.25		131.42	171.17	156.73	2
73		120	155	139	129.28	133.25		131.55	171.44	156.97	2
74		120	155	139		133.13		131.54	171.27	156.87	3
75		120	155			133.23		131.44		156.76	3
76	124	120	155.9	139.57	129.39	133.3		131.44	171.2	156.73	2
77	124.29	120.15		139.46		133.19	137.83	131.5		156.73	3
78		120.14		139.57		133.38		131.43		156.79	2
79	124.24	120.21		139.58	129.31	133.25	137.73	131.42	171.27	156.8	3
80	124		155	139	129.34	133.27	137.74	131.45	170	156	3
81	124	120	155	139	129.43	133.26	137.73	131.42	171.19	156.69	3

	D17mit117		D17mit155		D17mit184		D17mit68		D17mit93		rs13465576
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype
82	124	120		139				131.44		156.73	2
83		120	155		129.37	133.22		131.4	171.23	156.7	2
84	124.36	120.23		139.58	129.4	133.24	137.84	131.43	171.07	156.63	3
85	123	119	155	139	129	133	137	131	171.42	156.97	2
86	124.45	120.3	155.95	139.69	129.3	133.16	137.86	131.58	170	156	3
87		120.31	155	139		133.26		131.5		156.79	2
88		119		139		133.22		131.5		156.76	3
89		120.22		139.57		133.18		131.47		156.8	3
90	124	120	155	139	129	133	137.74	131.48	171.29	156.73	3
91	124.25	120.19	155.77		129.36	133.22	137.75	131.43	171.13		3
92	124.37	120.24		139	129.34	133.22	137.84	131.54	171.32	156.76	3
93		120.24	155.87		129.39	133.28		131.52	171.18		2
94		120.14		139.58		133.17		131.43		156.73	2
95	124.25		155.79		129.36		137.75		171.24		1
96		120.21		139.58		133.26		131.46		156.83	2

Appendix Table A1.3.2: *Tir1* genotypes and associated Survival of C3H/HeJ x C57BL/6 infected with *T. congolense* strain IL1180.

Marker	rs13465576	D17mit93	D17Mit68	D17mit117	D17Mit184	D17mit155
Total Number of Mice Genoytyped	96	95	96		95	95
Observed Genotypes at Locus						
Total C3H/HeJ	13	14	12	11	13	19
Total C57BL/6	41	39	48	41	39	36
Total heterozygote	42	42	36	44	43	40
Expected Genotypes at Locus (if in Hardy Weinberg Equilibrium)						
Hardy Weinberg P value	0.00013	0.00079	6.8×10^{-8}	6.1×10^{-5}	0.00057	0.015
Mean Survival (days)						
C3H/HeJ Genotype	62.4	90.1	56.1	57.5	72.6	97.5
C57BL/6 Genotype	116.0	116.9	121.7	123.5	118.7	112.5
Heterozygotes	108.4	100.0	99.9	100.5	102.7	105.1

Appendix Table A1.4.1: Genotyping data for *Tir2*

Genotyping data for C3H/HeJ x C57BL/6 F2 crossed mice with the most extreme survival times after infection with *T. congolense* strain IL1180. Microsatellite (D5mit[xxx]) loci are listed as the allele sizes detected by PeakScanner (Applied Biosystems) after PCR and electrophoresis on an ABI-3130xl capillary sequencer. SNP (rs) loci are summarised as:
1 = C3H/HeJ homozygote ; 2 = C57BL6 homozygote ; 3 = heterozygote ; 0 = null allele (no data)

	D5mit255		D5mit169		D5mit201		D5mit81		rs46746692	rs47415520	rs46742668	rs31694652
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype	Genotype	Genotype	Genotype
1	132.8		116		103.39		193		1	1	0	2
2	132.87		116.91	112.2	103.39	109.49	193	207	3	2	2	2
3	132.86			112.94	103.38		193.66		1	1	1	2
4	132.87		116.92	113.19	103.29		193.6		1	3	0	2
5	132.88		116		103.37		193.64		1	1	1	2
6	132.87		115.74		103.36				1	2	0	0
7	132.87	120.29	115.8		103.37				1	1	1	2
8	132.88	120.22	116.78		103				1	1	0	2
9	132.9	120.38	116.87	112.1	103.3		193.64		1	1	1	2
10	132.76	120.19	116.88	113.34	103.39	109.49	193.74	207.76	3	3	3	0
11	132.74		116.78	113.92	103.4				0	0	3	2
12	132.75		116.92	113.25	103.29		193.65		1	1	1	2
13	132.87		116.86	113.24	103.29		193		1	1	1	2
14	132.83		116.86	113.24	103.3		193.87		1	3	3	2
15	132.84		116.8	113.13	103.39	109.51	193.73	207.66	0	3	3	2
16	132.8	120.31	116.89	108.9	103.3		193.68		0	3	3	2
17	132.81		115.8	108.52	103.26				1	3	0	0
18	132.79	120.23	116.87	113.21	103.37		193.7		1	3	3	2
19	132.74			112.05	103.29	109.39	193.74	207.74	3	2	2	2
20	132.67			113.21	103.29				1	2	2	2
21	132.78	120.21	116.83	112.13	103				1	3	0	0
22	132.75		116.85	113.19	103.38	109.48	193.74	207.72	3	2	2	2
23	132.64		116		103.39	109.51			3	2	2	2
24	132.75			113.31		109.47	193.68	207.67	0	2	2	2
25	132.81		116.86	113.22	103.4		193.75		1	1	0	2

	D5mit255		D5mit169		D5mit201		D5mit81		rs46746692	rs47415520	rs46742668	rs31694652
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype	Genotype	Genotype	Genotype
26	132.77			113.22	103.4				1	3	0	2
27	132.66		116.74	113.89	103.3		193		1	1	0	2
28	132.66		116.83	113.18	103.38		193.7		0	0	2	2
29	132.67	120.12		112	103.3		193.54		0	1	1	2
30	132.66			112	103.39		193.74		1	1	1	2
31	132.77	120.17	116.89	113.12	103.4	109.53			0	0	0	0
32	132.7	120.08	116.82		103.36				1	3	3	2
33	132.8		116.84	113.16	103.39		193.64		1	2	2	2
34	132.77		116.87	113.11	104.28		193.74		1	1	0	2
35	132.75			112	103.3		193.68	207.6	1	3	3	2
36	132.79		116.89	113.12	103.3	109.53		207.74	3	2	2	2
37	132.64		116.86	113.24	103.31		193		1	3	3	2
38	132.74			112		109.46	193.69		2	2	0	2
39	132.72		116.89	113.14	103.38				1	1	1	2
40	132.61	120.07	115.82	108.93	103.29				1	1	1	2
41	132.66		116.83	109.03	103.27		193.72		1	1	1	2
42	132.78			112.06	103.29		193.66		0	0	1	2
43	132.66			113	103.4	109.54	193.59		3	3	3	2
44	132.65		116.82	113.15	103.29				1	1	1	2
45	132.72		116.85		103.31	109.46	193.64	207.64	3	2	2	2
46	132.75		116.92	113.15	103.29				1	3	1	2
47	132.78	123.23	116.84	113.16	103.3				1	2	0	2
48		123	116	112	103.39	109.51		207.65	3	2	2	2
49	132.8			113.22	103.39		193.75		1	1	1	2
50	132.77	120.17	116.92	113.29	103.32		193.67		0	0	0	2
51	132.74		115.82	108.3	103.29		193.77		1	0	0	0
52	132.81		116.83	113.18	103.31				1	2	2	2
53	132.65		116.83		103.28				1	1	1	2
54	132.64	120.03		112	103.28		193.68		1	3	3	2
55	132.75			113.21	103.27	109.44	193.7		3	2	2	2
56	132.69			113.18	103.29		193.7		1	2	2	2
57	132	120		113.18	103.39	109.5	193.7	207.68	3	3	3	2

	D5mit255		D5mit169		D5mit201		D5mit81		rs46746692	rs47415520	rs46742668	rs31694652
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype	Genotype	Genotype	Genotype
58	132.68	120.11	116.83		103.41	109.56	193.74		0	0	0	0
59	132.64			113.16	103.3		193.63		1	3	3	2
60	132.64			113.2	103.31		193.69		1	1	1	2
61	132.66	120.07		113.23	103.31				1	3	0	0
62	132.69		116.77	113.13	103.31		193.77		1	0	0	0
63	132.69			112	103.32		193.64		1	3	3	2
64	132	123	116.83		103.3				1	1	1	2
65	132.69			112.13	103.39		193.64		1	1	1	2
66	132.77			113.05	103.29		193.74		1	1	1	2
67	132.65		116.78	113.11	103.29				0	3	3	2
68	132.66	120.1		113.15	103.29		193.7		1	2	3	2
69	132.77	120.17	116.81	113.12	103.3		193.65		1	3	3	0
70	132.75		116.77	113.08	103.3		193.72		1	3	0	2
71	132.65		116.89	113.2	103.3	109.43	193.69	207.66	3	2	2	2
72	132	123	116.78	113.09	103.32		193		1	2	2	2
73	132.69	120.16	116.87	113.23	103.31	109.47			3	3	3	2
74	132.66	120.07	116.92	113.16	103.22				1	0	0	0
75	132.77		116.83	113.05	103.31				0	0	1	2
76	132.77			112.13	103.31	109.46	193.72	207.73	3	3	3	2
77	132.64	120.13	115.76		103.31		193.7		1	1	0	0
78	132.66		116.84	113.16	103.27		193.59		1	3	3	2
79	132.77		115.76		103.29		193.76		1	1	0	2
80	132.69		116.77		103.33	109.41	193.71		3	2	0	0
81	132.65		116.84	113.16	103.31				1	2	0	0
82	132.77			113.07	104.31		193.74		1	1	0	2
83	132.64			112	103.31		193.73		1	3	0	2
84	132.94		116.8	113.11		109.43	193		2	2	2	2
85	132.88	120.22		112	103.32	109.38	194.79	207.76	3	2	2	2
86	132.8	120.31	116.86	113.22	103.32	109.49	193.83	207.76	3	3	3	2
87	132.75		116.89	113.22	103.31				1	0	0	0
88	132.7	120.08		112.08	103.32	109.48	193.7	207.61	3	2	0	2
89	132.61	120.07	116.83	113.05	104.27		193.72		1	2	2	2

	D5mit255		D5mit169		D5mit201		D5mit81		rs46746692	rs47415520	rs46742668	rs31694652
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	Genotype	Genotype	Genotype	Genotype
90		121.22	116.78		103				1	0	0	0
91	132.69		116.78	113.11	103.29		193.72		1	1	1	2
92	133.76		116.81	113.13	103		193.76		1	1	1	2
93	132.86	120.25		113.16	103.39		193.74		1	0	2	0
94	132.69		116.88		103.31		193.59		1	1	1	2
95	132.75	120.21	116.98		103.31		193.91		1	1	1	2
96		121.34		112		109.4			2	2	2	2

Appendix Table A1.4.2: *Tir2* genotypes and associated Survival of C3H/HeJ x C57BL/6 infected with *T. congolense* strain IL1180.

Marker	rs46746692	rs47415520	rs46742668	rs31694652	D5mit81	D5mit255	D5mit201	D5mit169
Total Number of Mice Genoytyped	84	83	67	79	68	96	96	96
Observed Genotypes at Locus								
Total C3H/HeJ	63	30	25	0	52	63	71	17
Total C57BL/6	3	26	21	79	2	3	4	30
Total heterozygote	18	27	21	0	14	30	21	49
Expected Genotypes at Locus (if in Hardy Weinberg Equilibrium)								
Hardy Weinberg P value	1.5×10^{-22}	0.0044	0.00041	9.9×10^{-44}	1.99×10^{-17}	6.06×10^{-20}	1.25×10^{-27}	0.17
Mean Survival (days)								
C3H/HeJ Genotype	107.5	102.0	109.2	n/a	103.6	103.7	106.4	96.6
C57BL/6 Genotype	99.0	113.2	115.3	107.9	140.0	133.5	78.0	105.1
Heterozygotes	106.9	100.3	98.8	n/a	104.1	108.1	107.5	109.1

Appendix Table A1.5.1: Genotyping data for *Tir3*

Genotyping data for C3H/HeJ x C57BL/6 F2 crossed mice with the most extreme survival times after infection with *T. congolense* strain IL1180. Microsatellite (D1mit[xxx]) loci are listed as the allele sizes detected by PeakScanner (Applied Biosystems) after PCR and electrophoresis on an ABI-3130xl capillary sequencer. SNP (rs) loci are summarised as:
1 = C3H/HeJ homozygote ; 2 = C57BL6 homozygote ; 3 = heterozygote ; 0 = null allele (no data)

	D1mit425		**D1mit94**		**D1mit155**		**D1mit215**	
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6
1	152.24	116.3	210.21	153.46	214.87		157	148
2	152		210.21		214	252	157	148
3		116.33		153.26		252.21		148
4	152.02		210.07		214.84		157	148
5	152.11	116.3	210.03	153		252.2	157.16	148.25
6		116.26		153.26	214	252	157	148
7		116.34		153.25		252.26	157	148
8		116		153.33		252.25	157	148
9	152.08	116.36	210	153.25	214.87	252.25	156.17	148.05
10	152.09		210	153.36	214	252	157.32	
11	152.08	116.31		153.33		252.24	157	148
12	152.17		210	153.25	214.89	252.24	157.21	149.2
13	152.15	116.43	210.14	153.43		252.41	157.42	149.21
14	152.13		210.15		215.03	252.39	157.37	149.38
15	152.02	116.33	210.1	153.28		252.14	157.23	149.28
16	150	116.3	210	153.26		252.3		
17	152		210	154	214.83		157	
18	152.11	116.29	210.04		214.87	252.21	157	
19	152	116.39	210.06	153.26	214.91	252.2		148
20		116.33		153.35	214	252		148
21	152	116	210			252		
22	152.06	116.32	210	153	214.88	252.25	157	148
23	152.08		209.95	153.35	214.76			

	D1mit425		D1mit94		D1mit155		D1mit215	
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6
24	152.17	116.33	209.99	152.21	214.87	252.17	157.12	149.28
25	152.15	116.38		153.36		252.43		148.07
26	152.13	116		153.36	214		157	148
27	151	116	210	153		252	157	148
28	152.02	116.36	210	153.26	214.83		157.28	149.2
29	152.02			153.26	214.84			148
30	152.11	116.29		153.26	214.92	252.19	157	
31	152	116		153.28		252		
32	152	116		153.36		252.18		
33		116.32	209.99			252.2	157	148
34	152.08	116.31	210.02		214.84	252.25		149.21
35	152.09	116.33	209.97	153.25	214.74	252.1	157.29	149.2
36		116.27		153.25	214.81	252.17	157.22	149.2
37	152.13	116.33		153.35		252.39	156.27	149.21
38	152.13		210.17		214.92	252.39	157.35	
39	152.02	116.31	210	153		252.22		
40		116.36	210	153.29		252.29		149.2
41	152.11		210	153		252.2	157	
42	152.09	116.29	210.05	153.25	214.86	252.27	157	147.95
43	152.1		210.01		214.78	252.18	157	149
44		116		153.25		252		
45		116.38		153		252.25	157	149
46	152.08			153.25	214	252	157	
47		152.09		153.32		252		
48		116.36	210	153.25		252.17	157	
49	152.15	116.38	210.12			252.41	156.22	
50	152.11		210.17		214.97	252.38	157.35	
51	152.04	109.14	210.02		214.76		157.32	149.31
52		116.29		153.26		252.28		149

	D1mit425		D1mit94		D1mit155		D1mit215	
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6
53		116		153.25		252.21		
54		116.4		153.26		252.18		148
55	152.09	116.29	210.02	153.23		252.2	157.18	
56	152.09	116	210	153		252.17		
57		116.23	210		214.79	252.18		149.2
58	152.06	116.31	210.01	153		252.24	158	149
59		150		153.25	214.76	252.17	157.22	149.31
60	152.09	116.33		153.25	214.78	252.17	157	148
61		116		153.36		252	157	149
62	152.13	116.34	210.13			252.3	157.35	
63	152.02	116.34		153.26		252.12		148.05
64		116.25	209.97	153		252.12		149.2
65	152.09	116.25	210	153.26	214.81	252.19	157.22	149.2
66		116.25		153.26	214.83	252.18	157	149.2
67	152.09	116.34	209.92	153.15	214.8	252.18	157	
68	152.09	116.29	210		214.88	252.25	157.29	
69	152.06	116.32	210	153.25	214.88	252.16	157.18	149.2
70	150	116.28	210	153.25		252.25	157	149
71		116.35	210	153.23	214.76	252.1		149.2
72		116.31		153.36	214.86		148	156
73	152.15		210		214	252	156.21	149.32
74		116		153.33		252		149
75		116.28		153.33		252.22		149.09
76		116.3	210	153.26		252.19	157.18	149.2
77		116.29	210	153.28		252.12	157.21	149.31
78	151.98	116.33		153.36	214.83	252.18		149.31
79	152		210	153	214.88		157.18	149.2
80	152		210			252.16		
81		116.35		153.22		252.18		

	D1mit425		**D1mit94**		**D1mit155**		**D1mit215**	
Mouse	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6	C3H/HeJ	C57BL/6
82		116		153.33	214.9	252.23		
83	152.08	116.35		152.31	214.81	252.17	156	149
84	152.08		209.96	153.25	193.71	207.64	157.21	149.21
85	151		210		214	252	156.27	
86	152.21	116.3		153.33	214.75		157.34	149.21
87	152.02		210		214.82		157.32	149.2
88	152.1	116.29		153.25	214	252.27	157.21	
89	152.08	116.33		153.28		252.2		147.95
90		116		153.33		252	157.21	149.2
91	152.1		210.02	153.36	214.79	252.18	157.14	
92	152.06		210.09	153.32	214.93	252.23	157.21	
93	150	116.38	210	153.25		252.17	157.21	149.21
94	153.16		210	153.22	214.87	253.23	156.17	147.93
95	152.17		210.01		216.89		157.25	
96		116.34		153.32		207.75	157.94	149.2

Table A1.5.2: *Tir3* genotypes and associated Survival of C3H/HeJ x C57BL/6 infected with *T. congolense* strain IL1180.

Marker	D1mit215	D1mit155	D1mit94	D1mit425
Total Number of Mice Genoytyped	83	96	96	96
Observed Genotypes at Locus				
Total C3H/HeJ	19	13	20	23
Total C57BL/6	17	44	38	30
Total heterozygote	47	39	38	43
Expected Genotypes at Locus (if in Hardy Weinberg Equilibrium)				
Hardy Weinberg P value	0.21	8.31E-06	0.0043	0.36
Mean Survival (days)				
C3H/HeJ Genotype	99.6	72.4	93.9	86.5
C57BL/6 Genotype	118.4	112.3	117.7	121.7
Heterozygotes	103.6	108.7	99.9	104.7

Appendix Table A1.6.1: Genotyping data for *Tlr4*

Genotyping data for C3H/HeJ x C57BL/6 F2 crossed mice with the most extreme survival times after infection with *T. congolense* strain IL1180. Microsatellite (D4mit[xxx]) loci are listed as the allele sizes detected by PeakScanner (Applied Biosystems) after PCR and electrophoresis on an ABI-3130xl capillary sequencer. SNP (rs) loci are summarised as:
1 = C3H/HeJ homozygote ; 2 = C57BL6 homozygote ; 3 = heterozygote ; 0 = null allele (no data)

Mouse	D4mit178		rs3023006
	C3H/HeJ	C57BL/6	Genotype
1		148.99	2
2		148	2
3	169.59	148.98	3
4	167.61	148.88	3
5	167.58	148.87	3
6	169.52	148.88	3
7		148.87	2
8		148.87	3
9		148.86	2
10		148.87	2
11		148.86	2
12	169.55		1
13	169.69	151.37	3
14	169.65	151.27	3
15	169.53	148.87	3
16	169.61	148.87	3
17	167.49	148.86	3
18	169.54	148.86	3
19	169.57		1
20	169.56	151.19	3
21		148	1
22		148.87	2
23	167.5		1

Mouse	D4mit178		rs3023006
	C3H/HeJ	C57BL/6	Genotype
24	169.5		1
25		148.99	2
26	169.61		1
27	169.58	148.87	3
28	169.65		1
29	169.59	148.86	3
30	169.53	148.87	3
31	169.48	148.86	3
32	167.46		1
33	169.62	148.75	3
34	167.59	151.19	3
35	169.43	148.86	3
36	167.5	148.86	3
37		149	2
38	169.66	148.99	3
39	165.46	148.88	3
40	169.53	148.88	3
41		148.87	2
42	169.55	151.18	3
43	169.58		1
44	167.48		1
45		148.86	2
46	169.57	148.86	3

Mouse	D4mit178		rs3023006
	C3H/HeJ	C57BL/6	Genotype
47	167.49	148.86	3
48	169.49		1
49	169.72	148.88	3
50	169.72		1
51	169.54	148.87	3
52	169.54	151.17	3
53		148.86	2
54		148.87	2
55	167.46	148.87	3
56		148	2
57	169.58	148.86	3
58	169.5		1
59	167.5	151.19	3
60	167.5	148.85	3
61	169.62	148.88	3
62	167.59	148.99	3
63		148.87	2
64	167.45		1
65	169.5	148.86	3
66		148.86	2
67	167.42	148.86	3
68		148.87	2
69		148.85	2
70	169.49		1
71	169.43	148.86	3
72		148.87	2
73	169.69	148.99	3
74	167.56		1
75	169.46	148.88	3
76	167.44		1
77	169.51	148.86	3
78	169.58	148.76	3
79	169.58	157	1
80	167.37		1
81	169.49	148.75	3
82		148.87	3
83	167.43	148.85	3
84	167.39	148.87	3
85	167.62	148.99	3
86		148.99	2
87	169.58	148.8	3
88	169.47	148.88	3
89		148.89	2
90	167.44	148.76	3
91	167.43	148.81	3
92	169.56	148.88	3
93	167.37	151.23	3
94	169.46	148.87	3
95	169.47		1
96		148	2

Appendix Table A1.6.2:
Tlr4 Genotypes and associated Survival of C3H/HeJ x C57BL/6 infected with *T. congolense* strain IL1180.

Marker	rs3023006	D4mit178
Total Number of Mice Genoytyped	96	96
Observed Genotypes at Locus		
Total C3H/HeJ	20	18
Total C57BL/6	22	25
Total heterozygote	54	53
Expected Genotypes at Locus (if in Hardy Weinberg Equilibrium)		
Hardy Weinberg P value	0.45	0.36
Mean Survival (days)		
C3H/HeJ Genotype	104.47	105.65
C57BL/6 Genotype	111.14	110.83
Heterozygotes	104.04	103.42

Appendix II: Sequence Capture and Sequencing of *Tir1*

Table A2.1: Primers used for verification of 454-identified and publicly available SNP
Primers used for PCR and subsequent direct capillary-based dideoxynucleotide chain-terminator sequencing of SNP and an indel. A 24bp insertion in the 3'-UTR of *Mdc1* in susceptible (A/J; BALB/c; C3H/HeJ and 129P3) strains of mice relative to resistant C57BL/6 was verified using the *Mdc1* primers. ENSMUSSNP[xxx] and rs[xxx] were nsSNP in public datasets that had incomplete data across all susceptible strains and were predicted to be damaging by Polyphen [143].

Gene / SNP rsID	Chr	Position (bp)	Primer Sequences (5'-3')
Mdc1	17	35996005 -	Primer 1: AGCTCATCGTGGATTTTTGG
3'-UTR 24bp deletion		35996199	Primer 2: GGGGCTGGGAATAAGGTTTA
rs13471968	5	77409316	Primer 1: GATAAACGCTTGGCTTGTCC
			Primer 2:CCTCTCCTCATCATAGTCGTCC
rs46908277	5	77423911	Primer 1: CTGGGAAATGCTGTTGTGG
			Primer 2:AACTCAGAAATCCACCTGCC
rs13469260	1	151708206	Primer 1: ACATAGAGGTGGACTGGATGG
			Primer 2:AATACTCCCTCGTCTGCCC
rs46572905	1	172810709	Primer 1: GAACCACAAGTCTCTGCCTCC
			Primer 2:TGAATGAGGGGTTTGAATGG
rs32565724	1	172837666	Primer 1: TCTGTAAGCTCATGGAGGGC
			Primer 2:AGACCCCAGTGTCCAGATCC
rs31938776	1	172896598	Primer 1: CCTTTGCTTATCTCCTCTTCTCC
			Primer 2:TAAGGTGCCATAGCTGGAGG
rs13468876	1	173155890	Primer 1: GGAAGGGGCTAAAAGCTGG
			Primer 2:AGCATCCCTAACTCCAACCC
rs45643169	1	173290449	Primer 1: GCAGTACCAATAAGTTCGGG
			Primer 2:ACTGTGATCTTTGACCCAGC
rs47563734	1	173445364	Primer 1: AGCTCTTGGATGGTGACAGG
			Primer 2:TGGGTAAGTTTGGGATTCAGG
ENSMUSSNP3206521	1	173445848	Primer 1: ATTGCTTGCTTTCCAGAACC

Gene / SNP rsID	Chr	Position (bp)	Primer Sequences (5'-3')
			Primer 2:TCTCCACACCCAGAAGAAGG
rs31569041	1	173461221	Primer 1: TCAAACACCTGCCTCTCTCC
			Primer 2:GTTGCTTGCTTTCCAGAACC
rs31537441	1	173504085	Primer 1: AATGGAAGAAGACAGAACAGGG
			Primer 2:AATCAGAGGGACAGAGTGGG
rs49701531	1	173510917	Primer 1: CAATTCCAGTCCCTGCTCC
			Primer 2:TAAAATGCCCAGAACCCTCC
rs50073880	1	174518055	Primer 1: TCTGGAACCTTGTCCTCCC
			Primer 2:TACCACTCTCGTTTCCTGGG
rs50777261	1	175200471	Primer 1: TTTTCCAAGGTTTCTCCAGC
			Primer 2:AACATCTGGGTAGCACAGCC
rs51259593	1	175262420	Primer 1: AGGCCACCAGAAAATGAACC
			Primer 2:TCAGTGTGGGACTTTGGGG
ENSMUSSNP4607197	1	175456777	Primer 1: GATCTTGTTGGAGTTTCGGG
			Primer 2:GCAATAACAGCCTCAGTCCC
rs45743507	1	175525403	Primer 1: CAATCCAGCATCCTCTGGG
			Primer 2:CTATGTCAAGGGCTGTTCGC
ENSMUSSNP2076364	1	175613358	Primer 1: CAAGCAATAAAGGCAAGTGG
			Primer 2:TAAGGGCTGGATACTGTTGC
ENSMUSSNP3208701	1	175625728	Primer 1: TTGAATGGAAACTTGAGGGG
			Primer 2:CTTGCTGTTGTCTGTGGTGG

Table A2.2: List of annotated non-synonymous SNP at *Tir1* from 454 resequencing and publicly available data

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
rs33493422	33273721	L/P	A/G	A	G			benign	Morc2b	MORC family CW-type zinc finger protein 2B (TCE6)
ss159819044	33274181	I/V	T/C	T	C	C	C	benign	Morc2b	MORC family CW-type zinc finger protein 2B (TCE6)
ss159825982	33283941	Y/C	A/G	A		G	G	possibly damaging	EG383232	ZFP421
ENSMUSSNP3469236	33313652	A/S	G/T	G				benign	Olfr239	olfactory receptor 55
ENSMUSSNP7093716	33313676	T/S	A/T	A				benign	Olfr239	olfactory receptor 55
rs46324535	33379368	R/Q	G/A		G	C	G	benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
rs48052544	33379588	N/D	A/G		G	A	G	benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
ENSMUSSNP6788142	33379630	S/P	A/G	A				benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
ENSMUSSNP3495720	33379638	V/A	A/G	A				benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
rs48958389	33379664	I/M	C/G		G	C	G	benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
ENSMUSSNP764983	33379785	K/T	T/G	T	G			benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
ENSMUSSNP6564304	33380661	S/L	G/A	G				benign	Zfp422-rs1	zinc finger protein 422, related sequence 1
ss159826028	33472308	N/S	T/C	T		C	C	benign	Zfp81	zinc finger protein 81
ss159826029	33472495	G/R	C/T	C		T	T	benign	Zfp81	zinc finger protein 81
Sudbery et al 2009	33523403	R/Q	C/T	C	T			benign	MGI107547	zinc finger protein 101 Gene

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
ss159819414	33777716	R/K	G/A	G	A	A	A	benign	Pram1	PML-RARA-regulated adapter molecule 1 (PRAM-1)
ss159819417	33781645	L/P	T/C	T	C	C	C	probably damaging	Pram1	PML-RARA-regulated adapter molecule 1 (PRAM-1)
ss159826476	33956791	S/P	T/C	T		C		possibly damaging	Kank3	KN motif and ankyrin repeat domain-containing protein 3 (Ankyrin repeat domain-containing protein 47)
ENSMUSSNP3734896	34020615	E/A	T/G	T				benign	CT030732.13	Kinesin-like protein KIFC1
rs13459236	34021818	Q/E	C/G		C	G	G	benign	CT030732.13	Kinesin-like protein KIFC1
ENSMUSSNP671840	34022803	I/V	T/C	T	C			benign	CT030732.13	Kinesin-like protein KIFC1
ss159819549	34048493	N/S	A/G	A	G		G	benign	Daxx	Death domain-associated protein 6 (Daxx)
rs50014799	34069285	H/Y	C/T	C	-		-	probably damaging	Rgl2	Ral guanine nucleotide dissociation stimulator-like 2 (RalGDS-like factor)(Ras-associated protein RAB2L)
rs33482449	34077878	T/A	A/G	A	G	G	G	benign	Wdr46	WD repeat-containing protein 46 (WD repeat-containing protein BING4)

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
ss159819632	34080331	L/F	C/T	C	T	T		benign	Wdr46	WD repeat-containing protein 46 (WD repeat-containing protein BING4)
ss159819661	34085850	Q/K	C/A	C	A	A	A	benign	Wdr46	WD repeat-containing protein 46 (WD repeat-containing protein BING4)
rs46297313	34086281	M/K	T/A	T	A			benign	Wdr46	WD repeat-containing protein 46 (WD repeat-containing protein BING4)
Sudbery et al 2009	34112408	V/A	A/G	A		G		benign		
rs16787210	34112420	H/R	T/C	T	C	C	C	probably damaging	CR974462.5-1	tenascin XB
rs51128061	34112723	H/Q	G/T	G						
rs51600946	34112725	H/D	G/C	G				benign	CR974462.5-1	tenascin XB
rs46902096	34112919	D/G	T/C	T	C			benign	CR974462.5-1	tenascin XB
rs49488597	34114733	A/E	G/T	G				benign	CR974462.5-1	tenascin XB
rs49259513	34114735	N/K	G/A/C	G	-			benign	CR974462.5-1	tenascin XB
Sudbery et al 2009	34114746	E/Q	C/G	C		G				
rs50972829	34114814	V/A	A/G	A	-			benign	CR974462.5-1	tenascin XB
rs48267223	34114833	G/R	C/T	C	-	-		possibly damaging	CR974462.5-1	tenascin XB
rs50588965	34114957	E/D	C/A	C	A			benign	CR974462.5-1	tenascin XB
Sudbery et al 2009	34115254	Q/H	C/A	C		A				
ENSMUSSNP2116722	34115436	M/V	T/C	T	C			benign	CR974462.5-1	tenascin XB
rs8237574	34115704	E/A	T/G	T	G	G	G	benign	CR974462.5-1	tenascin XB
rs47311850	34119278	R/H	G/A	G	A			possibly damaging	AA388235	hypothetical protein LOC433100

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
rs48590366	34119293	Q/R	A/G	A				benign	AA388235	hypothetical protein LOC433100
rs46706640	34119320	R/M	G/T	G	T			benign	AA388235	hypothetical protein LOC433100
rs50038295	34119383	G/D	G/A	G	A			possibly damaging	AA388235	hypothetical protein LOC433100
rs51272593	34119394	I/V	A/G	A	G			benign	AA388235	hypothetical protein LOC433100
rs46685594	34119426	R/S	G/C	G	C			benign	AA388235	hypothetical protein LOC433100
rs49851601	34119473	F/S	T/C	T	C			possibly damaging	AA388235	hypothetical protein LOC433100
rs45732289	34119478	A/P	G/C	G				benign	AA388235	hypothetical protein LOC433100
rs46016319	34119481	V/M	G/A	G	A			benign	AA388235	hypothetical protein LOC433100
rs50711935	34119574	R/W	C/T	C	T			benign	AA388235	hypothetical protein LOC433100
rs33344611	34134067	V/I	C/T	C	T	-	T	benign	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs33280712	34134075	M/T	A/G	A	G			possibly damaging	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs52050947	34134463	D/G	T/C	T				benign	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
rs33293753	34134481	H/R	T/C	T	C			possibly damaging	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs8237942	34136452	G/E	C/T		C	T	C	possibly damaging	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs8237968	34136684	K/T	T/G	T						
rs8237969	34136685	K/E	T/C	T				benign	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
Sudbery et al 2009	34136703	G/R	C/G	C			G	benign	MGI95904	Histocompatibility 2, K1, K region Gene
Sudbery et al 2009	34136753	K/R	T/C	T		C		benign	MGI95904	Histocompatibility 2, K1, K region Gene
rs51419043	34136895	E/K	C/T	C				benign	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs48191051	34136924	V/D	A/T	A						
rs49568909	34137198	V/A	A/G	A	G			benign	H2-K1	H-2 class I histocompatibility antigen, K-B alpha chain Precursor (H-2K(B))
rs49854621	34158807	T/A	T/C	T	C			benign	Ring1	E3 ubiquitin-protein ligase RING1 (EC 6.3.2.-)(RING finger protein 1)(Polycomb complex protein RING1)(Transcription repressor Ring1A)

Variation ID	Position (bp)	Peptide shift	Allele	C57BL/6	A/J	BALB/c	C3H/HeJ	Polyphen	Gene Name	Description
ss159819951	34165995	V/M	C/T	C	T		T	benign	Slc39a7	Zinc transporter SLC39A7 (Solute carrier family 39 member 7)(Histidine-rich membrane protein Ke4)
ss159819959	34167349	A/T	C/T	C	T	T	T	benign	Slc39a7	Zinc transporter SLC39A7 (Solute carrier family 39 member 7)(Histidine-rich membrane protein Ke4)
rs29537070	34169377	V/L	G/T	G	T			benign	Rxrb	Retinoic acid receptor RXR-beta (Retinoid X receptor beta)(Nuclear receptor subfamily 2 group B member 2)(MHC class I regulatory element-binding protein H-2RIIBP)
rs51544304	34188746	M/V	A/G	A	G	G	G	benign	Col11a2	Collagen alpha-2(XI) chain Precursor
ENSMUSSNP6670506	34195857	A/T	G/A	G	A			benign	Col11a2	Collagen alpha-2(XI) chain Precursor
ss159826904	34197872	N/D	A/G	A		G		benign	Col11a2	Collagen alpha-2(XI) chain Precursor

Appendix III: Additional Analyses for candidate gene number reduction and SNP annotation

To fully characterise candidate genes at the QTL and to enumerate all functional SNP contained therein, a number of additional analyses were conducted by colleagues within the team [149]: The physical boundaries of the QTL had to be identified from the mapping distances (i.e. converting centimorgans to base pairs); the novel 454 SNP data had to be combined with public datasets such as from Perlegen [131] and the Mouse Genomes Project [135] and subsequently annotated for potential functional damage; and ancestral haplotypes were derived from the combined sets of SNP to assign genes to having a shared ancestral haplotype with the resistant strain and thus remove them from lists of potential candidates.

Identification of boundaries

The two independent F6 and F12 mapping populations have reduced the 95% CI of the QTL to exceptionally small regions, particularly at *Tir1*, the QTL of largest effect, where the physical size of the 95% CI was 930 Kbp for the combined data from the A/J × C57BL/6 and BALB/c × C57BL/6 F6 crosses (Main text: Table 2.7). This was only twice the mean distance between markers at this locus (400 Kbp) and consequently the main limitation in identifying the boundaries of the QTL is in estimating the position of the peak. Initial numbers of candidate genes were estimated using the physical position of the peak marker in the F6 and F12 AIL studies [128, 129] as the most likely position of the peak of the QTL. The physical size of the 95% confidence interval (CI) was estimated by using Mouse Genome Informatics data to find the median Kbp/cM ratio for the intervals between the ten flanking markers (which were spaced at ~0.3Mbp intervals). This ratio varied between 0.69 - 5.43 Mpb/cM and was used to convert the 95% CI in cM to Kbp.

Analysis of Public Datasets

Lists of published non-synonymous SNP (nsSNP), SNP in splice sites; and regulatory regions and SNP that cause gain or loss of stop codons were obtained from BioMart. nsSNP were annotated using Polyphen [143] in order to identify those most likely to modify gene function. Polyphen classifies nsSNP as benign, possibly damaging or probably damaging according to the likelihood that the polymorphism will modify protein activity. 'Damaging' implies a change of activity or function but this change could be beneficial to the animal. Additionally, phastCons conservation scores for SNP positions [270] were obtained from UCSC [2], which measure how conserved a position is amongst 30 mammalian species and are on a scale between 0-1 with the most conserved positions scored as 1. In this manner, nsSNP that occur at sites that are relatively conserved across species, having a phastCons score close to 1, are likely to have a greater effect on the function of a gene than at sites that are known to be polymorphic.

454 SNP validation

In order to validate SNP calls, 454-generated SNP were compared against those released in a recently published set sequenced on the Solexa/Illumina platform from flow-sorted mouse chromosome 17 for A/J [135], and similar, publicly available SNP from the concurrent Mouse Genomes Project (Wellcome Trust Sanger Institute) for BALB/c, C3H/HeJ and 129P2 mouse breeds [92]. Only 3 out of 36,784 (0.014%) of the homozygous calls (coverage > 1; alternative allele frequency (AAF) > 80%) were discordant between the two datasets. The 454 data included 53 - 71% of SNP in the Illumina data depending on the coverage required to call a SNP and the Illumina data contained 94 - 97% of SNP in the 454 data.

Haplotype Block Analysis

Whilst there are large numbers of reported SNP for A/J, 129X1/SvJ and 129S1/SvImJ due to the Celera sequencing project [271] and for BALBc/ByJ and C3H/HeJ from the Perlegen project [131], relatively few SNP are publicly available for the 129P3 strain. The 454 resequencing of the *Tir1* region indicated that approximately 50% of the resequenced region could be excluded from the QTL if the allele carried by 129P3

mice at this locus was known. If a QTL was identified at *Tir1* in a 129P3 × C57BL/6 cross then the QTL gene could be assumed to be within the three blocks where 129P3 differed from C57BL/6. If no evidence of a QTL was found then these regions could be excluded from the QTL on the assumption that 129P3 carried the same allele as C57BL/6 at this locus. This analysis indicates that mapping QTL for response to infection in a 129P3 x C57BL/6 cross should significantly refine the list of candidate genes. The availability of this haplotype data makes it possible to make more rational choices about the selection of strains for mapping experiments. This strategy has been used before with a much more limited SNP set [272] but the corresponding online resource is no longer available.

The correlation of Jukes-Cantor distances calculated from our 454 data and the published Perlegen dataset was only modest ($r = 0.63$). 32% of our 454 SNP loci were also in the Perlegen set, however the low correlation between the two sets shows that SNP discovery was uneven in one or both sets and inspection of the SNP distribution suggests that this was certainly the case in the Perlegen set. The uneven distribution of SNP discovery makes it much harder to undertake a consistent analysis across the genome using a single threshold for assigning alleles to haplotype blocks. However the high positive predictive value for identifying shared haplotypes suggests that this procedure should reliably exclude regions where C57BL/6 shares haplotypes with the susceptible strains. Nevertheless other more robust data types such as CNV and potentially functional SNP should still be surveyed in regions where haplotype does not correlate with phenotype. The more complete mouse resequencing projects that are currently underway should increase the predictive power of this approach substantially.

Appendix IV: Supplementary Multilocus Microsatellite Genotyping Data

Table A4.1: Primers used for multilocus microsatellite PCR of 31 Ugandan *T. b. rhodesiense* isolates. Forward and reverse primers are listed along with their chromosome and locus. A fluorescent dye is listed when used for capillary-based sequencer sizing.

Chromosome	Dye	Locus	FORWARD PRIMER (5' – 3')	REVERSE PRIMER (5' – 3')
1	FAM	1/18 inner	TATAATGCGTTTGTGAGAAT	TGTGAGAATGGTACTCACGCGCTG
		1/18 outer	GAAGGGAGGGAACAGAAGCAGGG	CAACGTTAGCACACAATTCCTGTG
2	FAM	2/5 inner	TATCGCGGTTATGTGGATTTGTGG	ATGGCGTGTATCACATTCGTGATG
		2/5 outer	CACAACAAAACTGCCATGAGGTAC	CCGTTGGCATTAGGCACAAGTA
2	FAM	PLC inner	TTAAGTGGACGACGAAATAACAACA	CAACGACGTTGGAAGAGTGTGAAC
		PLC outer	TTCAAACACCGTCCCCCTCAATAAT	CCACTGACCTTTCATTTGATCGCTTTC
3	FAM	5L5/2 inner	GAGCGTACATTGCAGGTAGTGCGTAGCG	GTACGTGGTTAACCACAACCTACT
		5L5/2 outer	ACGAAGAAACGAAGCAAAGAAG	GGAAACTGCTTAAACTTGCGTGAG
4	FAM	M12C12 inner	AAAACCTCATCCAGTCGCACTGG	TGGACACACAGAAGCCTACCG
		M12C12 outer	TACCCTCATCAAGTGGTCG	AGTGTGGTGGTGCGTGCAAACTTGG
5	FAM	JS2 inner	AGTAATGGGAATGAGCGTCACCAG	GATTGGCGCAACAACTTTCACATACG
		JS2 outer	GATCTTCGCTTACACAAGCGGTAC	CTTTCTTCCTTGGCCATTGTTTTACTAT
8	FAM	Tb8-393863	CCAGCAGAATGATGCAAAGA	TGGTAGCTTGGCGTCTTACC
8	FAM	Tb8-1074322	AGACGAAGCAGCGAGAAGAC	TCTCATTACTGCTCTGTTTTTGC
1	FAM	Tr401-1	GTGAAAAACGAAAGGCAACG	TGAGTTCAACAATCTTTTATTCC
2	FAM	TB2/21	CTGTGTGTTGCTTGTTCATA	AGTTTAACAGCACTTCCATTT
7	FAM	Tr407-1	AACAATATCTGACAATGAGGATGG	GTTTAGATGGGTGGAAAGGGTAGG
11	FAM	TB11/13	CAAGAACTCTGCATTGAGC	ATCTGTTGGCGATGGTGA

Table A4.2: Microsatellite genotyping data for eleven informative genome-wide loci in BAPS data format. Genotypes were manually assigned to a 'bin' of of alleles with similar sized microsatellites. As the genome is diploid, each sample is displayed across two rows, representing each allele at a locus. Sample IDs are identical to those shown in the main text (Table 4.2).

Sample ID	5L5	PLC	JS/2	18	5	11/13	401/1	407/1	Ch8_001	Ch8_002	2/21	BAPS cluster (K=3)	Zymodeme
1	1	1	1	1	1	1	1	1	3	3	2	2	Z375
1	1	3	2	1	3	1	2	1	3	3	3	2	
2	1	2	1	1	3	1	1	1	2	3	2	2	B17
2	1	3	1	1	3	1	2	1	3	3	3	2	
3	1	2	1	1	1	1	1	1	3	3	2	2	Z375
3	1	3	2	1	3	1	2	1	3	3	3	2	
4	1	1	1	1	3	1	1	1	1	3	1	1	Z366
4	1	2	1	4	3	2	2	2	3	4	3	1	
5	1	2	1	1	1	1	1	1	3	3	3	2	Z310
5	1	2	2	1	3	1	2	1	3	3	3	2	
6	1	2	2	1	1	1	1	1	2	3	2	2	Z310
6	1	3	3	1	1	1	2	1	3	3	3	2	
7	1	1	2	1	1	1	1	1	2	2	2	2	B17
7	1	1	3	1	3	1	2	2	3	3	3	2	
8	1	1	3	1	2	1	1	1	1	3	1	1	Z366
8	1	1	4	1	3	2	2	1	3	4	3	1	
9	1	1	2	1	1	1	1	1	3	3	2	2	Z309
9	1	1	2	1	3	1	2	1	3	3	3	2	
10	1	1	2	1	1	1	1	1	3	2	3	2	B359
10	1	1	2	3	3	3	2	1	3	3	3	2	
11	1	2	1	1	1	1	1	1	3	3	3	2	Z310
11	1	5	2	1	3	1	2	1	3	3	3	2	
12	1	2	1	1	1	1	1	1	2	1	2	2	B17
12	1	3	1	1	3	1	2	1	3	3	3	2	
13	1	2	1	1	1	1	1	1	3	3	2	2	Z310
13	1	3	2	1	3	1	2	1	3	3	3	2	
14	1	1	1	1	3	1	1	1	1	3	1	1	Z366
14	1	3	2	1	3	2	2	2	3	4	3	1	
15	1	1	1	1	3	1	1	1	1	3	1	1	Z366
15	1	2	2	1	3	2	2	2	3	4	3	1	
16	1	1	1	1	3	1	1	1	1	3	1	1	Z366
16	1	2	2	4	3	2	2	2	3	4	3	1	
17	1	2	1	1	1	1	1	1	2	3	2	2	B17
17	1	3	1	1	3	1	2	1	3	3	3	2	
18	1	2	1	1	1	1	1	1	3	3	2	2	unknown
18	1	3	2	1	3	1	2	1	3	3	3	2	

Sample ID	5L5	PLC	JS/2	18	5	11/13	401/1	407/1	Ch8_001	Ch8_002	2/21	BAPS cluster (K=3)	Zymodeme
19	1	2	1	1	2	1	1	1	1	3	1	1	Z366
19	1	3	2	4	3	2	2	2	3	4	3	1	
20	1	2	1	1	1	1	1	1	3	2	2	2	Z311
20	1	3	2	1	1	1	2	1	3	3	3	2	
21	1	2	1	1	1	1	1	1	2	2	2	2	B17
21	1	3	1	1	1	1	2	1	3	3	3	2	
22	3	3	1	1	1	1	1	1	1	3	3	3	Z377
22	3	3	3	1	5	2	3	2	1	3	4	3	
23	1	3	1	1	1	1	1	1	3	3	2	2	Z310
23	1	3	2	1	3	1	2	1	3	3	3	2	
24	1	3	1	1	3	1	1	1	3	2	2	2	unknown
24	1	3	2	1	3	2	2	2	3	3	3	2	
25	1	3	1	1	2	1	1	1	1	3	1	1	Z366
25	1	3	2	1	3	2	2	2	3	4	3	1	
26	1	3	1	1	1	1	1	1	2	3	2	2	Z310
26	1	3	2	1	1	1	2	1	3	3	3	2	
27	1	2	1	1	1	1	1	1	2	3	3	2	B359
27	1	5	2	1	3	3	2	1	3	3	3	2	
28	1	3	1	1	1	1	1	1	3	2	2	2	B17
28	2	4	1	1	1	1	2	1	3	3	3	2	
29	1	2	1	1	1	1	1	1	3	3	2	2	Z375
29	1	3	2	1	3	1	2	1	3	3	3	2	
30	1	3	1	1	2	1	1	1	3	3	2	2	Z310
30	1	3	2	1	3	1	2	1	3	3	3	2	
31	1	3	1	1	3	1	1	1	1	3	1	1	B376
31	1	3	2	1	4	2	2	2	3	4	3	1	

Appendix V: Additional *T. b. rhodesiense* virulence phenotype data

Previous studies on cytokine-driven pathology

The causes of pathology have been attributed to the pro- and anti-inflammatory responses of the host. Early pro-inflammatory responses are required to control the initial peak of parasitemia [35, 273], however the subsequent expression of cytokines such as IFNG are suspected to be a major cause of pathology [274]. Furthermore, IFNG also acts as growth factor for the parasite [275]. In order to mediate inflammation-related pathology, a corresponding anti-inflammatory response involving cytokines such as TNFA and IL-6 is necessary. TNFA, whilst being toxic to the parasites, also contributes to immunosuppression [276] and correlates with severity of human disease [94].

Survival and cytokine response to parasite zymodemes in CD-1 mice

To determine whether different zymodemes elicit differences in levels of TNFa and IFNg, serum levels of these cytokines were previously monitored for seven weeks in *Busoga* and *Zambesi* zymodeme infections by Tanja Beament [210]. TNFa (Figure A5.1A) and IFNg (Figure A5.1B) concentrations in the serum of sacrificed CD-1 mice infected with Z310 *T. b. rhodesiense* parasites were significantly higher than those infected with B17 zymodeme parasites at week 3 (Mann-Whitney; $p < 0.01$), but were not statistically significant at other time points. Similarly, the relationship between survival time and bloodstream parasitemia was previously investigated for B17- and Z310-zymodeme infections in mice. Parasitemia levels and the overall condition and behaviour of the mouse according to six "wellness" characteristics were monitored in CD-1 mice. Figure A5.2 shows parasitemia throughout the course of infection in CD-1 mice, as described by Tanja Beament, LSTM [210]. Mice infected with B17-zymodeme isolates survived for longer periods despite a higher first peak of parasitaemia at 3-5 days post-infection, compared to Z310. B17-infected mice did not appear ill according to the wellness characteristics and were never sacrificed prior to the endpoint of the experiment (38 days post-infection). Z310-infected CD-1 mice showed a generally

higher, more variable level of parasitaemia (with a lower initial peak of parasitaemia) and showed stronger symptoms associated with infection; 72% of these mice were humanely sacrificed earlier due to deterioration in health. Z310-infected mice had a significantly higher parasitemia at the time of sacrifice than B17-infected mice (Mann Whitney; $p < 0.001$). Mice infected with Z310 parasites lived for a significantly shorter length of time than mice infected with B17 trypanosomes ($p < 0.01$).

Figure A5.2 shows that for experimental infections of *Zambesi 310* in CD-1 mice, an increase in parasitemia towards the end of infection appears to correlate with survival. This, however, cannot the only factor influencing virulence, as parasitemia was erratic and not significantly different throughout infection [210]. Furthermore, both A/J mice, and humans, appear to be more susceptible to the *Busoga* strain group of *T. b. rhodesiense.*

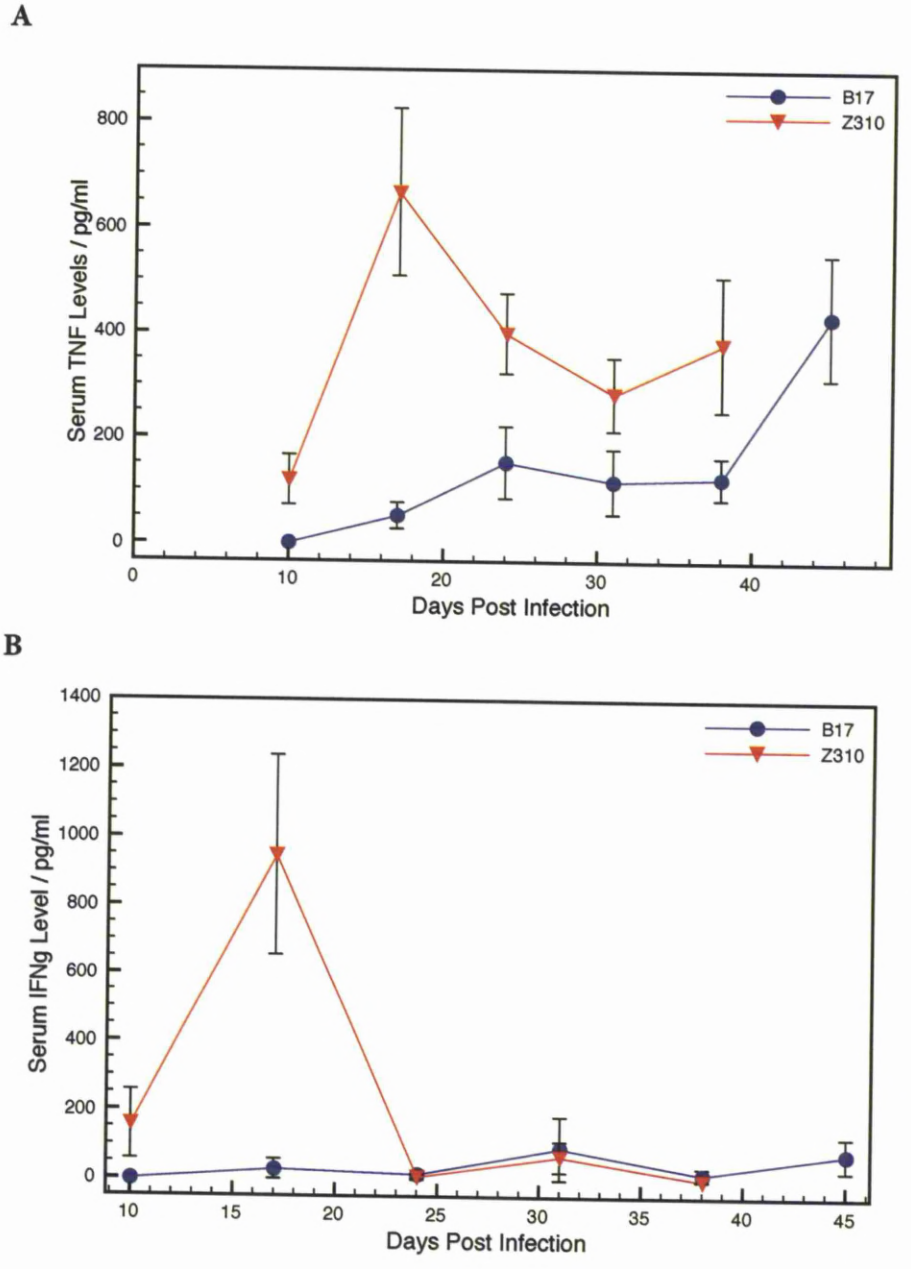


Figure A5.1: Mean serum levels (pg/ml ± standard error) of TNFa (A) and IFNg (B) throughout *T. b. rhodesiense* infection in CD-1 mice infected with three isolates representing each of two zymodemes. Six mice were tested in each case, except for Z310-infections after day 31 (five mice) and day 38 (three mice). No Z310 mice were available after day 45 as all had to be humanely sacrificed due to severity of illness. Data courtesy of Tanja Beament and Wendi Bailey; LSTM [210].

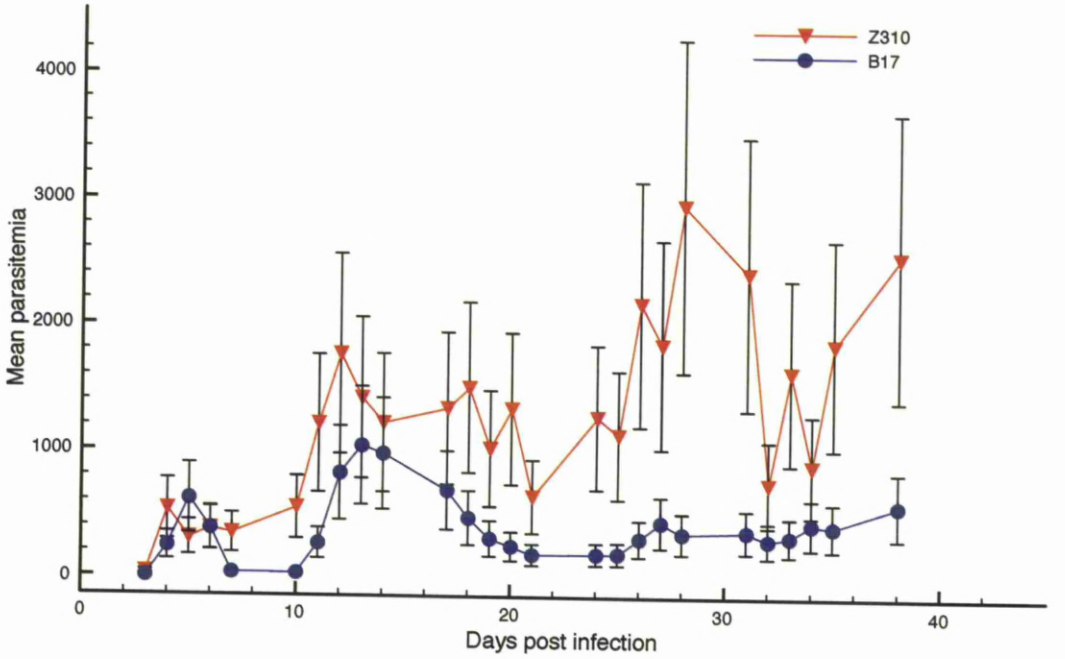


Figure A5.2: Parasitemia in experimental *T. b. rhodesiense* infections in CD-1 mice
Line represents mean parasitemia CD-1 mice infected with Z310 and B17 zymodeme *T. b. rhodesiense* parasites ± standard error (25 fields; thick film; x400 objective). A total of 30 mice per zymodeme were infected with three isolates representing each of two zymodemes. Six mice were sacrificed per week, not including mice that were sacrificed due to substantial symptoms. Number of mice available is shown (Bar chart, secondary Y-axis) [210].

Appendix VI: *T. b. rhodesiense* SNP Genotyping and Validation Primer Data

Table A6.1: CAPS-based SOLID SNP validation
CAPS loci, associated primers and confirmation status of 33 predicted SOLID SNP loci between *T. b. rhodesiense* Z310 and B17. Genomic coordinates are as per the *T. brucei brucei* TREU 927 v.4 reference sequence (GeneDB, [61]). Confirmation status represents whether the SNP locus was validated or not. In the case of unconfirmed SNP loci, whether the locus was within the coding sequence of a VSG element is highlighted.

Position (bp)	Forward Primer	Reverse Primer	Confirmed?
144608	TGCCCTGGAGGTAAATGAAG	TCCCGAATAAGCTCCTCCTT	Y
939154	AACCTTCCCCATCCGTATTC	CTTCCGCTCAGTGGAGAATC	Y
1274873	GGCAATGCGTAAAAAGGTTG	TCCGTCATGTTGCGTATGAG	Y
1277045	GATGATGGACAGACGAGTGC	CCTCCACATAGCGTCCATTC	Y
1277953	GCGCTTGAATGGGGAATA	TGGCAGATCAAAGAGCAAAA	Y
1501007	CAGCGCTGGTTCTTCTCG	GCCTCGAAGAAACGTGAGAG	Y
1807834	GGTCACAGAAGCGTGGTTTC	AAGGCTGAAGCCGTGTGTAG	Y
4093840	TCCCCAACTCTTCCGTGTAG	AAACCAGGTGAAATGGGTGA	Y
6051486	AGAAGGATACTCCGCCCTTT	TAGCGTCTTGCTCCTTCGTC	Y
8219383	CACCAATGCACACTGAACAA	AACGAGAAGTGCGAGTAACATAA	N
9749341	ACAGGCACCCTACTTTCAACT	CGCTTGGAAGCATGTCACTA	N
11067004	TCGCTGAAGGATTTCTTCGT	TGACAAAGAGCCAAAAGAGG	N
11395646	ACATGAACCGCACACAATTT	AACATTTGTCACATATATCATAACAGC	Y
11884938	AGCTGCTGCTTTTGTGAGTTT	CACAGGTGAAGTTCTCCCCTA	N (VSG)
12141648	GGGTGAAGGTGATGGAGACA	CCCTCTCCTTCTTCCACTGA	Y
12346430	GGGCTGCAGTGGTAATTGTC	GTCGGTACATCCGCTCATC	Y
13553833	TCACAATCGTCTGCACAACA	GCTGTCGAAGTTGGAGGAAA	N
13649484	CCACTACGGAGCCAAAAGG	GGTGTTCCCGAATTTTCTCA	Y
15069964	TGGTCAAGCAACCACTTTTT	TGCTACTACAGAGATAGTGGAAGTCA	N (VSG)
15525835	TTGTGTCTCGTACCGTGGTG	GTGCACACACGCACACATAG	Y
16352364	GCCAATCCTCATCCAATTTC	AACGCTGACACGCTACAACA	Y
17925806	AGAGGGAAACGAAGACATCC	CTGCAAATGAGCAGACCGTA	Y
19260317	GCGGTATAGCACCTCTCCAA	ATGTGGTACAACCCCTCCTG	Y
20477885	CACGATTCCCGAAGGTTTTA	CCGCTGTTAACGGAAGTCAT	N
20985877	TGAAGTTGAACATTGCGTCA	AGCGGGCTTTTGTGCTAAAC	N
21209754	TGTACCAGGGTGTGCAGAGA	TGCCGTGTGTGTCAAAATA	Y
21709517	GAGACCTGGGGAAACCAACT	CATAGGCTTCCGATGCAAGT	N
21858003	ACACACACGCTCACGACTTC	GCGAACCTTCCAGAAAACAT	Y
22409948	TGAAGGAGCCAAGGGAGTAA	GCGGGATTCTAAGAAAACTGG	Y
23086223	CCAAAAGACAACGCGGTATG	TTCACCGCTGCTATGACATC	N (VSG)
23293496	AGATGTGGTCATTTTCCATCAA	CGTAATCGCTGGTTCATGTG	N (VSG)
25476959	TCCAATACGGGGTCGATAAA	ACAGCGGCTGCAGAAGTAAT	N
25812336	TTTGTTGTTCTTTCACGCAGA	AGCATTTACGTTGCAGCTTG	N (VSG)

Table A6.2: KASPAR-based SNP genotyping loci

A list of SNP loci selected across all eleven megabase chromosomes in order to genotype all *T. b. rhodesiense* samples isolated by J. Wendi Bailey from the 1980s/1990s Ugandan outbreak. Loci are shown in the context of a 100bp window with a square bracket around the SNP tested, using either the IUPAC code for degenerate bases (e.g. [W]), or each of the individual bases separated by a forward slash (e.g. [A/T]). SNP were detected to be either homozygous (Hom) or heterozygous (Het) in the Z310 or B17 isolate resequenced.

SNP position (bp)	Sequence	Chromosome	Homozyous / Heterozygous
	GCTAGGATGAAATAATGAAGTTCGACTGTCAGCGTCATCATCATCATCA[W]CA		
10340074	TCAACATCATCAACATCATCGTCATTCTCATCGCGGCAAACGGCAGCA	1	Z310 Hom. B17 Het
	GCTGCACTTGCGGGACCTTTTCCGACTAAATTCTATGGTGCCATCGTGT[R]GGG		
11826710	TCTTGCCATATCCGGACTGATGACCTCATTTTTGGCTATTGTTATAC	2	Z310 Hom; B17 Het
	AGTTGCTTAGGTAAAGGTTTGTGAGATTCCGGCACTCGCACCCATTCAA[T/C]C		
11987611	AAACCTCCCTCAGGCTCACACAATCGTTAAGCTCCACTGTCATCAGATT	3	Z310 Hom.
	GCGCACTCTCTCCGCAAGGTCCTCACCCACGGGTTTGCACGATGCCGTA[A/T]G		
13727977	GTGAAGCACACGCAACACATATGTTTGTGCGACAGTTCTCATCCGTGAA	4	Z310 Hom
	GTGTTGTACGCCACCCGTCTGCAGAAGGAGCGCTTTGCCGCAGCGGCTC[C/T]G		
16296739	GACGAAGCGAAGGCGTTACAGGCATTTGAAGCGGCAAAGGCGGATATCG	5	Z310 Hom
	CAGGTAGTTGTCCATGTCATGCTGGAAGTGCTGAATCAGTGAGAACCGA[G/A]T		
16852700	GTCGCTGGCGAGCGGCACTGTTCCAGTTGTGATCAGCTGTCCAGGAGTA	6	Z310 Hom
	GAATTTATCCACCGAATGGCAACAGACAAGATTATCATTTGCTTTTACA[C/T]AC		
20080306	TTATAATGATAGGTATTATTATATTTGCTACGTTAAAGTTTTTGCAAA	7	Z310 Hom
	CTGCGTCGAGTGTGTGATCAGTTGCTTTCWCTCGATCCAGCTAAACGGC[Y]GTC		
22586191	ACTGCGACAGCTCTTCCAACAGCCTTTCATTCGCCGTGGCTTAGACT	8	Z310 Hom; B17 Het
	TTCCCGTTTTCGCACTCTCTCGTCTCTTTTTCCATAAGCGTGTCCAGCC[A/G]TG		
24286830	CACTTTCATTCGTAAGGTTCCATATCTTACCGTCCTCCTTAGTAACTA	9	Z310 Hom
	CAGCAACGCCATGGACGCAAACCCCTGCTCATTACACCTAACGATTAAA[Y]AG		
3988952	ACAGGAGGCAGCATCTGCAGCAGACCGACGCCAGCAGGAGCAACGAAA	10	Z310 Hom
	TTTCCTTTTTGTTGTTTTTTTTTTTTCCCTCCCCCTTCCCCGCCGTTTG[T/A]TGTT		
5360361	GTTTTTTCTTCACTCTTGATTTTATTCCTTTCCTCTCTCTCTCTCT	11	Z310 Hom
	AATGCGAAAGTTGTAACCGTGTTAAAGTGGATTACAAATTTCATTGAAA[C/A]TT		
9614469	TTTTTTGCTGAAGSAAAACATACTGCAACAACAATCKTTACAGCCAAT	1	B17 Hom
	CAAGCCGCCTGAATATGTGATCCCGCTTGTTGCCGCAATAACTACCTTG[A]CAC		
11391293	TTTGTTCTATTCTGGTAGTAAGTGACGGCAAGCTGGGGATCCTTACC	2	B17 Hom
12634050	ATTAGCCGCGAAAGCTTGAGTTGGAGCCTCGTTAATATTGAACTCAAAG[K]CAG	3	B17 Hom; Z310 Het

SNP position (bp)	Sequence	Chromosome	Homozyous / Heterozygous
	AGTTCTTATCAGCCTTTGAAATGCAGGTATGCTTCACTGTATGTGAT		
	TTCTTAGCATATTCTACAGAACTGCATAGTCGTCGTCGCTAACGTGGTG[R]AAA		
15020177	ACAATAAGATTRGGATTACTRCTTACGAATGTATGCAGGGGCCTTTG	4	B17 Hom. Z310 Het
	CTCCATTTCCATTTTCTTGATCTTCTTCTTTAGCTTCTTCATTTTCGCC[Y]CGATC		
15764227	TGCTCGGTCTCATCGCGGTGATTTTGTGTGGATTCCTGTAGGCGT	5	B17 Hom; Z310 Het
	ATGATCAGCCTGTTCATCGTTTGGGCTTTTATACAGTTTTCTATCTCCT[G]CCAT		
18027932	AAAGATATGTTCTTTTTCTCTTTGTGGGGTTATATTTTCATTTTCA	6	B17 Hom
	GCGGCGGCGGGCCGGCGTCTTCTGTGCGAAGGACAACGCCCGCCCCCCA[T/C]		
18625702	GGTGGTCCAGCGGCGTTTTCGCGGGTGATGATTCTCGTGTTAATGGGAGG	7	B17 Hom
	CGGACCTCGTCATTGGTTATATATGATGATGCTCTCAAAAGAACATTCG[G/T]CT		
20902551	CGGGTGCCTCGGCACCCGACAACGAGCCCGCGGATGGGGCGGATGTTT	8	B17 Hom
	GAGGCATATGACAGTCGCGGTCGGCCGAACGCTAAAAAAAAAGAAAAACT[T/G]		
23754764	ATCGCCAATAATCTTACATCAACAAGGCGTTTTTCTTTGTGTTTCTTCTG	9	B17 Hom
	TTTTCACGGAAGTTAAGCAATCTCGGGAGGCAGAGCTTTTGTCTCTATG[R]AAC		
673532	TACTGGGTACAGCTCCGACGTGAAACGCGGGCGCTCTCTGCTGAGGT	10	B17 Hom. Z310 Het
	TGCACTCTTTTTTTTTCTGTCTCAAGAAATTCGTTCGCTTCACCTAGTT[T/A]AGC		
8082221	ATAAATTTCACGACAAAATTTCATCGCCACCACCACCTCCTCCCTAT	11	B17 Hom
	CACCACCTACGAAAAGCACAGCGCCGCCCAACTGCGCAGCCGAATGCAC[Y]A		
9613492	GCCGMAAACCCCGCCCCGCAGCACAAGACCGGAACCGCGGAGGAAGCGA	1	B17 Het
	CCGGTCGCAGTCCGTGCCGTTTCTCAGCGTGGCCACGGGGCCGTTTCGT[M]CC		
11405143	TTTTTTCGCTCTCCCTCTTTGGCAAATGACGGCACAGAAGAGATGACT	2	B17 Het
	AAGTTTTTTTGCTGCACGTCTCCAGCACACTAGTTKTTTCCYGGGTTGA[R]GCTC		
12262672	ATTATTTCCCATTATATAACGATATTCCCAAGCCACAGAGGGAAAG	3	B17 Het. Z310 Hom
	GAAACGTTACAAAGGCGGGCAATTCGTTCAACACATTTTCGAAGCCGCC[R]GG		
13649484	GAAAATGCGAATACTTCTCCTTGCAGTATGATTTAGATGACCATTGAT	4	B17 Het. Z310 Hom
	GTGGTGGAGATTGTTGRTTCCTTAATGAACTATGTGTTTGTTTATCATA[W]TTTY		
15882477	GATGCGTTTCAGAGCTGCATGATGGGCGTGCTGCGTGGGTGTGGAA	5	B17 Het
	GTTCGAATCACTTACAAGCGCAYGGAAGTGCTCCGCCCGAACGTAAACC[K]GG		
20861527	TCCTTTTCCTTCAGCCTGCTCCGTTCAGACCGTGATTCATCTCCTTGC	8	B17 Het
	CTACGCTTTCCGTGAAGAACTTCGACTAAGGCTCCGGATAGAGCAAACT[Y]GTA		
21377395	GTAACACTATGTCGGCTGAGGAGCTGGAAAAACTCAACAAACAGTTT	8	B17 Het
	CATCTCATCACCATCCATGTCAACAACCGTCCCACTCACCTTAATGCGA[Y]CM		
21642832	AAGCTTGGACGTGTTGTAGACACAAAGCGCGGTGTTAATGCACCTGCG	8	B17 Het
21642834	TCTCATCACCATCCATGTCAACAACCGTCCCACTCACCTTAATGCGAYC[M]AA	8	B17 Het

SNP position (bp)	Sequence	Chromosome	Homozyous / Heterozygous
	GCTTGGACGTGTTGTAGACACAAAGCGCGGTGTTAATGCACCTGCGTT		
	TTGGAGCATATGTTGTCGCGCGACTGGGCAGCGCCCCTACGTGGAAGTC[Y]CTC		
22171609	CTTAGGAAAGCGACACGATGCAGTTCAGCAAATTCAGCAAATAACCG	8	B17 Het
	ACCCTTTACGAGATCCCGCGCACGCTCATCAAACCAATTGGGGAATTTA[R]GC		
24718735	CGGCCCGCAAGAATCTTTTCATACGTCCGAATTGGGGTTTCATCAAAA	9	B17 Het
	CTGACGAAAGACCTTGTATTGGGTGATGAATGTAATGGTGCTCATCATC[W]TCT		
3832579	AGACATGACCTTTCCAATACACAATGGCGTCATCCAGAATATGGACG	10	B17 Het
	GTTCTCCTTAAGGTGGACAACAAACCAGTTTTTGTTCAGTAAATGTACG[R]CGA		
6096006	GAAGTGGGACGAGGTGCTCCCATGGTTGGACGTGGAGCTACATAAGA	11	B17 Het
	CACAGCTTTTTTTGCCGCTGTGGCTGCTGCTGCTGCCGCTGATGATGAT[R]ATG		
11331716	ATAATGATGAGGAAGCCGGTAATTCCCCCAATGTCCCTTTATTGTCT	2	Z310 Het
	CGGGGCTCCAGTTTCCTYCATTTTGGTGGTTGTGAGCACATCWGTACTT[Y]CCT		
12546628	TTGCCTTAGACGCGCCTCTCATCCACGAAAACGCACTCGCTACTCGT	3	Z310 Het
	ACTCCTTTTCTTTTCCCTTGAGAGCGACGGGTGGAAACTCGGCGCACCG[S]TGG		
13745548	CGCCGTAACTTCTTTTCGTGAGCGCCGCCTGCGGTCCTCTACGGTGT	4	Z310 Het
	GTCTAGTGGAGGCGCCGCTGCTGCTGCTCCTGCCGCCGGTGGCGCACCT[S]CTG		
15668532	CCGCTGCTCCTGCAAAGGAGGAAGAGGAAGATGATGATGATATGGGC	5	Z310 Het
	ACACGATCGCACCTCACCAGTGACTGAAGCACAAACAAATCGTCGCCGC[S]AA		
16096240	CGGAACGTGGGAGCACCACGTTCACTTCTCCATTTGGATATTGACCGC	5	Z310 Het
	GTTATGCCGCGACAAGCTGCAAATACCTTTCTACCAGTAATTGCTCTAC[M]AG		
16374695	GCGCTASGACAGATGCCCCRTGAGACATTACGGGCAAGACATTACCAA	5	Z310 Het. B17 Hom
	TATTTTCGCGTTCCACGCCCCTGGTATCCTTGTCGTCCGCGGTCCTTTC[M]CCC		
16946362	GCGCCCTCGGCCGTTGTGGCCCCTTCTTTATTATATGGATGAAGTTT	6	Z310 Het
	ATACTGCAAAGGCACCTGTGCAAGSGGCCACCACGTGATAGCGAACTTA[W]TA		
19707953	ATCGGTCGGTRACTGAGGTTGAGGGTACTCAGCCCCGTCGTCTTTCAG	7	Z310 Het
	TTGGCCCAACTTCAAAGCAGCTCTAACATTACCGCCGCTACGACCAACG[R]AA		
22957610	ACGGTGCGACTAATGAGAGCGGATGCGGTGGCTCCACAAGTAGTTCTA	8	Z310 Het
	TGGCTACTTTTGACAGTGTTATGCACTGCAGGTATTCGTGCTGACCCAA[W]TAA		
25400302	GGAGGAGTGTCCGGATTTGGGAGGTGATCCATCAGCTGACAACGTAC	9	Z310 Het
	AAGCTTAATGAGGAGTCGTTGGAGCGTTTCTTTGCTATTGTGAACGAGT[Y]CCTT		
186991	TCGGTTAACCCGWTACAGGTTGAAGCGGTTAGCTTTRCAGTTCATA	10	Z310 Het
	CCAACTGCCACATCAGCGAATGGTACAGCTCTGCGAGGAGGTCCAACGT[M]CT		
661569	GAGAATTTTACAGCAGSCGAYCACAACTCCACGGGGTGTCGCGTTGAC	10	Z310 Het. B17 Hom
1174392	GTTCCACAAACGAACTCAACACCAGGTTTTTCAAGTGCCCTCAGCAAAC[R]GG	10	Z310 Het. B17 Hom

SNP position (bp)	Sequence	Chromosome	Homozyous / Heterozygous
	GTACTGTTTTTGGTGGGTGCTGAAGGTCCGCATCCATTACGAGAATGA		
1769896	TGGAAACAATGGCACCACCGTGACGGCCGCTCTTATGGCTCATAAACAC[R]GA GTTTCGTGGCGCACCAAAACTGGAACGAAACAACCGGATTATCTCGGC	10	Z310 Het. B17 Hom
6567536	AAGTRTGTCTTTAATTTTTTTTTTGTWTCATCCCTCATCCGTTCACCCGC[M]CAA AGTGTTCCAAGGTGTCGATCATCATGCAGAKGTTGGCCCCTGGTGCT	11	Z310 Het

Appendix VII: *T. b. rhodesiense* SNP comparison data

Table A7.1: Comparison of BIOSCOPE and BOWTIE mapping algorithms for mapping SOLID *T. b. rhodesiense* sequences to the T. b. brucei TREU927/4 reference.

	Z310	B17
Mean coverage at SNP loci		
Bowtie	53x	52x
Bioscope	89x	94x
Number of SNP		
Bioscope (total)	203,049	209,415
Bioscope (>10x coverage)	190,658	197,472
Bowtie (total)	132,389	137,665
Bowtie (>10x coverage)	116,065	121,202
Total shared loci (>10x coverage)	100,674	105,300

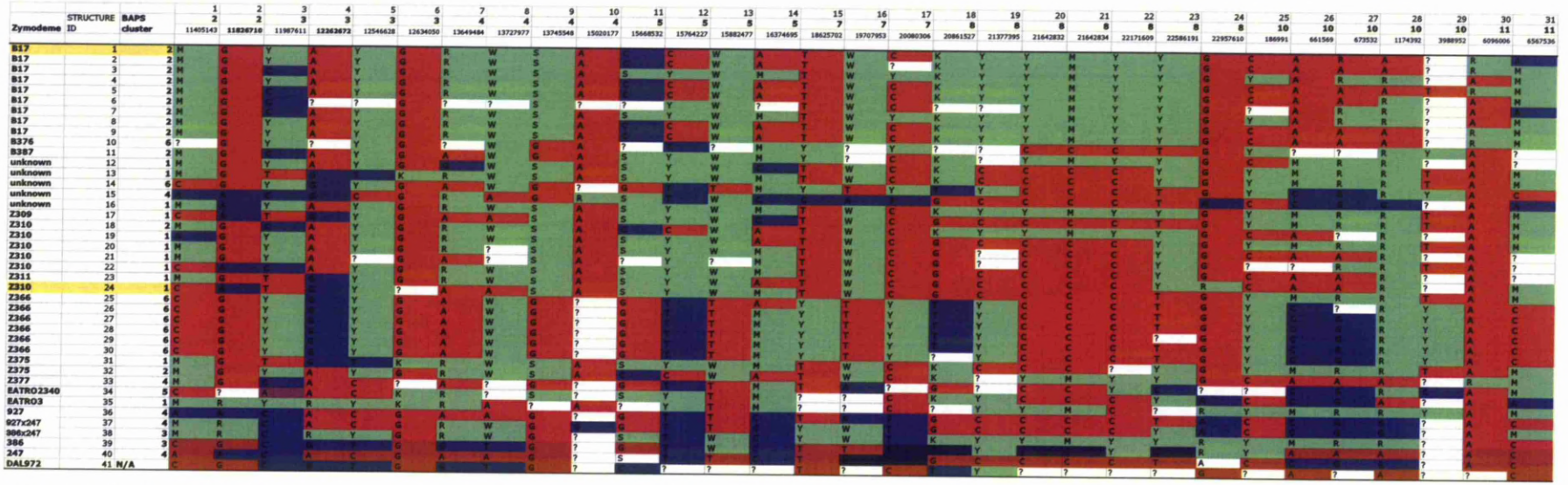


Figure A7.1.1: KASPAR Genotyping Data. Raw SNP data from KASPAR genotyping and next generation sequencing data. 31 informative SNP loci for 40 isolates are presented by allele at a given locus. IUPAC single letter codes for bases are used in all cases. Ambiguous bases (i.e. YRKWSM) represent heterozygous sites. Missing data is represented by a "?". Colour coding is as follows: Green represents heterozygous sites; Other colours (blue and red) represent the two homozygous alleles therein. Any third allele (i.e. not heterozygous, or the two, split homozygous alleles) is represented by purple. Samples tested by KASPAR genotyping are listed according to zymodeme (or unknown, if the data was not known or unavailable). Next generation sequencing data are listed by other identification as follows: EATRO3 = 1960 Ugandan *T. b. rhodesiense*; EATRO2340 = 1977 Kenyan *T. b. rhodesiense*; TREU927 = *T. b. brucei* TREU927; 927x247 = *T. b. brucei* TREU927 x STIB247 cross; 247 = *T. b. brucei* STIB247; 386 = *T. b. gambiense* (Type 2) STIB386 (All Illumina GA second-generation sequencing data). DAL972 = *T. b. gambiense* (Type 2) DAL972 (Sanger sequencing data). Highlighted strains (yellow) are those chosen for SOLID sequencing. SNP loci 20 and 21 are the heterozygous SNP predicted to cause non-synonymous polymorphisms in the isocitrate dehydrogenase gene.

Figure A7.1.2: Map of KASPAR genotyping loci.

Black lines represent loci designed for KASPAR genotyping that failed to amplify. Green lines represent loci that were successfully genotyped by KBiosciences Ltd.

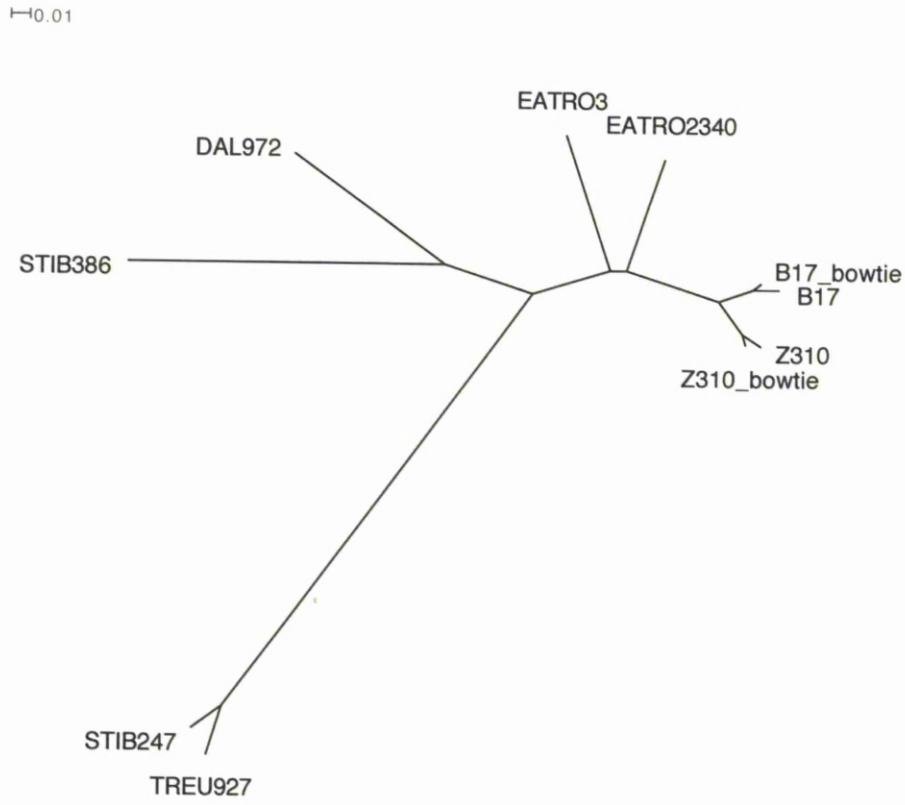


Figure A7.2.1: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 6,172 loci on chromosome 1. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

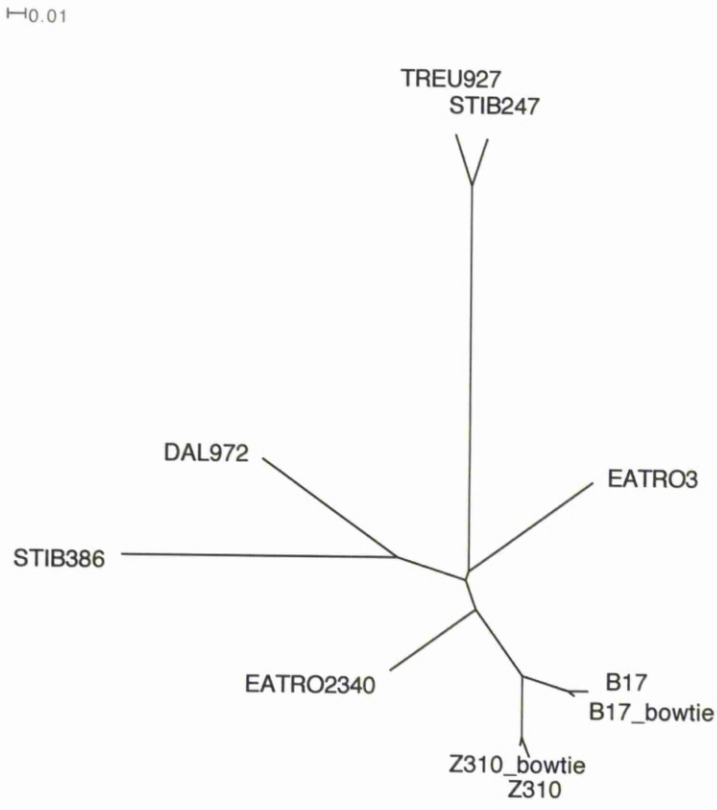


Figure A7.2.2: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 8,177 loci on chromosome 2. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

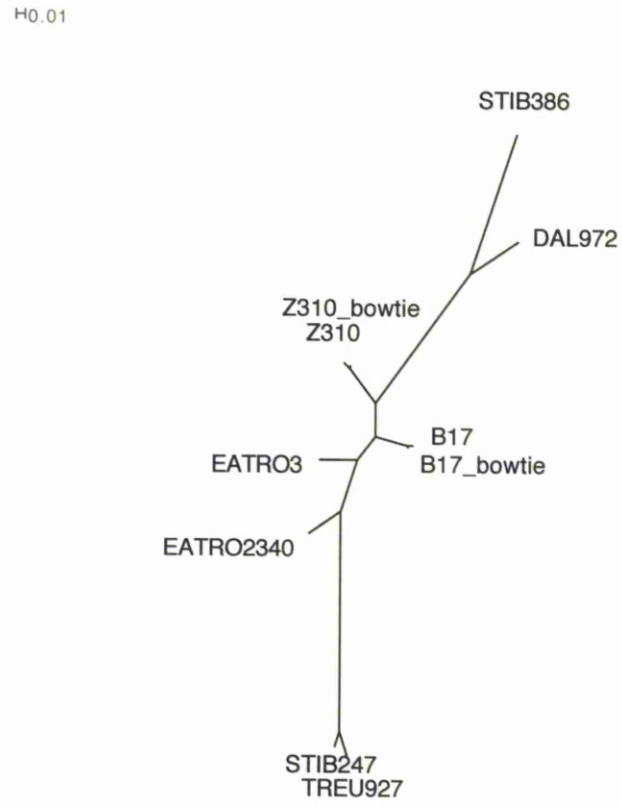


Figure A7.2.3: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 6,938 loci on chromosome 3. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

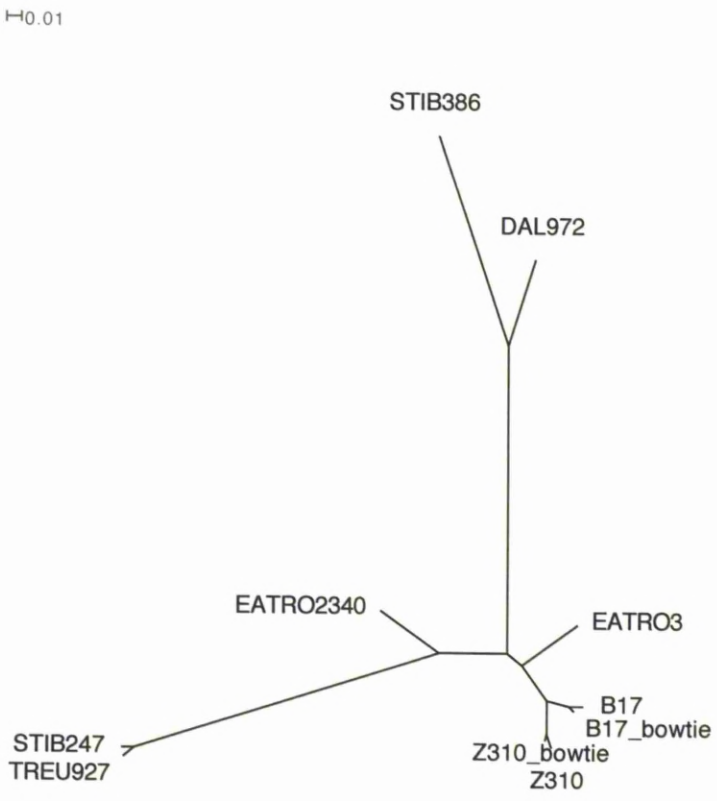


Figure A7.2.4: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 6,964 loci on chromosome 4. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

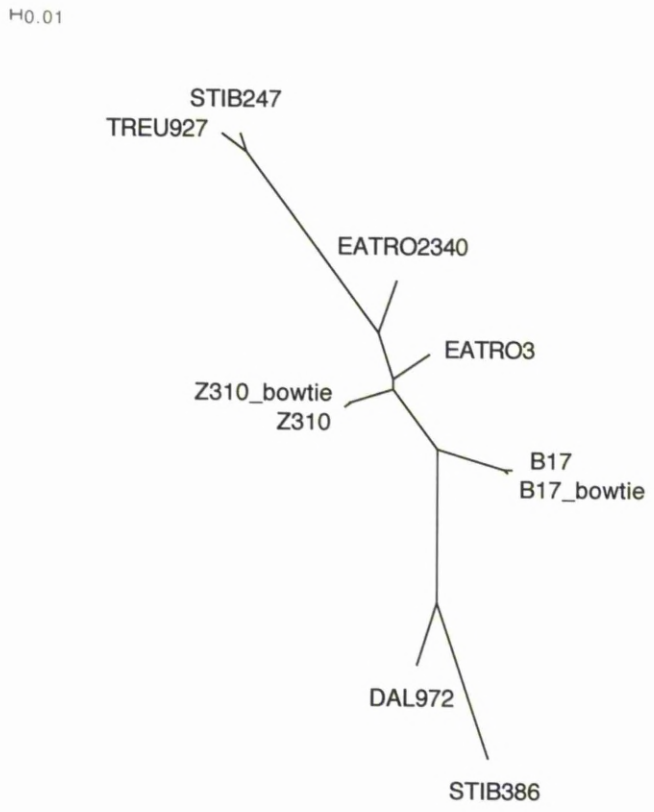


Figure A7.2.5: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 7,502 loci on chromosome 5. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

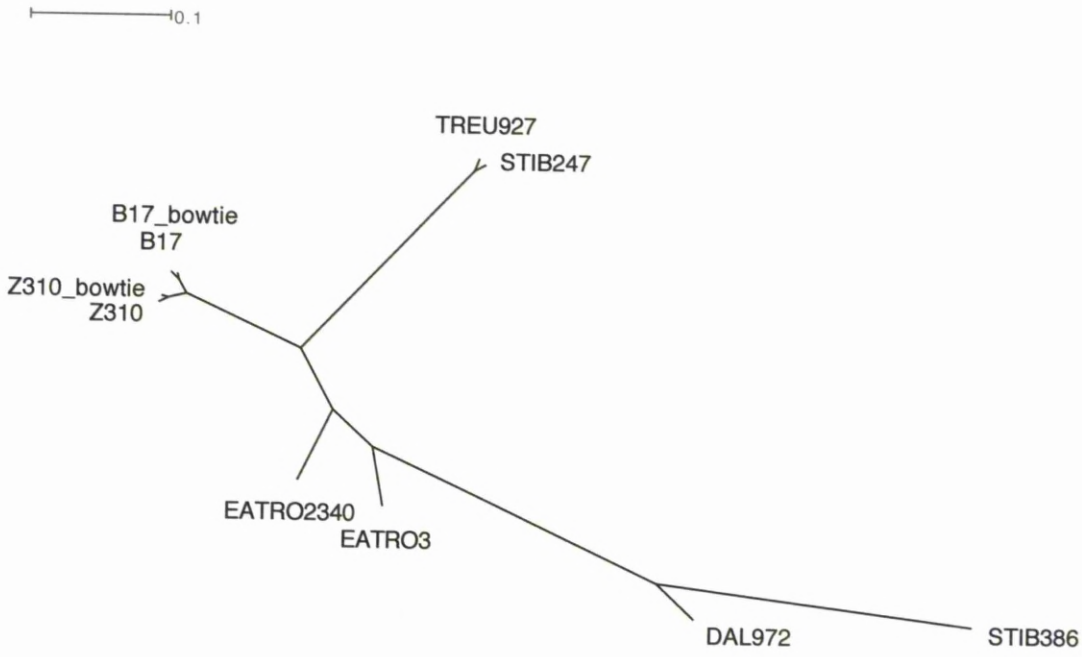


Figure A7.2.6: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 7,517 loci on chromosome 6. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

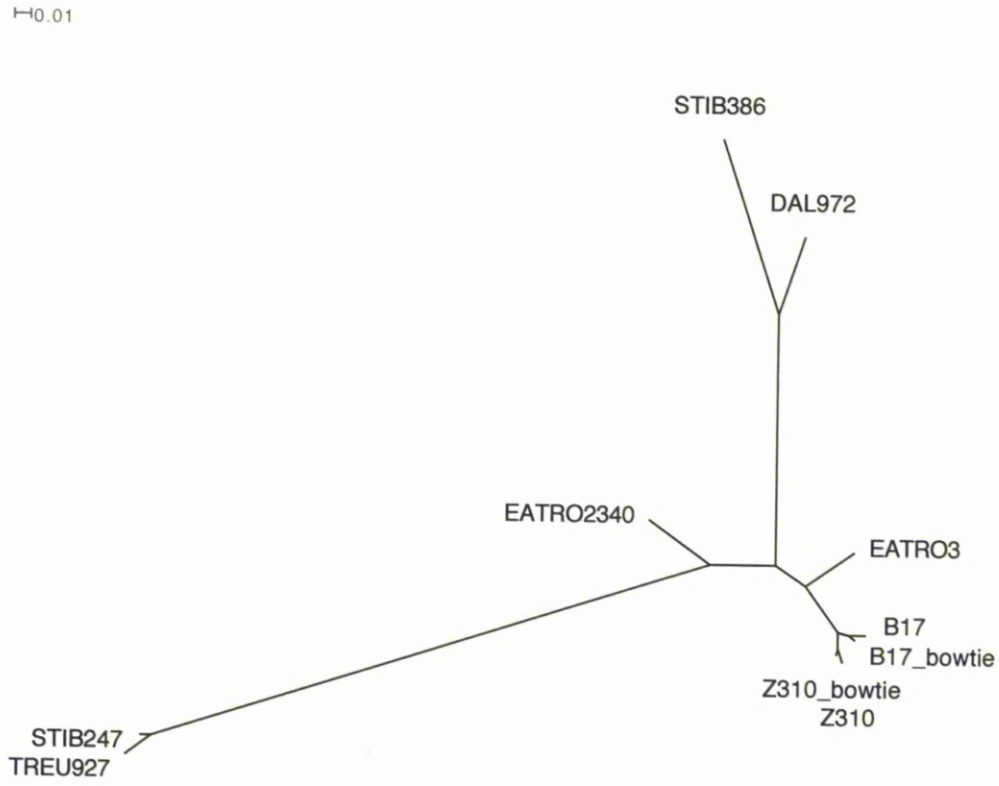


Figure A7.2.7: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 7,945 loci on chromosome 7. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

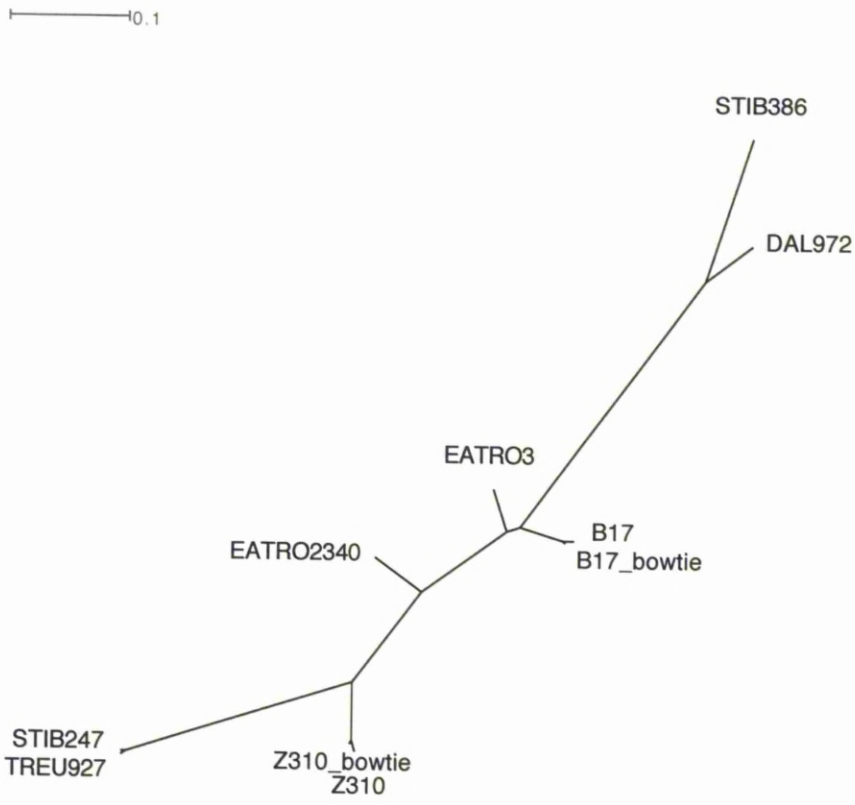


Figure A7.2.8: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 9,154 loci on chromosome 8. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

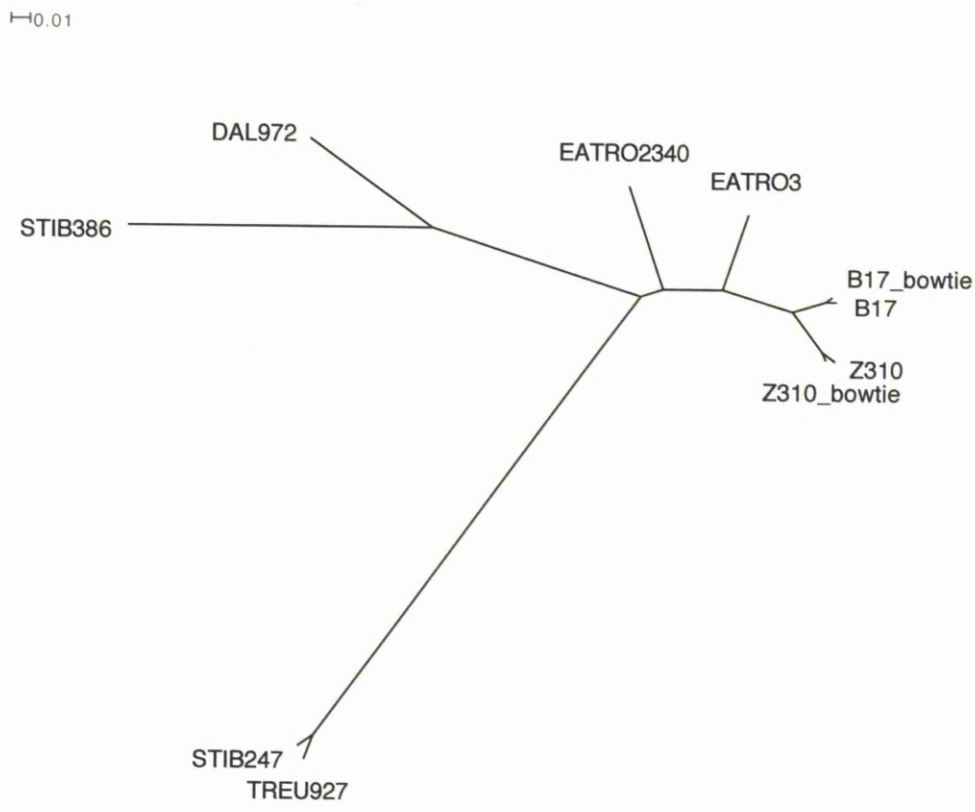


Figure A7.2.9: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 12,375 loci on chromosome 9. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

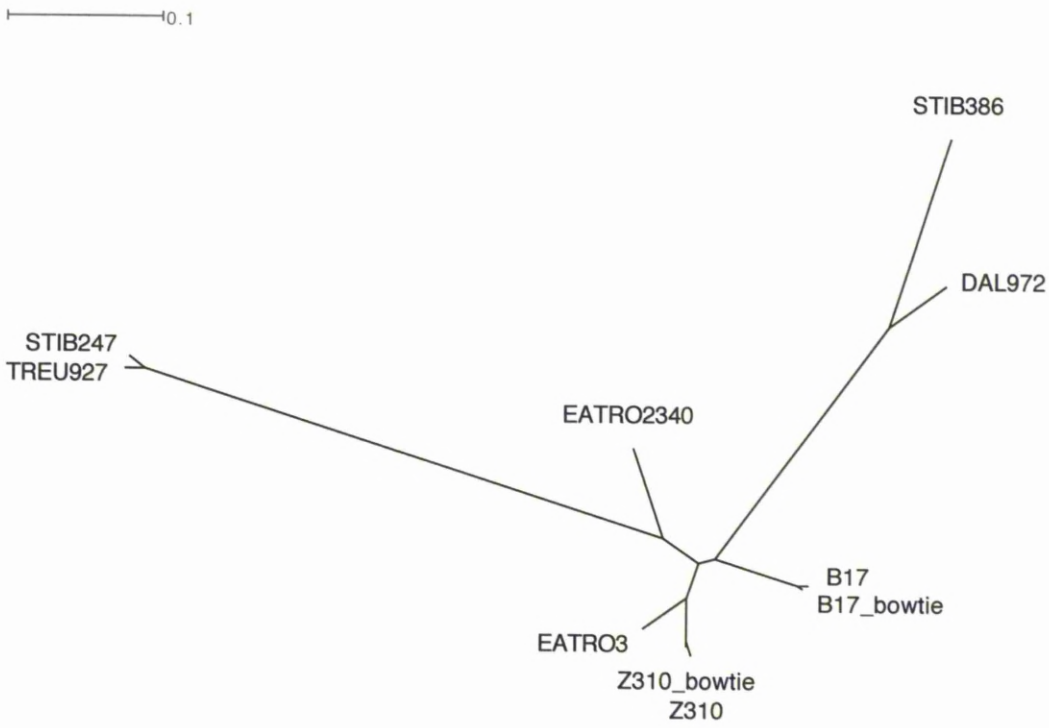


Figure A7.2.10: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 14,920 loci on chromosome 10. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

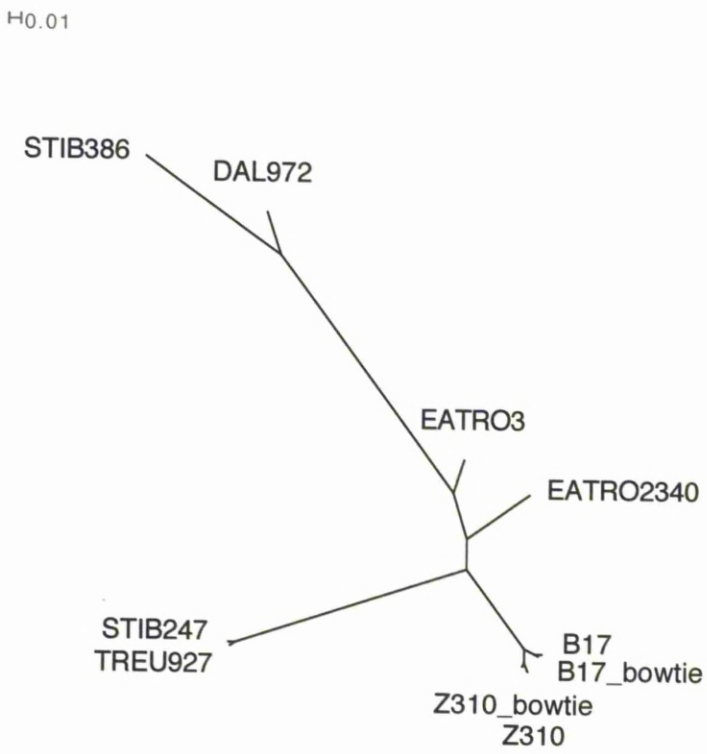


Figure A7.2.11: SPLITSTREE Jukes Cantor Neighbour Joining Tree for 21,830 loci on chromosome 11. "B17" and "Z310" are BIOSCOPE mapping results; "B17_bowtie" and Z310_bowtie" are the BOWTIE mapping results of the same SOLID data.

Appendix VIII: Accompanying Research Paper

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

A Comprehensive Genetic Analysis of Candidate Genes Regulating Response to *Trypanosoma congolense* Infection in Mice

Ian Goodhead[1], Alan Archibald[2], Peris Amwayi[3], Andy Brass[4,5], John Gibson[3¤a], Neil Hall[1], Margaret A. Hughes[1], Moses Limo[6], Fuad Iraqi[3¤b], Stephen J. Kemp[1,3], Harry A. Noyes[1]*

1 Centre for Genomic Research, School of Biological Sciences, University of Liverpool, Liverpool, United Kingdom, **2** The Roslin Institute, University of Edinburgh, Roslin, United Kingdom, **3** International Livestock Research Institute, Nairobi, Kenya, **4** Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom, **5** School of Computer Science, University of Manchester, Manchester, United Kingdom, **6** Egerton University, Njoro, Nakuru, Kenya

Abstract

Background: African trypanosomes are protozoan parasites that cause "sleeping sickness" in humans and a similar disease in livestock. Trypanosomes also infect laboratory mice and three major quantitative trait loci (QTL) that regulate survival time after infection with *T. congolense* have been identified in two independent crosses between susceptible A/J and BALB/c mice, and the resistant C57BL/6. These were designated *Tir1*, *Tir2* and *Tir3* for *Trypanosoma infection response*, and range in size from 0.9–12 cM.

Principal Findings: Mapping loci regulating survival time after *T. congolense* infection in an additional cross revealed that susceptible C3H/HeJ mice have alleles that reduce survival time after infection at *Tir1* and *Tir3* QTL, but not at *Tir2*. Next-generation resequencing of a 6.2 Mbp region of mouse chromosome 17, which includes *Tir1*, identified 1,632 common single nucleotide polymorphisms (SNP) including a probably damaging non-synonymous SNP in *Pram1* (PML-RAR alpha-regulated adaptor molecule 1), which was the most plausible candidate QTL gene in *Tir1*. Genome-wide comparative genomic hybridisation identified 12 loci with copy number variants (CNV) that correlate with differential gene expression, including *Cd244* (natural killer cell receptor 2B4), which lies close to the peak of *Tir3c* and has gene expression that correlates with CNV and phenotype, making it a strong candidate QTL gene at this locus.

Conclusions: By systematically combining next-generation DNA capture and sequencing, array-based comparative genomic hybridisation (aCGH), gene expression data and SNP annotation we have developed a strategy that can generate a short list of polymorphisms in candidate QTL genes that can be functionally tested.

Citation: Goodhead I, Archibald A, Amwayi P, Brass A, Gibson J, et al. (2010) A Comprehensive Genetic Analysis of Candidate Genes Regulating Response to *Trypanosoma congolense* Infection in Mice. PLoS Negl Trop Dis 4(11): e880. doi:10.1371/journal.pntd.0000880

Editor: Christian Tschudi, Yale School of Public Health, United States of America

Received July 7, 2010; **Accepted** October 12, 2010; **Published** November 9, 2010

Funding: IG is supported by a Royal Society Wolfson Merit award to NH and a University of Liverpool funded Studentship (http://royalsociety.org). Funding to NH was provided by MRC grant G0900753 (http://www.mrc.ac.uk). Funding to HAN, SJK, JG, AB was provided by Wellcome Trust (GR066764MA to SK); (http://www.wellcome.ac.uk/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: harry@liv.ac.uk

¤a Current address: The Centre For Genetic Analysis and Applications, C.J. Hawkins Homestead, University of New England, Armidale, Australia
¤b Current address: Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Tel Aviv, Israel

Introduction

African trypanosomiasis is a disease of both livestock and humans, largely caused by three species of *Trypanosoma* parasites. Two subspecies of *T. brucei*: *T. b. gambiense* and *T. b. rhodesiense*, cause severe disease in humans, whilst disease in livestock is mainly caused by two other species: *T. vivax* and *T. congolense*. The diseases affect over ten million Km^2 of Africa and it is estimated that some thirty percent of Africa's 160 million cattle are at risk of infection. Losses of livestock and crop production are estimated at over $1 billion per annum [1].

Some indigenous breeds of cattle, notably N'Dama (*Bos taurus*), have the ability to tolerate the effects of an infection by *Trypanosoma* parasites, and remain productive. Other, introduced, breeds are much more susceptible, and quickly show the classic symptoms of infection, such as anaemia, fatigue and muscle wastage [2]. This effect is under genetic control, and ten quantitative trait loci (QTL) have been mapped in F2 crosses between the N'Dama and susceptible Boran cattle (*Bos indicus*) [2].

Scientists are aided by a mouse model of trypanotolerance, as African trypanosomes also infect laboratory mice in which susceptibility is measured by survival time after infection, which varies between inbred lines. Whilst C57BL/6 mice survive for a relatively long period after infection with *T. congolense* (110 days), some other strains, such as A/J (16 days), 129/J (23 days), BALB/c (49 days) and C3H/HeJ (59 days) mice are relatively susceptible [3,4,5]. Mapping studies, initially undertaken in two independent F2 crosses: C57BL/6JOlaHSD (C57BL/6) × BALB/cOlaHsd

Author Summary

About one-third of cattle in sub-Saharan Africa are at risk of contracting "Nagana"—a disease caused by *Trypanosoma* parasites similar to those that cause human "Sleeping Sickness." Laboratory mice can also be infected by trypanosomes, and different mouse breeds show varying levels of susceptibility to infection, similar to what is seen between different breeds of cattle. Survival time after infection is controlled by the underlying genetics of the mouse breed, and previous studies have localised three genomic regions that regulate this trait. These three "Quantitative Trait Loci" (QTL), which have been called *Tir1*, *Tir2* and *Tir3* (for *Trypanosoma Infection Response* 1–3) are well defined, but nevertheless still contain over one thousand genes, any number of which may be influencing survival. This study has aimed to identify the specific differences associated with genes that are controlling mouse survival after *T. congolense* infection. We have applied a series of analyses to existing datasets, and combined them with novel sequencing, and other genetic data to create short lists of genes that share polymorphisms across susceptible mouse breeds, including two promising "candidate genes": *Pram1* at *Tir1* and *Cd244* at *Tir3*. These genes can now be tested to confirm their effect on response to trypanosome infection.

(BALB/c) and C57BL/6JOlaHSD × A/JOlaHsd (A/J), identified three major QTL regulating survival time [6]. These were mapped to mouse chromosomes 17, 5 and 1 and have been designated *Tir1*, *Tir2* and *Tir3* respectively for *Trypanosoma Infection Response*. These loci were further refined to five smaller regions using advanced intercross lines of the same crosses that were extended to the F6 and then F12 generations, in which *Tir3* was resolved into three smaller regions, termed *Tir3a*, *Tir3b* and *Tir3c* [7,8]. Whilst these studies substantially reduced the size of the 95% confidence interval of each of the QTL to between 0.9 and 12 cM, each one still includes 17 to 650 candidate genes.

Moving from well defined QTL regions to QTL genes is still a major challenge: over 2,750 such quantitative trait loci have been mapped in mice and rats but fewer than 1% have been characterised at the molecular level [9]. However, new sequencing technologies are making it possible to identify a large proportion of the differences between common inbred mouse strains. At present this is possible for defined areas of the genome, but public data sets will soon be available for the whole genome. We have used a combination of these methods and resources to demonstrate how large QTL regions can be reduced to tractable short lists of candidate genes for functional analysis.

We have mapped QTL in a C57BL/6 × C3H/HeJ cross so that we now know whether four mouse strains carry either the susceptible or the resistant allele at each QTL. This will reduce the number of polymorphisms that correlate with phenotype at any given QTL. The haplotype structure of the QTL regions has been determined using the 8 million public SNP from 16 mouse strains in the Perlegen set and identified regions where haplotypes correlate with survival time in the four mouse strains studied. Copy number variations (CNV) have been shown to be responsible for a significant number of quantitative traits [10]. We have used array comparative genomic hybridisation (aCGH) to identify CNV in QTL regions that correlate with survival in the four mouse strains. We have also correlated CNV with existing gene expression data from three of the mouse strains [11] to identify CNV that putatively cause expression differences. Finally we have sequenced one of the QTL regions in four strains of mice to identify SNP that correlate with phenotype and validated these against an additional publicly available dataset [12,13]. We have also used Polyphen to identify the non-synonymous SNP in the QTL regions that are most likely to change the activity of the protein.

By combining additional mapping with haplotype analysis, aCGH and resequencing we have reduced the initial long list of genes within QTL regions to a short list of candidate genes with defined genetic differences that correlate with phenotype. It is now practical to test the function of these genes and polymorphisms to determine their role in response to infection with *T. congolense*. The Perlegen and aCGH data is already publically available for many mouse strains and the Wellcome Trust Sanger Institute is resequencing the genomes of the common laboratory mouse strains so this strategy will soon be applicable to many QTL without further experimental work [13,14,15].

Methods

Ethics statement

All animal work was undertaken under IACUC ref no 2003.19. The ILRI IACUC complies voluntarily with the UK Animals (Scientific Procedures) Act 1986 that contains guidelines and codes of practice for the housing and care of animals used in scientific procedures. All animals on survival experiments were regularly monitored to check for signs of terminal illness, and any showing such signs were euthanised by UK Schedule 1 procedures.

C3H/HeJ × C57BL/6 cross

C57BL/6JOlaHSD (C57BL/6) and C3H/HeJ mice were obtained from Harlan Laboratories. Mice were infected with 4×10^4 *T. congolense* strain IL1180 intra-peritoneally (ip) as previously described [6]. Any mice that did not develop a microscopically proven parasitaemia were removed from the study.

345 F2 C3H/HeJ × C57BL/6 mice were phenotyped for survival time after infection with *T. congolense* strain IL1180. 94 animals that had extreme survival times (≤62 days and >140 days) were selected for genotyping using the markers shown in Table S1 in Supporting Text S1. Selective genotyping significantly reduces genotyping costs with little loss of power to detect QTL, however it does give exaggerated estimates of effect sizes [16]. The F2 mice were also genotyped at the *Tlr4* locus since C3H/HeJ carries a proline to histidine mutation at position 712 of the *Tlr4* gene that makes this mouse strain insensitive to LPS and might modify response to infection with *T. congolense* [17].

PCR reactions were performed using Reddymix (Thermo) with 20 ng of template DNA. Cycling conditions were as follows: 95°C, 50 secs; [Tm −5]°C, 50 secs; 65°C, 50 secs; 30× cycles. PCR products, including negative controls, were resolved by ethidium bromide stained agarose-gel electrophoresis and visualised under UV-light. SNP were genotyped by sequencing PCR products using primers shown in Table S2 in Supporting Text S1. Unincorporated primers and residual nucleotides were degraded using ExoSAP-IT (USB Corp, Ohio, USA) and sequencing products generated using Big-Dye v3.1 terminators (Applied Biosystems, Foster City, USA). Cycle sequencing products were ethanol precipitated and subject to electrophoresis on an Applied Biosystems ABI-3130XL capillary sequencer. Microsatellite and SNP genotyping data was viewed using PeakScanner (Applied Biosystems) and GAP4 [18] respectively.

Allocation of strains to haplotypes

Strains were allocated to haplotypes as previously described [19,20]. Briefly, Perlegen SNP and haplotype boundaries were

downloaded from Perlegen [15]. Strains were allocated to haplotypes for each haplotype block using a local Perl script that extracted all alleles from the Perlegen dataset within a haplotype block, substituted them into the C57BL/6 reference sequence and submitted the resulting aligned sequences to the Jukes-Cantor algorithm in DNADIST in PHYLIP to calculate genetic distances between each pair of strains [21]. Strains were given a binary "barcode" with all possible pairs of strains assigned a 1 or a 0 depending on whether the genetic distance for that pair was above or below a threshold value. Strains that had the same "barcode" were allocated to the same haplotype number. C57BL/6 was used as the reference strain for block allocation and assigned to haplotype one; Succeeding strains were allocated to the same haplotype block as another strain that they shared a haplotype with or, if there was none, to the next available haplotype number (Full details are available in Supporting Text S1; Allocation of strains to haplotypes).

CNV discovery

Array CGH was performed using the Agilent Mouse Genome CGH Microarray 244A platform. Dye-flip replicates were carried out on the C57BL/6 reference strain and three test strains (129P3/J, A/J and BALB/cJ) and analysed as previously described [22]. Overlapping aberrations were grouped into CNVR (t-test analysis, $P \leq 0.05$, Overlap 0.9) using the Agilent CGH analytics software (v 4.0) and using the ADM-2 algorithm (threshold 6.0) using centralization (threshold 6.0, bin size 1) and Fuzzy Zero [23]. CGH array data have been submitted to the NCBI Gene Expression Omnibus database (GEO) [GEO: GSE9669].

DNA capture and sequencing

Genomic DNA for BALB/cJ (Jackson #000651), 129P3/J (Jackson #000690), A/J (Jackson #000646) and C3H/HeJ (Jackson #000659) were obtained from the Jackson Laboratories and submitted to Nimblegen for sequence capture [24]. Capture probes were designed to cover 4.5 Mbp of non-repetitive sequence between 30,637,692 bp and 36,837,814 bp on Mmu17 (NCBI37). 385,000 60mer probes were tiled at approximately 5 bp intervals leading to a mean of 12 probes over each base. Captured DNA was sequenced on a Roche 454 FLX Genome Sequencer using Titanium chemistry (Roche). Sequence assembly and SNP calling was performed using the Newbler mapping algorithm, which aligned 454 reads against the Ensembl C57BL/6 reference (NCBI37) and outputs lists of SNP and associated coverage metrics.

As pyrosequencing is known to miscall sequences either across, or either side of, homopolymeric tracts (long stretches of a single nucleotide), discrepancies were removed from subsequent analysis if they were within 13 bp of a homopolymeric tract ≥ 5 bp [25]. SNP were additionally filtered to those with at least an eight-fold coverage and occurring in at least 87.5% of the reads sequenced across any polymorphic position. 14,440 high-confidence genotypes were submitted to dbSNP with SSIDs ss159831440-ss159845897. 454 reads were submitted to the European Short Read Archive under Accession number ERA000179.

SNP were aligned against coding sequences and non-synonymous SNP were identified. SNP positions were compared to the mouse regulatory build to test for SNP that may alter transcription factor binding sites or promoter regions [26,27]. A 24-bp insertion in *Mdc1* in susceptible strains was amplified by PCR and verified by agarose gel electrophoresis, but could not be shown to have any functional effect (data not shown).

Identification and annotation of single nucleotide polymorphisms (SNP)

SNP outside the *Tir1* region were obtained from the 8 million Perlegen SNP set [15]. phastCons conservation scores for SNP positions [28] were obtained from UCSC [29]. These scores are a measure of how conserved a position is amongst 30 mammalian species and are on a scale between 0–1 with the most conserved positions scored as 1.

SNP within exons were annotated using the Ensembl SNP annotation API to identify non-synonymous SNP (nsSNP) and SNP in splice sites. nsSNP in the 454 data were identified with a local Perl script. Publically available functional SNP were also obtained from BioMart and the Wellcome Trust Sanger Institute website [12].

nsSNP were annotated with Polyphen [30] using the Polyphen batch submission tool. Publicly available functional SNP identified at QTL for which complete genotypes were not available were confirmed in C57BL/6, A/J, BALB/cJ and 129P3 mice using PCR and dideoxynucleotide sequencing as described for genotyping. Sequences which showed evidence of multiple copies were cloned using TOPO-TA cloning kit (Invitrogen) and sequenced.

Measurement of gene expression

Gene expression was measured for A/J, BALB/c and C57BL/6 mice before infection and at four time points post infection on Affymetrix 450_2 microarrays as previously described [11]. All microarray data has been deposited at ArrayExpress under the accession number E-MEXP-1190. The expression data and plots like those presented here are also available for all genes on the microarrays from the authors' website [31].

Results

Refining numbers of candidate genes within the *Tir* QTL

Determination of QTL boundaries and initial candidate gene identification. Different locations of the *Tir2* and *Tir3a,b,c* QTL have been published at the F6 and F12 generations [7,8]. QTL have also been physically mapped using congenic mice [32]. The congenic data supports the F6 location in one case (*Tir2*) and the F12 location in one other (*Tir3a*). Consequently we have annotated genes under both definitions of QTL positions and discuss their relative merits, case by case, below. In order to refine the number of candidate genes within *Tir* QTL it is necessary to first convert the 95% confidence intervals of the QTL from centiMorgan (cM) positions to megabase (Mbp) positions. Whilst the exact assignment of physical boundaries to the QTL is not possible, we have used the physical position of the peak marker in the F6 and F12 advanced intercross studies [7,8] as the most likely position of the peak of the QTL. We estimated the physical size of the 95% confidence interval (CI) by using Mouse Genome Informatics data to find the median Kbp/cM ratio for the intervals between the ten flanking markers (which were spaced at ~0.3 Mbp intervals). This ratio varied between 0.69–5.43 Mpb/cM and was used to convert the 95% CI in cM to Kbp. These positions are then used to identify the candidate genes contained within the QTL prior to further refinement (Table 1).

Identification of QTL in C3H/HeJ mice

By increasing the number of breeds known to carry susceptible alleles at the QTL, candidate gene lists can be refined to remove those genes that are in QTL for *T. congolense* infection response but have the same ancestral haplotype as the resistant strain in at least one susceptible mouse breed. The three major *Tir* QTL have only been identified in C57BL/6, A/J and BALB/c mice, with

Table 1. Physical locations of QTL and counts of candidate genes.

QTL	*Tir1*	*Tir2*-F6	*Tir2*-F12b	*Tir3a*-F6	*Tir3a*-F12	*Tir3b*-F6	*Tir3b*-F12	*Tir3c*-F6	*Tir3c*-F12
Chromosome	17	5	5	1	1	1	1	1	1
Peak marker	D17Mit16	D5Mit114	D5Mit58	D1Nds2a	DiMit286	D1Mit102	D1Mit102-DiMit105	D1Mit113	D1Mit107-DiMit16
95%CI (cM)	0.9	12	1	1.8	6	10	7	8	2
Median Mbp per cM	1.04	1.77	1.46	3.9	1.93	5.49	1.92	0.69	1.16
Start (Mbp)	33.27	71.02	73.45	100.54	124.71	121.63	148.15	170.96	164.3
End (Mbp)	34.2	92.3	73.91	107.57	136.19	176.56	161.44	176.51	166.6
Size (Mbp)	0.93	21.25	1.46	7.03	11.56	54.93	13.44	5.54	2.23
# Genes	43	210	27	20	127	650	113	143	35
Number of Candidate Genes (H1)	0	42	12	10	33	144	30	54	8
Number of Candidate Genes (H2)	27	74	14	10	63	355	61	122	8
454 Sequencing Data									
Common SNP (d)	194								
Common nsSNP	2								
Additional Data from Illumina Comparison [13]									
nsSNP	0								
5′-UTR SNP	0								
Synonymous SNP	2								

Positions were interpolated using NCBI37 from peak marker positions and 95% confidence intervals. The physical position of the D1Nds2 marker is not known, so its position was estimated from the intervals between its flanking markers. Lists of the genes with different haplotypes are shown in the Supplementary Data S2: *GenesAndHaplotypes.xls*. [a] Number of SNP common to the three susceptible strains of mice: A/J; BALBc/J and C3H/HeJ. [b]At Tir2-F12 we have estimated the physical 95% confidence interval around the D5MIT58 peak marker and this 1.46 Mb region contained 27 genes, however the exact position of the peak is hard to identify since both D5MIT58 and DMIT258 are at 41 cM in the MGI map although they are 7 Mb apart on the physical map.
Numbers of candidate genes were calculated under two hypotheses: Hypothesis 1: all four susceptible strains have the same haplotype as each other and different from C57BL/6. Hypothesis 2: All susceptible strains have a different haplotype from C57BL/6 but not necessarily the same as each other. Hypothesis 1 is a special case of hypothesis 2 and all genes included under hypothesis 1 are also included under hypothesis 2. Only A/J and BALB/c are known to carry susceptibility alleles at *Tir2* and so at this locus only the correlation of C57BL/6, A/J and BALB/c was considered. [e] nsSNP loci submitted to dbSNP.
doi:10.1371/journal.pntd.0000880.t001

C57BL/6 carrying the resistant allele at each locus. To that end, we measured survival after infection in an inter-cross between another susceptible breed, C3H/HeJ, and C57BL/6 mice. For the cross, the mean survival times of parental founder lines for the C3H/HeJ × C57BL/6 F2 cross were 63 days for C3H/HeJ and 87 days for C57BL/6. Out of the 345 F2 C3H/HeJ x C57BL/6 mice that were phenotyped, we selectively genotyped the 94 mice (51♂ and 43♀; $p = 0.41$) that had the most extreme survival times (Table S1 in Supporting Text S1) with microsatellite and SNP markers across the three known QTL. Table 2 shows that C3H/HeJ carries alleles that reduce survival time at the *Tir3* QTL on Mmu1 and the *Tir1* QTL on Mmu17. No QTL was discovered on Mmu5 in the region of *Tir2*.

Refining numbers of candidate genes by allocation of alleles to haplotype blocks

Over eight million SNP and haplotype block boundaries derived from them have been published for the whole mouse genome [15], however the haplotype alleles carried by each strain are not

Table 2. Loci regulating survival after *T. congolense* infection in the C3H/HeJ × C57BL/6 cross.

Chr	F-value	LOD score	95% CI (cM)	QTL position (cM)	QTL effect days	Peak marker
17	17.22	6.344	17	16	32	D17mit81
1	9.13	3.614	47	94.9	24	D1mit356

94 mice were genotyped with markers across known QTL regions but not elsewhere in the genome. QTL effects are the mean number of days difference in survival between mice that are homozygous for the alternate alleles at a QTL. Positive QTL effects indicate that longer survival was associated with C57BL/6 alleles. The QTL effects are likely to be biased upwards as a consequence of selective genotyping of the extremes of the phenotypic distribution [16]. Phenotype distribution is shown in Figure S1 in Supporting Text S1.
doi:10.1371/journal.pntd.0000880.t002

available on a genome-wide basis. Full details of the allocation of strains to haplotype blocks and associated figures are available in Supporting Text S1; Allocation of Strains to Haplotypes. Results are briefly presented here:

In order to identify haplotype alleles that correlated with phenotype we obtained the Jukes-Cantor distance between each pair of the four mouse strains (C57BL/6, A/J, BALB/c and C3H/HeJ) for each haplotype block across each QTL. The distribution of the natural logarithm of Jukes-Cantor distances was approximately normal and the fifth percentile of the distribution corresponding to a distance of 5×10^{-5} was selected as a threshold (Figure S2 in Supporting Text S1). Strains were allocated to the same haplotype allele if the Jukes-Cantor distance between them at that block was less than 5×10^{-5}.

Under the assumption that where QTL coincide in multiple crosses, it is likely that it is due to the same polymorphism in all breeds tested, there are only two possible distributions of resistant and susceptible haplotypes. In the present case, C57BL/6 was the only strain carrying a haplotype for longer survival at any QTL, therefore either: all susceptible breeds have the same haplotype that is different from C57BL/6 (hypothesis one); or C57BL/6 has a unique haplotype that differs from all susceptible breeds but that these might differ amongst themselves (hypothesis two).

A list was compiled of the genes for which the susceptible strains were on a different haplotype or immediately upstream or downstream of a different haplotype from the resistant (C57BL/6) strain (Supplementary Data S2: *GenesAndHaplotypes.xls*). Table 1 shows the number of candidate genes in each locus under the two nested hypotheses: H1) Short survival time is caused by a common deleterious allele in all three susceptible strains; or H2) that the difference in survival is attributable to a beneficial allele in the single long surviving line (C57BL/6) and the susceptible lines may carry any non C57BL/6 haplotype. Under H1 the number of candidate genes was reduced from 1193 to 283 and under H2 the number was reduced from 1193 to 651.

A null allele of the *Tlr4* gene in C3H/HeJ does not affect survival

A functional toll like receptor 4 (*Tlr4*) gene is necessary for maximal control of *Trypanosoma cruzi* in mice [33] and there is evidence that the GPI anchor of *T. brucei* VSG has endotoxin like properties that could stimulate *Tlr4* [34]. C3H/HeJ has a polymorphism in the *Tlr4* gene, on mouse chromosome four, which ablates its function, making these mice insensitive to LPS [17]. We used this spontaneous mutation to discover whether *Tlr4* was as important in the response to *T. congolense* as to *T. cruzi*. Since all previous mapping had been done in mice with intact *Tlr4* genes, no QTL could have been detected at this locus even if *Tlr4* does modulate the response to infection. The C3H/HeJ × C57BL/6 mapping population could therefore be used to discover whether this gene (or a closely linked one) is involved in the regulation of survival time after infection. Mice were genotyped with a microsatellite marker linked to the functional polymorphism and sequenced across the polymorphic position. There was no association with either of these markers and survival time, indicating that the *Tlr4* pathway does not affect survival after *T. congolense* infection in mice (Table S1 in Supporting Text S1).

Comparative genomic hybridisation and gene expression

To assess the impact of copy number variation regions (CNVR) upon the expression of genes that may influence response to *T. congolense* infection we performed array-based comparative genomic hybridisation (aCGH) on the complete genome of three mouse strains: 129P3, A/J and BALB/c, relative to C57BL/6. The expression of genes within CNVR in A/J, BALB/c and C57BL/6 mice over the course of infection was evaluated using a previously described dataset [11].

Genome-wide, one hundred and twenty-nine CNVR involving three or more probes were common to A/J, BALBc/J and 129P3/J. These encompassed a total of 317 genes, and ranged in size from 400 bp to 6.4 Mbp, although 96% were smaller than 1 Mbp. Twelve CNVR containing the complete coding sequences of genes and that had corresponding differences in gene expression, were common in all susceptible breeds of mice tested. Lists of the genome-wide CNVR is shown in Table S8 in Supporting Text S1.

One significant CNVR was detected close to the peak of *Tir3c* in the F6 population (D1Mit113: 173,734,611 bp). A two to four-fold reduction in C57BL/6 copy number relative to A/J, BALB/c and 129P3/J encompassed, or overlapped with, the coding sequences of *Itln1* (intelectin 1), *Cd244*, and *AC083892.19-1* and may affect the nearby *Ly9* (lymphocyte antigen 9) (173,441,746-173,499,029 bp; 11 probes; $p=0.0003$; Figure 1A). There were expression differences in *Cd244* (Figure 2A), but not *Itln1* or *Ly9* [31], over the course of infection between resistant C57BL/6 and susceptible A/J and BALB/c. *AC083892.19-1* was not on the expression microarray. This CNV region has also been previously reported by Graubert *et al* [14] who showed that an additional susceptible strain, C3H/HeJ, carries the same variant as A/J and BALB/c.

No common CNVR were detected within *Tir1* or *Tir2*. The CNVR that was previously reported to be the cause of differential expression of Glyoxalase 1 (*Glo1*) [35] and is 2.8 Mbp from the peak of *Tir1*, was detected as a two to fourfold reduction in copy number for C57BL/6 and BALB/c relative to A/J and 129P3 (Chr17: 30,176,153 bp–30,650,413 bp; 68 probes; $p<0.001$; Figure 1B). Since the CNVR did not correlate with phenotype, this polymorphism is unlikely to contribute to the difference in response to infection.

Identification of functional SNP

Lists of published non-synonymous SNP (nsSNP), SNP in splice sites; and regulatory regions and SNP that cause gain or loss of stop codons were obtained from BioMart. nsSNP were annotated using Polyphen [30] in order to identify those most likely to modify gene function. A complete list of annotated SNP is in Supplementary Data S1: *AnnotatedFunctionalSNP.xls*. Polyphen classifies nsSNP as benign, possibly damaging or probably damaging according to the likelihood that the polymorphism will modify protein activity. 'Damaging' implies a change of activity or function but this change could be beneficial to the animal.

Tir1. The physical size of the 95% CI for *Tir1* based on the combined data from the A/J × C57BL/6 and BALB/c × C57BL/6 F6 crosses [7] was 930 Kbp and contained 43 genes. *Tir1* was not reassessed with the F12 data. Assessing the Perlegen dataset against the smallest *Tir1* definition, none of the genes had haplotypes that correlated with phenotype under hypothesis 1, but there were 27 genes that correlated with phenotype under hypothesis 2 (Supplementary Data S2: *GenesAndHaplotypes.xls*). SNP that might modify phenotype at *Tir1* are discussed under sequencing of *Tir1* below.

Tir2. The *Tir2* QTL contained 210 genes in the 21.25 Mb (F6) QTL or 27 genes in the 1.46 Mb (F12) region, which was a subset of the F6 region. Congenic mice that were bred to physically map the *Tir2* QTL had a region of C57BL/6 DNA in an A/J background between 75.1 Mb and 89.7 Mb on chromosome 5 [32]. This was within the large F6 QTL (71.0–92.3 Mbp) but distal to the much smaller F12 QTL (73.5–73.9 Mbp). Since the QTL was physically mapped in the congenic

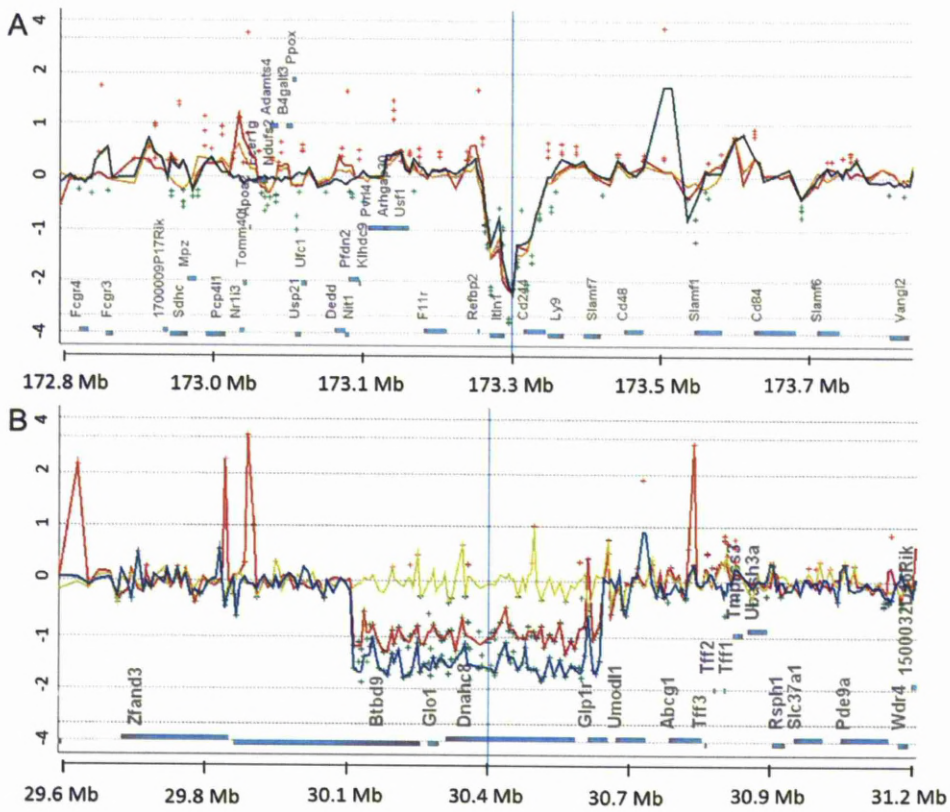


Figure 1. CNV plots from Agilent DNA Analytics software. A: Reduced copy numbers in C57BL/6 of *Itlnb* and *Cd244* near *Tir3c* relative to two susceptible breeds of mice (Chr 1: 172,831,532–173,931,532 bp). B: CNV data at the proximal end of *Tir1* showing a deletion of *Glo1* and *Dnahc8* in C57BL/6 and BALB/c relative to A/J and 129P3. (Chr 17: 29,854,972–30,954,972). Probes are plotted at their genomic position relative to their respective $\log_2$ fluorescence intensity ratios (Y-axis) along with genes on the x-axis (filled blue rectangles). Green dots are negative ratios and red dots positive ratios (threshold 0.5). Lines are a moving average over a 10 Kbp window for A/J (blue); 129P3 (red) and BALBc (yellow). Genomic positions are based on mouse build mm8 (NCBI36).
doi:10.1371/journal.pntd.0000880.g001

mice, they are expected to provide a more accurate prediction of location than genetic mapping methods. There were 21 and 52 genes consistent with hypotheses 1 and 2 respectively within the congenic region (Supplementary Data S2: *GenesAndHaplotypes.xls*). There were probably damaging nsSNP in *Srp72* (signal recognition particle 72 kDa) and *Ugt2b38* (UDP glucuronosyltransferase 2 family, polypeptide B38) (Supplementary Data S1: *AnnotatedFunctionalSNP*). The SNP in *Ugt2b38* and *Srp72* had phastCons scores of <0.1 and 0.998 respectively indicating that the *Srp72* was in a highly conserved position. Therefore the *Srp72* SNP was the SNP with the greatest probability of having an effect on gene function in the *Tir2* congenic region, although what this might be and whether it would modify response to *T. congolense* is not known.

Tir3a. The F6 *Tir3a* locus, at around 103 Mbp on chromosome 1, is within a region that was tested for its effect on survival after *T. congolense* infection by breeding congenic mice that had a fragment of C57BL/6 origin between 93–123 Mbp on an A/J background [32]. There was no difference in survival between mice that carried the region derived from C57BL/6 and littermate controls without the C57BL/6 region indicating that the F6 region was not likely to contain the QTL gene. The F12 *Tir3a* locus was distal to the congenic region and is consequently a more likely candidate region for this QTL than the F6 QTL. It contains 33 and 63 candidate genes under hypotheses 1 and 2 respectively. These include *IL10, Cd55* (complement decay-accelerating factor) and *Cxcr4* (CXC chemokine receptor 4), which all have plausible roles in the response to infection but there were no published SNP in exons of any of these and no SNP in conserved intergenic regions. *Thsd7b* (Thrombospondin type-1 domain-containing protein 7B Precursor) was the only gene in the region with a probably damaging (Polyphen) SNP and this SNP was also in an evolutionary conserved position. However there are no published studies of *Thsd7b* and expression levels are low in all tissues measured [36].

Tir3b. The *Tir3b* region was the largest QTL in the F6 (54.9 Mb) and F12 (13.4 Mb) and contains 650 and 113 genes respectively, of which 144 and 30 have haplotypes that correlate with phenotype. The F6 *Tir3b* QTL overlaps the *Tir3a* and *Tir3c* loci but exclusively contained *Ptprc* (protein tyrosine phosphatase, receptor type, C; Leukocyte common antigen Precursor, CD45 antigen), which had a probably damaging nsSNP in a highly conserved position (phastCons score 1). *Tir3b* F12 and F6 both contained *Soat1* (Sterol O-acyltransferase 1), which had a probably

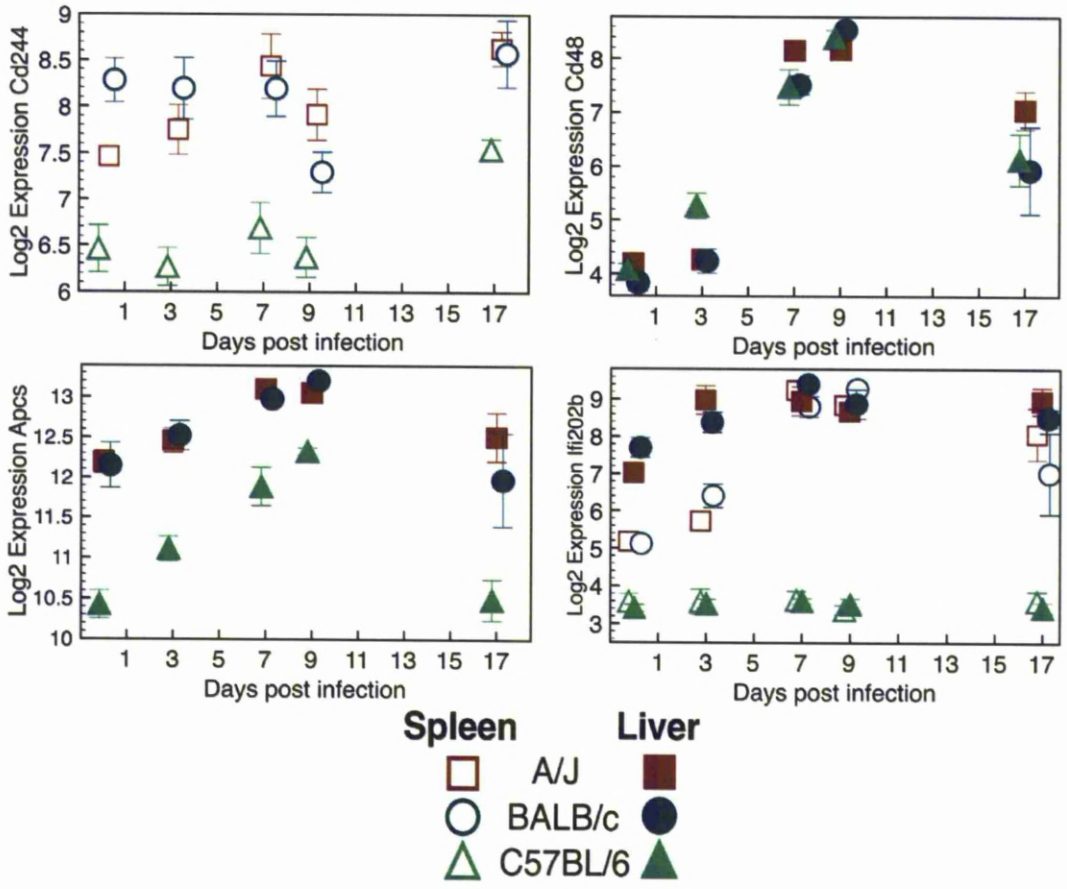


Figure 2. Expression of A/J OlaHsdnd (A/J), BALB/cJ OlaHsdce (BALB/c) and C57BL/6JOlaHSD (C57BL/6) mouse genes in the *Tir3c* locus at five time points in the course of infection (0 days; 3 days; 5days; 9 days; 17 days). Graphs include a small x-axis offset to improve spatial clarity. **A** *Cd244* in the spleen, **B** *Cd48* in the liver, **C** *Apcs* in the liver **D** *Ifi202b* in liver and spleen. *Cd244* expression was low in liver in all strains until Day 7 when it rose above background and C57BL/6 had slightly lower levels than A/J or BALB/c (data not shown).
doi:10.1371/journal.pntd.0000880.g002

damaging SNP. *Soat1* expression increases eight-fold after infection with *T. congolense* in A/J, BALB/c and C57BL/6 [31], and expression was up to four-fold higher in C57BL/6. *Soat1* is clearly responding to infection and the probably damaging SNP could affect its function and may be contributing to the difference in expression.

Tir3c. There were 122 and 8 genes at the F6 and F12 *Tir3c* loci that had haplotypes that correlate with phenotype, 54 and 8 of which had identical haplotypes in all the susceptible strains. *Cd244* (natural killer receptor 2B4) has a haplotype and expression pattern that correlates with phenotype (Figure 2A), as well as a CNV that may be the cause of the observed expression differences. It is a strong candidate for being a QTL gene at *Tir3c*. CD244 binds CD48 on lymphocytes and *Cd244* is about 60 Kbp from *Cd48*, which has a probably damaging nsSNP (rs31533394).

Additional candidate genes at *Tir3c* were *Apcs* (serum amyloid P-component; *Sap*) and *Ifi202b* (interferon activated gene 202B). The expression of *Apcs*, a major acute phase protein, rose after infection in all strains, but was consistently lower in C57BL/6 (Figure 2C). This was associated with a SNP (rs47990301) in a regulatory region that correlated with expression and phenotype and a SNP in a splice site in the 5′-UTR (rs47985673). Likewise, expression of *Ifi202b* increased to high levels after infection in A/J and BALB/c but remained at the threshold of detection in C57BL/6 in both liver and spleen. The *Ifi200* cluster, which includes *Ifi202b*, is at the distal end of *Tir3c* and contains genes that are all IFN-inducible and contain a highly conserved 200 amino acid motif [37]. *Fcgr3*, a low affinity immunoglobulin receptor that is associated with chronic inflammation [38], had a probably damaging nsSNP that correlated with phenotype. *Arhgap30* a little known rho-GTPase that is most highly expressed in macrophages and monocytes, had a probably damaging SNP (rs31539487) that correlated with phenotype in all strains tested. Similarly, we confirmed nsSNP in *Klhdc9* (rs45643169); *Darc* (Duffy blood group, chemokine receptor; rs51259593); *Slamf8* (signalling lymphocytic activation molecule F8; rs50073880) and *E430029J22Rik* (ENSMUSSNP3208701) that correlated with phenotype within this QTL region.

Sequence capture and sequencing of *Tir1*

DNA from across the *Tir1* QTL was sequenced in order to characterise novel SNP and to improve the identification of

alternate alleles for each haplotype block. DNA from four mouse breeds: 129P3, A/J, BALB/c and C3H/HeJ; was captured on Nimblegen arrays with probes for a 6.2 Mbp region of mouse chromosome 17 between 30,637,692 and 36,837,814 (NCBI37). 1.7 Mbp of repetitive sequence was excluded. Captured DNA was sequenced on a Roche 454 Genome Sequencer FLX using Titanium chemistry. 1,308,175 reads were mapped to the C57BL/6 reference sequence giving an average ~15× coverage of each sequenced strain (mean read length 282 bp; total sequence ~370 Mbp).

As 454 pyrosequencing is known to suffer from sequencing errors within, or close to long homopolymeric tracts, SNP were filtered to exclude those that were within a 13 bp window of homopolymeric tracts ≥5 bp. Furthermore, SNP were additionally filtered for those that were not outside regions covered by capture probes even if they were within the *Tir1* region. After filtering, 14,440 SNP loci were identified, 3,618 of which were not in dbSNP build 128. 1,588 loci were common to A/J, BALB/c and C3H/HeJ, but differed from C57BL/6. Furthermore, upon adding data for 129P3, there were 466 SNP loci common to all four sequenced mouse strains. Summary statistics for all SNP are available in Table S4 in Supporting Text S1.

Figure 3 shows a circular plot of all SNP called by the Roche/454 mapping algorithm (Newbler) against the C57BL/6 reference. Haplotype blocks can be seen as clusters of high-densities and low-densities of SNP. Whilst at this resolution it is not easy to see haplotype blocks in the A/J, BALB/c or C3H/HeJ data, one haplotype block stands out in the 129P3 data where 81 common SNP clustered within a 430 Kbp region (33,245,853–33,675,688 bp).

In order to validate SNP calls, 454-generated SNP were compared against those called in a recently published set sequenced on the Solexa/Illumina platform from flow-sorted mouse chromosome 17 for A/J [13], and similar, publicly available SNP from the concurrent Mouse Genomes Project (Wellcome Trust Sanger Institute) for BALB/c, C3H/HeJ and 129P2 mouse breeds [12]. Only 3 out of 36,784 (0.014%) of the homozygous calls (coverage >1; alternative allele frequency (AAF) >80%) were discordant between the two datasets Table S7 in Supporting Text S1. The 454 data included 53–71% of SNP in the Illumina data depending on the coverage required to call a SNP and the Illumina data contained 94–97% of SNP in the 454 data (Figure S5 in Supporting Text S1). Full details of the comparison are available in Supplementary Data S3; SNP validation.xls.

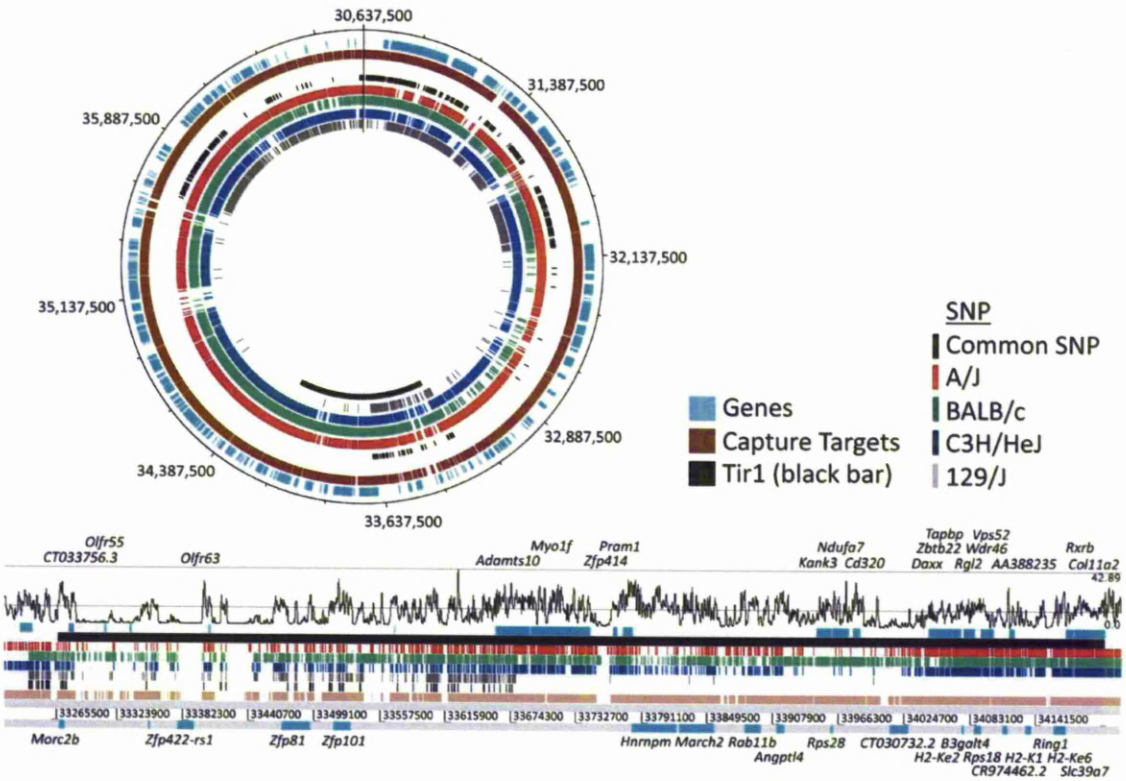


Figure 3. Array-based sequence capture and next generation sequencing of a 6.2 Mbp region of Mmu17 in four breeds of mice: A/J; BALB/c; C3H/HeJ and 129/J (Mmu17:30,637,692–36,837,814 bp). Plot is circular for ease of display [56]. *Tir1* is highlighted in black on the inside track. Genomic positions are in Mbp. The outer tracks (blue and brown) show genes and designed capture probes, respectively. The four, coloured, inner tracks show SNP called in each of the four sequencing experiments, with the black tick marks highlighting areas of common SNP. Haplotype blocks can clearly be seen as clustering of high- and low- density regions of SNP. A magnified region around *Tir1* is displayed underneath the circular plot. Tracks are identically coloured and include a moving average (window 1 Kbp) of sequence read coverage across the region (top). Genes in the region are displayed for the forward strand (above) and reverse strand (below).
doi:10.1371/journal.pntd.0000880.g003

Structural polymorphisms

Using all available data, the *Tir1* region contained 80 nsSNP loci that correlated with phenotype. There were seven "possibly damaging" (Polyphen) nsSNP and "probably damaging" nsSNP in PML-retinoic acid receptor alpha regulated adaptor molecule 1 (*Pram1*) (rs33399614), *Rgl2* (Ral guanine nucleotide dissociation stimulator-like 2) and *CR974462* (Table 3). Nine genes contained splice site polymorphisms (See Supplementary Data S1: *AnnotatedFunctionalSNP.xls*).

Regulatory polymorphisms

Differences between the susceptible strains and C57BL/6 were aligned to the Ensembl mouse regulatory build (NCBI37: Ensembl 54). Ten differences were predicted to fall within regions of accessible chromatin and may affect transcription factor binding regions. Furthermore, 13 differences mapped to within 2500 bp of the upstream region of genes that may be associated with promoter regions. In total, 14 genes may be affected by SNP in this way (Table S5 in Supporting Text S1). Of the 13 genes for which microarray data was available, however, only phosphodiesterase 9A (*Pde9a*) showed any differences in gene expression, and these correlated with alleles of a SNP (rs33223038). A/J differed from C57BL/6 and BALB/c at this locus in both SNP genotype and *Pde9a* expression, but since this did not correlate with phenotype, it was discounted as a candidate SNP. There were also SNP in non-essential splice sites in nine genes that may modify their exon usage (Supplementary Data S1: *AnnotatedFunctionalSNP.xls*).

Correlation of haplotype assignments using 454 and Perlegen data

Jukes Cantor distances were calculated for each haplotype block in the *Tir1* region using the 454 and Perlegen datasets. A more detailed description of the analysis is presented in Supporting Text S1. Shared haplotypes had high positive predictive value and specificity for shared SNP alleles but low negative predictive value and sensitivity (Table S3b in Supporting Text S1), indicating that having shared haplotypes is a good indicator of shared SNP alleles but that the converse is not true. This means that assignments will be accurate where C57BL/6 has been assigned the same haplotype allele as susceptible strains but less accurate where C57BL/6 has been assigned to a different haplotype block allele from the susceptible strains. Therefore the data may be reliable way of excluding loci as candidate QTL regions but less accurate for including loci. The correlation between the distances calculated from the 454 and Perlegen SNP sets was modest ($r = 0.63$). The slope of the regression line was 0.67 reflecting the greater number of SNP in the 454 dataset. A high degree of scatter was observed in a plot of distances based on Perlegen and 454 data (Figure S3 in Supporting Text S1). The scatter suggests that SNP coverage is uneven in one or both datasets, and therefore increasing SNP density should increase the reliability of haplotype calls. Inspection of a plot of SNP coverage in the two data sets shows that the ratio of the number of SNP that were found in the two data sets varied substantially between haplotype blocks (Figure 4 and Figure S4 in Supporting Text S1).

Plots like those shown in Figures 4 and S3 can be obtained for any region of *Tir1* from our website [31]. Plots of SNP and haplotypes and tables of Jukes Cantor distances between alleles at each haplotype block based on Perlegen data can be obtained for any part of the mouse genome at the same site.

Discussion

The survival time phenotype for mapping murine QTL associated with response to *T. congolense* infection was selected in the 1990's because the large variance between strains made it more likely that there would be QTL of large enough effect to be identifiable. This prediction proved correct [6], however survival is likely to have a remote and complex relationship with the underlying quantitative trait genes (QTG). Given that trypanosomiasis is a systemic blood stream infection and the remote relationship between survival and the underlying QTG it is almost impossible to prioritise candidate genes on the basis of known functions. We have previously measured parasitaemia, anaemia and fifteen clinical chemistry phenotypes, in inbred and congenic mice, in order to identify correlations between survival and other traits that might be more proximally related to gene function, however no such associations have been found [32]. Therefore in this study we have identified the allele carried at each QTL in an additional strain (C3H/HeJ), formally identified the physical boundaries of the QTL and enumerated CNV and functional SNP that fall within those boundaries.

QTL mapping

The mapping studies showed that C3H/HeJ mice carry susceptible alleles at the *Tir1* and *Tir3* loci. No QTL were

Table 3. nsSNP loci within the extended *Tir1* definition.

Position	C57BL/6	A/J	BALB/c	C3H/HeJ	Phast Cons	Gene	Polyphen Consequence	Peptide shift
33,283,941	A		G	G	<0.1	*Zfp421*	possibly damaging	Y/C
33,781,645	T	C	C	C	<0.1	*Pram1*	probably damaging	L/P
33,956,791	T		C		<0.1	*Kank3*	possibly damaging	S/P
34,069,285	C	T	T	T	0.928	*Rgl2*	probably damaging	H/Y
34,112,420	T	C	C	C	<0.1	CR974462.5	probably damaging	H/R
34,114,833	C	-	-		<0.1	CR974462.5	possibly damaging	G/R
34,119,278	G	A			<0.1	AA388235	possibly damaging	R/H
34,119,383	G	A			<0.1	AA388235	possibly damaging	G/D
34,119,473	T	C			0.337	AA388235	possibly damaging	F/S
34,134,481	T	C		C	<0.1	H2-K1	possibly damaging	H/R

Genes within *Tir1* (Mmu17:33271855–34203529 bp) with damaging nsSNP that correlate with survival phenotype. A full list of annotated SNP is available in Supplementary Data S1: *AnnotatedFunctionalSNP.xls*.

doi:10.1371/journal.pntd.0000880.t003

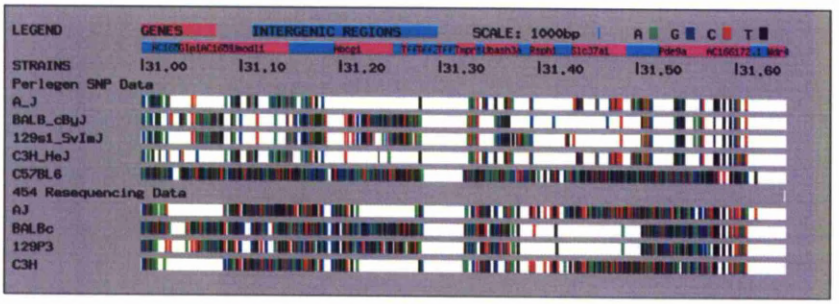


Figure 4. SNP plots of *Tir1* between 31 and 31.65 Mbp. The C57BL/6 row represents the reference allele for all loci that are polymorphic in either the Perlegen set or our 454 set. The SNP density is much greater in the 454 data set, in which haplotype blocks are clearly identifiable by eye. It appears that SNP are much better represented in the Perlegen data set in some regions than in others. Between 31.2 and 31.30 the two data sets are very similar with high density of SNP in BALB/c and 129 substrains in each dataset. However in the region between 31.1 and 31.20 the SNP in BALB/c and 129 are relatively much sparser in Perlegen than in the 454 data.
doi:10.1371/journal.pntd.0000880.g004

observed at the *Tir2* locus. The *Tir1* locus as defined by previous fine mapping studies is just proximal to the major histocompatibility complex (MHC) (Table 2), and the conversion of genetic distances to physical positions presented here shows that *Tir1* includes three classical class I MHC H2K genes. However previous studies have found no correlation between MHC haplotype and response to infection [4] consistent with the QTL gene not being a classical MHC molecule.

The mapping population was also screened for an association between *Tlr4* and survival; no association was found. This observation implies that the presence or absence of a functional *Tlr4* gene has no effect on survival, but does not preclude the pathway from *Tlr4* to *Nfkb* (nuclear factor kappa-B) from responding to infection. *Tlr4* could still participate in the regulation of anaemia and parasitaemia, which are not correlated with survival [11].

Identification of physical boundaries of QTL

The two independent F6 and F12 mapping populations have reduced the 95% CI of the QTL to exceptionally small regions, particularly at *Tir1*, the QTL of largest effect, where the physical size of the 95% CI was 930 Kbp for the combined data from the A/J × C57BL/6 and BALB/c × C57BL/6 F6 crosses (Table 1). This was only twice the mean distance between markers at this locus (400 Kbp) and consequently the main limitation in identifying the boundaries of the QTL is in estimating the position of the peak.

Identification of functional nsSNP

Resequencing of the QTL region on the Roche 454 platform at Liverpool to 15× coverage discovered 3,618 novel SNP loci that were deposited in dbSNP. Comparison with a resequencing project on the Illumina platform at the Wellcome Trust Sanger Institute to 22× coverage [13] showed 99.98% consistency in SNP calls even when no minimum coverage criterion was applied for calling a SNP. Both data sets contained large numbers of SNP called as heterozygotes with alternative allele frequencies between 25–80%. These loci from both data sets were associated with significantly higher sequence coverage in our data indicating that the majority were likely to be due to mapping artefacts probably caused by CNV. The 454 data contained only 71% of the SNP discovered by the higher coverage Illumina data but both methods discovered the same set of nsSNP. The 454 data discovered an additional 3% of SNP that were not in the Illumina data.

Utilising all SNP from the 454, Perlegen and Illumina data sets, three probably damaging nsSNP were identified in genes at the peak of the *Tir1* QTL that correlated perfectly with phenotype (Table 3). Two nsSNP were in *Pram1*; the *Pram1*$^{537L/P}$ polymorphism was scored as probably damaging by Polyphen. The *Pram1*$^{103R/K}$ polymorphism was classed as benign by Polyphen but lies within a proline rich domain (PRINTS: PR01217) that is involved in binding the "SH3 domain of hematopoietic progenitor kinase 1 (HPK-1)-interacting protein of 55 kDa (HIP-55)," which is known to stimulate the activity of HPK-1 and c-Jun N-terminal kinase (JNK)" [39]. *Pram1* is almost exclusively expressed in myeloid cells [36] and specifically in granulocytes in terminal stages of differentiation [40] where it is induced by retinoic acid. It was thought that *Pram1* might be a negative regulator of neutrophil differentiation since it is repressed in acute myeloid leukaemia. The deletion of *Pram1*, however, has no effect on neutrophil differentiation and maturation but does disrupt reactive oxygen intermediate production and degranulation by neutrophils [41]. This may affect the early, pro-inflammatory response to infection or downstream TNFa signalling, which has been shown to be differentially expressed in susceptible and resistant mice [42]. C57BL/6 appears to have the derived allele of *Pram1*$^{537L/P}$ since A/J, BALB/c and C3H/HeJ had the same allele as Hominidae and dogs. Since C57BL/6 tend to have a more inflammatory phenotype, it is possible that the polymorphisms lead to a gain of function with stronger binding to HIP55 leading to faster and more persistent ROI induction and a more inflammatory state.

The other probably damaging SNP at *Tir1* were *CR974462* and *Rgl2*. There is no annotation for CR974462. *Rgl2 (Rif)* is a small GTPase that is most highly expressed in macrophages and B cells and appears to be involved in *Ras* mediated signalling [43]. The *Rgl2*$^{147H\rightarrow Y}$ polymorphism could affect the *Ras* pathway that plays a key role in leukocyte activation and is therefore a plausible candidate gene.

The Fas death domain-associated protein (*Daxx*) gene, which we have previously reported to contain a deletion of a single aspartate residue in susceptible mice [44], is also under the peak of *Tir1*. *Daxx* is within the MAPK pathway, which was found to respond to *T. congolense* infection in microarray data. However a new Polyphen analysis of the aspartate deletion in *Daxx* indicates that this polymorphism will be benign in effect. The aspartate deletion is within a run of 11 aspartate residues and a region where 35/41 residues are acidic [44]. Therefore this polymorphism is probably less significant than the probably damaging ones reported here.

Regulatory polymorphisms could also cause the phenotypic difference: one SNP (rs33223038) was identified in Ensembl as being in a regulatory region upstream of *Pde9a* but although this SNP correlated with differential expression it did not correlate with survival differences between susceptible and resistant mouse breeds. There were also SNP in non-essential splice sites in nine genes.

Copy number variation at *Tir* QTL

CNV have previously been shown to be a major cause of quantative trait differences [10]. We used Agilent whole mouse genome aCGH arrays to identify CNV between C57BL/6 mice and A/J, BALB/c and 129P3 mice. The aCGH data highlighted a CNVR containing three genes close to the peak of the *Tir3c* QTL: *Cd244*; *Ly9* and *Ithl1* (Figure 1A). A nearby gene *Cd48* had a probably damaging nsSNP. *Cd244*, *Cd48* and *Ly9* are important genes involved in the production and regulation of IFNg by NK and T cells. CD244 binds CD48 on lymphocytes and is involved in NK:NK cell and NK:T cell interactions leading to NK and T cell proliferation [45], which are important mechanisms in innate resistance to protozoan infection [46,47].

Splenic expression of *Cd244* differed between strains with the resistant C57BL/6 mice having the lowest expression consistent with the low copy number of *Cd244* in C57BL/6. *Cd48* expression increased 16-fold in liver after infection with *T. congolense*, but this occurred in all strains tested (Figure 2). Since CD48 and CD244 directly interact, it is possible that the QTL is a consequence of the combined effect of the probably damaging nsSNP in *Cd48* and the CNV in *Cd244*. Differences in expression could not be seen in *Ithb* or *Ly9*.

The large number of genes in *Tir3c* that had CNV, nsSNP or haplotypes that correlated with phenotype may make it difficult to identify the QTL gene at this locus. It is possible that the QTL is not a consequence of a single polymorphism but the combined effect of multiple polymorphisms in an extended haplotype. However the CNV at *Cd244* was the most substantial DNA polymorphism in the region making *Cd244* a strong candidate QTL gene. Inserting an additional copy of Cd244 into the C57BL/6 background, so that it had a similar gene dosage to the susceptible strains, could test the effect of this CNV on the response to infection.

Haplotype block analysis

We have previously used this strategy to show a strong association between upstream haplotype differences and high confidence ($p<0.005$) differences in gene expression [19] and also short listed genes under QTL for differences in response to *Heligmosomoides bakeri* infection [20]. We reduced the number of candidate genes in this study by about 76% and 45% under hypotheses H1 and H2 from the 1193 genes that were under the 95% confidence intervals of the QTL. There were 283 genes where A/J, BALB/c and C3H/HeJ had the same haplotype different from C57BL/6 and 651 genes where C57BL/6 differed from the other three. The large number of genes that had haplotypes that correlated with phenotype is mainly because: 1) C3H/HeJ, A/J and BALB/c are more similar to each other than to any other strain based on analysis of 673 genome wide SNP in 55 strains [48]; 2) we used the stringent criterion that a gene was included if any haplotype block between the two neighbouring genes correlated with phenotype; 3) The high positive predictive power of the method means that whilst it is probably very reliable for excluding loci where susceptible strains share a haplotype block with the C57BL/6 resistant strains, it assigns too many haplotype blocks to different alleles.

Whilst there are large numbers of reported SNP for A/J, 129X1/SvJ and 129S1/SvImJ due to the Celera sequencing project [49] and for BALBc/ByJ and C3H/HeJ from the Perlegen project [15], relatively few SNP are publicly available for the 129P3 strain. The 454 resequencing of the *Tir1* region indicated that approximately 50% of the resequenced region could be excluded from the QTL if the allele carried by 129P3 mice at this locus was known. If a QTL was identified at *Tir1* in a 129P3 × C57BL/6 cross then the QTL gene could be assumed to be within the three blocks where 129P3 differed from C57BL/6. If no evidence of a QTL was found then these regions could be excluded from the QTL on the assumption that 129P3 carried the same allele as C57BL/6 at this locus. This analysis indicates that mapping QTL for response to infection in a 129P3 × C57BL/6 cross should significantly refine the list of candidate genes. The availability of this haplotype data makes it possible to make more rational choices about the selection of strains for mapping experiments. This strategy has been used before with a much more limited SNP set [50] but the corresponding online resource is no longer available.

Our objective was to identify the SNP that were most likely to have an impact on function. These were considered to be nsSNP that altered the physical properties of the protein as judged by Polyphen analysis, SNP in essential splice sites and CNV and regulatory SNP that correlated with changes in expression. It should be emphasised, however, that many types of SNP can underlie QTL, for example the QTL SNP at the *Idd5* locus appears to be a synonymous SNP that gives rise to a splice variant [51]. This SNP would not have been identified as a high priority by our pipeline. Furthermore, although we have substantially complete sequence coverage of the *Tir1* locus, at other loci we have used the Perlegen data, which is estimated to be about 45% complete [15]. Therefore although the candidate QTL SNP presented here are the most likely given the available data and annotation, both SNP data and annotation is incomplete and other candidates may be discovered in the future.

The correlation of Jukes-Cantor distances calculated from our 454 data and the published Perlegen dataset was only modest ($r=0.63$). 32% of our 454 SNP loci were also in the Perlegen set, however the low correlation between the two sets shows that SNP discovery was uneven in one or both sets and inspection of the SNP distribution suggests that this was certainly the case in the Perlegen set. The uneven distribution of SNP discovery makes it much harder to undertake a consistent analysis across the genome using a single threshold for assigning alleles to haplotype blocks. However the high positive predictive value (Table S3 in Supporting Text S1) for identifying shared haplotypes suggests that this procedure should reliably exclude regions where C57BL/6 shares haplotypes with the susceptible strains. Nevertheless other more robust data types such as CNV and potentially functional SNP should still be surveyed in regions where haplotype does not correlate with phenotype. The more complete mouse resequencing projects that are currently underway should increase the predictive power of this approach substantially.

QTL involved with resistance to other parasitic diseases overlap with the *Tir* QTL, raising the possibility that polymorphisms discovered here may be involved in the response to other parasites. *Leishmania* resistance 1 (*Lmr1*) [52], *Plasmodium chabaudi* resistance QTL 3 (*Char3*) [53] and *Heligmosomoides bakeri* nematode resistance 2 (*Hbnr2*) [54] all overlap with *Tir1*. Similarly, the *Tir3c* QTL overlaps with a QTL for murine resistance to *Plasmodium berghei*-driven experimental cerebral malaria (*Berr1*) [55].

Thirteen genes around the peak of *Tir1* show conserved order and sequence homology to a ~311 Kbp region of BTA7

(15,412,179:15,723,462 bp) where there is a QTL in cattle that regulates the level of parasitaemia in cattle infections with *T. congolense* [2]. This region includes *Pram1*, which has a probably damaging mutation that correlates with phenotype in mice and was the most plausible candidate gene in *Tir1* and is therefore a candidate QTL gene in cattle as well. However since trypanotolerance QTL cover approximately 15% of the bovine genome it would be expected that at least one of the five murine QTL would coincide with a bovine QTL by chance ($p = 0.56$).

Conclusions

By linking genes to haplotypes, we have reduced the number of candidate genes in *Tir1* to 43. Within these there were three genes with probably damaging nsSNP; CR974462.5, *Rgl2* and *Pram1*. CR974462.5 is an anonymous gene in which the effects are hard to predict. *Pram1* regulates oxidative stress in neutrophils and *Rgl2* is involved in *Ras* signalling which can regulate inflammation. *Pram1* is the closest to the peak of the QTL and has the best understood function making it the most attractive candidate at *Tir1* however *Rgl2* is also a plausible candidate. Probably damaging polymorphisms were identified in *Srp72* in *Tir2* and *Thsd7b* in *Tir3a* but little is known of their functions so it is hard to interpret these observations. *Ptprc (Cd45)* and *Soat1* in *Tir3b* had probably damaging polymorphisms in conserved nucleotides, CD45 is the common leukocyte antigen and has multiple roles in cytokine signaling and cell regulation making it plausible candidate. *Tir3c* has a CNVR encompassing *Cd244*, which is differentially expressed and has a haplotype that correlates with phenotype in the four strains tested. Since gene dosage is lower in C57BL/6 it will be possible to test this hypothesis by inserting an extra copy of the *Cd244* gene into the C57BL/6 background. Several other genes in *Tir3c* had haplotype and nsSNP that correlated with phenotype but none had such a distinct CNV and such strong differential expression.

By combining haplotype analysis, array-CGH, gene expression and next-generation DNA capture and sequencing, we have identified a small number of genetic polymorphisms that may be responsible for differences in response to *T. congolense* infection, demonstrating that this approach can systematically reduce the number of candidate genes under QTL to generate a short enough list of genes to test for function.

Supporting Information

Data S1 AnnotatedFunctionalSNP.xls. A comprehensive annotation of publicly available SNP (NCBI build 37) across QTL regions including Polyphen annotation.
Found at: doi:10.1371/journal.pntd.0000880.s001 (1.97 MB XLS)

Data S2 GenesAndHaplotypes.xls. Haplotype block alleles across QTL regions
Found at: doi:10.1371/journal.pntd.0000880.s002 (1.02 MB XLS)

Data S3 SNP validation.xls. Comparison of resequencing of Tir1 region on the Illumina system at the Wellcome Trust Sanger Institute and on the 454 system at the University of Liverpool.
Found at: doi:10.1371/journal.pntd.0000880.s003 (0.04 MB XLS)

Text S1 Supporting Text referred to in the main text. Includes: additional methods on haplotype analysis and genotyping markers and primers; and additional SNP and CNV data.
Found at: doi:10.1371/journal.pntd.0000880.s004 (1.30 MB DOC)

Acknowledgments

We thank Leanne Wardlesworth of the University of Manchester Core Services unit for excellent technical assistance.

Author Contributions

Conceived and designed the experiments: IG AA AB JG NH ML FI SJK HAN. Performed the experiments: IG PA MAH ML FI. Analyzed the data: IG AB JG HAN. Contributed reagents/materials/analysis tools: AA NH HAN. Wrote the paper: IG HAN.

References

1. Kristjanson P, Swallow B, Rowlands G, Kruska R, de Leeuw P (1999) Measuring the costs of African animal trypanosomosis, the potential benefits of control and returns to research. Agricultural Systems. pp 1–20.
2. Hanotte O, Ronin Y, Agaba M, Nilsson P, Gelhaus A, et al. (2003) Mapping of quantitative trait loci controlling trypanotolerance in a cross of tolerant West African N'Dama and susceptible East African Boran cattle. Proc Natl Acad Sci U S A 100: 7443–7448.
3. Morrison W, Murray M (1979) *Trypanosoma congolense*: inheritance of susceptibility to infection in inbred strains of mice. Experimental Parasitology 48: 364–374.
4. Morrison W, Roelants G, Mayor-Withey K, Murray M (1978) Susceptibility of inbred strains of mice to *Trypanosoma congolense*: correlation with changes in spleen lymphocyte populations. Clinical and Experimental Immunology 32: 25–40.
5. Murray M, Morrison WI, Whitelaw DD (1982) Host susceptibility to African trypanosomiasis: trypanotolerance. Adv Parasitol 21: 1–68.
6. Kemp SJ, Iraqi F, Darvasi A, Soller M, Teale AJ (1997) Localization of genes controlling resistance to trypanosomiasis in mice. Nature Genetics 16: 194–196.
7. Iraqi F, Clapcott S, Kumari P, Haley C, Kemp S, et al. (2000) Fine mapping of trypanosomiasis resistance loci in murine advanced intercross lines. Mammalian Genome 11: 645–648.
8. Nganga JK, Soller M, Iraqi FA (2010) High resolution mapping of trypanosomosis resistance loci *Tir2* and *Tir3* using F12 advanced intercross lines with major locus *Tir1* fixed for the susceptible allele. BMC Genomics 11: 394.
9. Flint J, Valdar W, Shifman S, Mott R (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nature Reviews Genetics 6: 271–286.
10. Cho EK, Tchinda J, Freeman JL, Chung YJ, Cai WW, et al. (2006) Array-based comparative genomic hybridization and copy number variation in cancer research. Cytogenet Genome Res 115: 262–272.
11. Noyes HA, Alimohammadian MH, Agaba M, Brass A, Fuchs H, et al. (2009) Mechanisms controlling anaemia in *Trypanosoma congolense* infected mice. PLoS One 4: e5170.
12. Mouse Genomes Project (Wellcome Trust Sanger Institute). http://www.sanger.ac.uk/resources/mouse/genomes/.
13. Sudbery I, Stalker J, Simpson JT, Keane T, Rust AG, et al. (2009) Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. Genome Biol 10: R112.
14. Graubert T, Cahan P, Edwin D, Selzer R, Richmond T, et al. (2007) A High-Resolution Map of Segmental DNA Copy Number Variation in the Mouse Genome. PLoS Genet 3: e3.
15. Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, et al. (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448: 1050–1053.
16. Darvasi A, Soller M (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. TAG Theoretical and Applied Genetics.
17. Poltorak A, He X, Smirnova I, Liu MY, Van Huffel C, et al. (1998) Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. Science 282: 2085–2088.
18. Bonfield JK, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23: 4992–4999.
19. Noyes HA, Agaba M, Anderson S, Archibald AL, Brass A, et al. (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.
20. Behnke JM, Menge DM, Nagda S, Noyes H, Iraqi FA, et al. (2010) Quantitative trait loci for resistance to *Heligmosomoides bakeri* and associated immunological and pathological traits in mice: comparison of loci on chromosomes 5, 8 and 11 in F2 and F6/7 inter-cross lines of mice. Parasitology 137: 311–320.
21. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package). *University of Washington, Seattle. Department of Genome Sciences.*
22. Rennie C, Noyes H, Kemp S, Hulme H, Brass A, et al. (2008) Strong position-dependent effects of sequence mismatches on signal ratios measured using long oligonucleotide microarrays. BMC Genomics 9: 317.

23. Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z (2006) Efficient calculation of interval scores for DNA copy number data analysis. J Comput Biol 13: 215–228.
24. Albert T, Molla M, Muzny D, Nazareth L, Wheeler D, et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Meth 4: 903–905.
25. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876.
26. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, et al. (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16: 123–131.
27. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
29. UCSC phastCons Conservation Scores. http://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/.
30. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894–3900.
31. University of Liverpool African Bovine Trypanosomiasis Website. http://www.genomics.liv.ac.uk/tryps/resources.html.
32. Rathkolb B, Noyes HA, Brass A, Dark P, Fuchs H, et al. (2009) Clinical chemistry of congenic mice with quantitative trait loci for predicted responses to *Trypanosoma congolense* infection. Infection and Immunity 77: 3948–3957.
33. Oliveira A-C, de Alencar BC, Tzelepis F, Klezewsky W, da Silva RN, et al. (2010) Impaired innate immunity in *Tlr4*(−/−) mice but preserved CD8+ T cell responses against *Trypanosoma cruzi* in *Tlr4*-, *Tlr2*-, *Tlr9*- or *Myd88*-deficient mice. PLoS Pathog 6: e1000870.
34. Alafiatayo RA, Crawley B, Oppenheim BA, Pentreath VW (1993) Endotoxins and the pathogenesis of *Trypanosoma brucei brucei* infection in mice. Parasitology 107 (Pt 1): 49–53.
35. Hovatta I, Tennant RS, Helton R, Marr RA, Singer O, et al. (2005) Glyoxalase 1 and glutathione reductase 1 regulate anxiety in mice. Nature 438: 662–666.
36. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences 101: 6062–6067.
37. Landolfo S, Gariglio M, Gribaudo G, Lembo D (1998) The Ifi 200 genes: an emerging family of IFN-inducible genes. Biochimie 80: 721–728.
38. Nigrovic PA, Binstadt BA, Monach PA, Johnsen A, Gurish M, et al. (2007) Mast cells contribute to initiation of autoantibody-mediated arthritis via IL-1. Proc Natl Acad Sci U S A 104: 2325–2330.
39. Denis FM, Benecke A, Di Gioia Y, Touw IP, Cayre YE, et al. (2005) PRAM-1 potentiates arsenic trioxide-induced JNK activation. J Biol Chem 280: 9043–9048.
40. Moog-Lutz C, Peterson EJ, Lutz PG, Eliason S, Cave-Riant F, et al. (2001) PRAM-1 is a novel adaptor protein regulated by retinoic acid (RA) and promyelocytic leukemia (PML)-RA receptor alpha in acute promyelocytic leukemia cells. J Biol Chem 276: 22375–22381.
41. Clemens RA, Newbrough SA, Chung EY, Gheith S, Singer AL, et al. (2004) PRAM-1 is required for optimal integrin-dependent neutrophil function. Mol Cell Biol 24: 10923–10932.
42. Uzonna JE, Kaushik RS, Gordon JR, Tabel H (1998) Experimental murine *Trypanosoma congolense* infections. I. Administration of anti-IFN-gamma antibodies alters trypanosome-susceptible mice to a resistant-like phenotype. J Immunol 161: 5507–5515.
43. Post GR, Swiderski C, Waldrop BA, Salty L, Glembotski CC, et al. (2002) Guanine nucleotide exchange factor-like factor (*Rlf*) induces gene expression and potentiates alpha 1-adrenergic receptor-induced transcriptional responses in neonatal rat ventricular myocytes. Journal of Biological Chemistry 277: 15286–15292.
44. Fisher P, Hedeler C, Wolstencroft K, Hulme H, Noyes H, et al. (2007) A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis. Nucleic Acids Research 35: 5625–5633.
45. Assarsson E, Kambayashi T, Schatzle JD, Cramer SO, von Bonin A, et al. (2004) NK cells stimulate proliferation of T and NK cells through 2B4/CD48 interactions. J Immunol 173: 174–180.
46. Scharton-Kersten TM, Sher A (1997) Role of natural killer cells in innate resistance to protozoan infections. Curr Opin Immunol 9: 44–51.
47. Harty JT, Tvinnereim AR, White DW (2000) CD8+ T cell effector mechanisms in resistance to infection. Annu Rev Immunol 18: 275–308.
48. Tsang S, Sun Z, Luke B, Stewart C, Lum N, et al. (2005) A comprehensive SNP-based genetic analysis of inbred mouse strains. Mamm Genome 16: 476–480.
49. Reuveni E, Ramensky VE, Gross C (2007) Mouse SNP Miner: an annotated database of mouse functional single nucleotide polymorphisms. BMC Genomics 8: 24.
50. Cervino AC, Gosink M, Fallahi M, Pascal B, Mader C, et al. (2006) A comprehensive mouse IBD database for the efficient localization of quantitative trait loci. Mamm Genome 17: 565–574.
51. Araki M, Chung D, Liu S, Rainbow DB, Chamberlain G, et al. (2009) Genetic evidence that the differential expression of the ligand-independent isoform of CTLA-4 is the molecular basis of the Idd5.1 type 1 diabetes region in nonobese diabetic mice. The Journal of Immunology 183: 5146–5157.
52. Roberts LJ, Baldwin TM, Curtis JM, Handman E, Foote SJ (1997) Resistance to *Leishmania major* is linked to the H2 region on chromosome 17 and to chromosome 9. J Exp Med 185: 1705–1710.
53. Burt RA, Baldwin TM, Marshall VM, Foote SJ (1999) Temporal expression of an H2-linked locus in host response to mouse malaria. Immunogenetics 50: 278–285.
54. Iraqi FA, Behnke JM, Menge DM, Lowe AM, Teale AJ, et al. (2003) Chromosomal regions controlling resistance to gastro-intestinal nematode infections in mice. Mamm Genome 14: 184–191.
55. Bagot S, Campino S, Penha-Goncalves C, Pied S, Cazenave PA, et al. (2002) Identification of two cerebral malaria resistance loci using an inbred wild-derived mouse strain. Proc Natl Acad Sci U S A 99: 9919–9923.
56. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

Appendix IX: Example Perl Scripts (attached CD)

Additional data file 1: Removal of SNP within 13bp of a homopolymeric tract ≥5bp.

Additional data file 2: 454 non-synonymous SNP identification.

Additional data file 3: SOLID SNP extraction (BIOSCOPE).

Additional data file 4: KASPAR genotyping data extraction script.

Additional data file 5: Artificial splitting of Sanger (FASTA) reads into pseudo next-generation sequencing reads for subsequent alignment using BOWTIE (50bp).