Bridging between parasite genomic data and population processes: Trypanosome dynamics and the antigenic archive

UNIVERSITY
of
GLASGOW

Erida Gjini

School of Mathematics and Statistics

University of Glasgow

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

May 2012

To my parents, Eduart and Fatbardha

Acknowledgements

Firstly, a huge thanks to my main supervisor Christina Cobbold, for her continuous support, un-ending encouragement, her patience in guiding me, and help in finding my way through the PhD. I learned a lot of mathematical modeling in these years and so much of it I owe to Christina, her constructive criticism, her ability to explain complex issues in a simple way and not let go of details. I am very lucky to have had two other great supervisors, Dan Haydon and Dave Barry, who brought me closer to the fascinating world of evolutionary biology and genetics, inspired and taught me to become an interdisciplinary scientist, and provided endless confidence boosts during difficult times. Christina, Dan, Dave, thank you for everything, including the generous financial support to attend various scientific conferences. I am very grateful to have been a PhD student of yours.

I would like to thank also my external examiner, Nick Savill from the University of Edinburgh, and my internal examiner Xiaoyu Luo from Glasgow University, for their insightful comments, suggestions and constructive criticism which helped to improve this thesis after my viva.

Many thanks go to Lucio Marcello, Liam Morrison, Jamie Hall and Lindsey Plenderleith at WTCMP, with whom I had interesting and stimulating discussions on trypanosomes at various stages of my PhD. I want to thank Catherine Higham and Mathew Denwood who introduced and explained Bayesian estimation to me in a very understandable way and enabled me to use it in practice. I am grateful to Tom Kwiatkowski in Edinburgh for mentioning Hidden Markov Models at a crucial moment during my PhD,

thus opening a whole new modelling avenue for me. Thanks to the references provided by Christian Althaus, I became more interested on body size scaling relationships and their application to infectious diseases.

I want to thank my colleagues in room 522 at the Maths Department, Dot, Tarig, Craig, Bhishan, David, Faiza, Moniba and Bart who provided a very warm office environment. Then Otti, whom I met regularly on the Edinburgh-Glasgow train during these years of commuting, who provided great company and advice, thank you. Thanks also to various friends in Edinburgh, Stephanie, Rodrigo, Marta, Daniele Sepe, Daniele Fanelli, Patricia, Antonella, Thoma, Sayak, the Italians, the Greeks, etc. who made it possible for me to have life and excitement outside the PhD.

I also want to thank my family. My parents for their love, care and faith in me and their contribution in my way of thinking about science. My sisters for their emotional support and fun times together. My grandmother for being the first mathematician of my life and a true source of inspiration.

Finally I want to thank Moreno, who was there from the beginning of this journey, who was actually the reason why I looked into Scotland for a PhD in the first place, who helped me in all ways to enjoy this experience, filling my life with love and warmth, sharing with me all ups and downs, and making me look always forward.

Author's Declaration

I declare that this thesis is a record of the original work carried out solely by myself in the School of Mathematics and Statistics at the University of Glasgow, United Kingdom, during the period of October 2007 to September 2011. The copyright of this thesis belongs to the author under the terms of the United Kingdom copyright acts. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis. This thesis has not been presented elsewhere in consideration for a higher degree.

The results of Chapter 2 have appeared in the following publication:

Gjini, E., Haydon, D.T., Barry, J.D., Cobbold, C.A. (2010). Critical interplay between parasite differentiation, host immunity, and antigenic variation in trypanosome infections. *The American Naturalist*, **176**(4): 424-39.

The results of Chapter 4 are presented in the following publication:

Gjini, E., Haydon, D.T., Barry, J.D., Cobbold, C.A. (2012). The impact of mutation and gene conversion on the local diversification of antigen genes in African trypanosomes. *Journal of Molecular Biology and Evolution*, (Accepted on May 4, 2012).

Erida Gjini

2012

Abstract

Antigenic variation processes play a central role in parasite invasion and chronic infectious disease, and are likely to respond to host immune mechanisms and epidemiological characteristics. Whether changes in antigenic variation strategies lead to net positive or negative effects for parasite fitness is unclear. To improve our understanding of pathogen evolution, it is important to investigate the mechanisms by which pathogens regulate antigenic variant expression. This involves consideration of the complex interactions that occur between parasites and their hosts, and top-down and bottom-up factors that might drive changes in the genetic architecture of their antigenic archives.

Increasing availability of pathogen genomic data offers new opportunities to understand the fundamental mechanisms of immune evasion and pathogen population dynamics during chronic infection. Motivated by the growing knowledge on the antigenic variation system of the sleeping sickness parasite, the African trypanosome, in this thesis, we present different models that analyze antigenic variation of this parasite at different biological scales, ranging from the within-host level, to between-host transmission, and finally the parasite genetics level.

First, we describe mechanistically how the structure of the antigenic archive impacts the parasite population dynamics within a single host, and how it interplays with other within-host processes, such as parasite density-dependent differentiation into transmission life-stages and specific host immune responses. Our analysis focuses first on a single parasitemia peak and then on the dynamics of multiple peaks that rely on stochastic switching between groups of parasite variants. We show that the interplay between the two types of parasite control within the host: spe-

cific and general, depends on the modular structure of the parasite antigenic archive. Our modelling reveals that the degree of synchronization in stochastic variant emergence (antigenic block size) determines the relative dominance of general over specific control within a single peak, and can divide infection scenarios into stationary and oscillatory regimes. A requirement for multiple-peak dynamics is a critical switch rate between blocks of antigenic variants, which depends on host characteristics, such as the immune delay, and implies constraints on variant surface glycoprotein (VSG) archive genetic diversification.

Secondly, we study the interactions between the structure and function of the antigenic archive at the transmission level. By using nested modelling, we show that the genetic architecture of the archive has important consequences for pathogen fitness within and between hosts. We find host-dependent optimality criteria for the antigenic archive that arise as a result of typical trade-offs between parasite transmission and virulence. Our analysis suggests that different traits of the host population can select for different aspects of the antigenic archive, reinforcing the importance of host heterogeneity in the evolutionary dynamics of parasites.

Variant-specific host immune competence is likely to select for larger antigenic block sizes. Parasite tolerance and host life-span are likely to select for whole archive expansion as more archive blocks provide the parasite with a fitness advantage. Within-host carrying capacity, resulting from density-dependent parasite regulation, is likely to impact the evolution of between-block switch rates in the antigenic archive. Our study illustrates the importance of quantifying the links between parasite genetics and within-host dynamics, and suggests that host body size might play a significant role in the evolution of trypanosomes.

In Chapters 4 and 5 we consider the genetics behind trypanosome antigenic variation. Antigen switch rates are thought to depend on a range of genetic features, among which, the genetic identity between the switch-off and switch-on gene. The subfamily structure of the VSG archive is important in providing the conditions for this type of switching to occur.

We develop a hidden Markov model to describe and estimate evolutionary processes generating clustered patterns of genetic identity between closely related gene sequences. Analysis of alignment data from high-identity *VSG* genes in the silent antigen gene archive of the African trypanosome identifies two scales of subfamily diversification: local clustering of sequence mismatches, a putative indicator of gene conversion events with other lower-identity donor genes in the archive, and the sparse scale of isolated mismatches, likely to arise from independent point mutations. In addition to quantifying the respective rates of these two processes, our method yields estimates for the gene conversion tract length distribution and the average diversity contributed locally by conversion events. Model fitting is conducted for a range of models using a Bayesian framework. We find that gene conversion events with lower-identity partners are at least 5 times less common than point mutations for *VSG* pairs, and the average imported conversion tract is short. However, due to the high frequency of mismatches in converted segments, the two processes have almost equal impact on the rate of sequence diversification between *VSG* sub-family members. We are able to disentangle the most likely locations of point mutations vs. conversions on each aligned gene pair.

Finally we model *VSG* archive diversification at the global scale, as a result of opposing evolutionary forces: point mutation, which induces diversification, and gene conversion, which promotes global homogenization. By adopting stochastic simulation and theoretical approaches such as population genetics and the diffusion approximation, we find how the stationary identity configuration of the archive depends on mutation and conversion parameters. By fitting the theoretical form of the distribution to the current *VSG* archive configuration, we estimate the global rates of gene conversion and point mutation. The relative dominance of mutation as an evolutionary force quantifies the high divergence propensity of VSG genes in response to host immune pressures.

The success of our models in describing realistic infection patterns and making predictions about the fitness consequences of the parasite anti-

genic archive illustrates the advantage of using integrative approaches that bridge between different biological scales. Even though quantifying the genetic signatures of antigenic variation remains a challenging task, cross-disciplinary analyses and mechanistic modelling of parasite genomic data can help in this direction, to better understand parasite evolution.

Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Introduction

Pathogens interact with their hosts in complex ways, using subtle strategies for immune evasion and for establishing chronic infection (Frank, 2002; Schimd-Hempel, 2008). Current research in immunology and microbiology shows that parasite evasion of host immunity occurs in a wide range of pathogens, employs a wide range of molecular mechanisms and generates a variety of pathogenic outcomes.

Among the most sophisticated parasite survival strategies is antigenic variation (Barbour *et al.*, 2006), where individuals within the proliferating parasite population change their surface molecules and so present different antigens to the host. This phenotypic variation is achieved in different ways: by mutation, recombination or by the expression of archived genetic variants.

At the within-host level, antigenic variation allows the parasite to avoid impending antibody responses and characteristically yields an oscillating parasite load over long periods (Barbour & Restrepo, 2000; Barry, 1997). By affecting infection duration, host infectiousness and immune history, antigenic variation over the course of individual infections, has also significant epidemiological implications at the population level, often contributing to the persistence of infectious diseases like malaria and sleeping sickness in many countries. More generally, the amount of new variation and the types of new variants generated have an impact on antigenic polymorphism and the pace of evolutionary change (Frank, 2002).

The phenomenon of immune evasion was discovered more than 100 years ago by Paul Ehrlich, who in 1908 reported for the first time on the 'disappearance of receptors' in African trypanosomes, a mechanism now known as antigenic variation. Since then, different aspects of antigenic variation have been studied at various levels of detail and in different systems, including the major parasite groups such as bacteria (*E.coli*, *Borrelia hermsii*), viruses (*HIV*) and protozoa (*Plasmodium falciparum*, *Trypanosoma brucei*).

One of the most intensively studied systems of antigenic variation is that of the parasite African trypanosome (see Barry & McCulloch (2001) for a review). *Trypanosoma brucei* is a protozoan, transmitted by the tsetse fly, found in Equatorial and sub-Saharan Africa, able to infect a wide range of host species, including humans, livestock and wild mammals. It generally causes the chronic disease, human sleeping sickness, very difficult to treat, and poses a heavy economic burden to the communities affected. Given the devastating effects of this disease, there is a real need for further studies, better control programmes and drug development.

With increasing availability of genomic data (Berriman *et al.*, 2005), recent advances in our understanding of trypanosome within-host processes and mechanisms of antigenic variation (Lythgoe *et al.*, 2007; Marcello & Barry, 2007b; Turner, 1999) are calling for an integrated framework, where parasite genetics and within-host parasite dynamics can finally be linked. This integration has the potential to deepen our understanding of the population dynamics of trypanosomes and similar pathogens within hosts, reveal their intrinsic dependence on parasite genetic processes, and ultimately highlight their consequences for between-host transmission and disease epidemiology. Such cross-scale understanding is becoming essential for controlling infectious disease (Haydon & Mathews, 2007). Within-host population ecology of antigen switching pathogens is not a new topic but increasing access to genetic data provides us with a rapidly widening opportunity to understand the evolutionary ecology of antigenic variation.

The main focus of this thesis are African trypanosomes and their antigenic diversity. Our aim is to investigate the role and consequences of trypanosome antigenic variation at different biological levels, starting from the within-host level, then moving towards between-host population dynamics and finally addressing some of the genetic processes that might be responsible for shaping the parasite diversity at the molecular

level. The generation and maintenance of the antigenic variation machinery is perhaps the most important life-history trait of this pathogen. Addressing the interplay of both top-down and bottom-up factors involved in this parasite trait remains a key challenge for understanding trypanosome evolution.

1.2 African Trypanosomes

1.2.1 Medical and economic significance

African trypanosomes pose a severe problem in much of sub-Saharan Africa because of the pathogenic effects of their infections and the huge socio-economic losses resulting from the disease. Disease in livestock is caused primarily by the three tsetse-transmitted trypanosome species: *T.congolense*, *T. vivax* and *T.brucei*. Possibly as a result of their co-evolution, trypanosomes are often non-pathogenic in many wild animals, and some breeds of cattle, such as N'Dama (Taylor, 1998).

There are three main subspecies of *Trypanosoma brucei*: *T.b. brucei*, *T.b. gambiense* and *T.b rhodesiense*. Although the three subspecies can infect wildlife and livestock, only *T.b. gambiense* and *T.b. rhodesiense* can infect humans, with significant differences in their epidemiology and hence prospects for control. The epidemiology of human African trypanosomiasis (HAT) has been comprehensively covered by Welburn *et al.* (2004) and an epidemiological update is available from the WHO (WHO, 2006). For a review on trypanosome molecular epidemiology see (Hide & Tait, 2004).

T.b. gambiense infection causes a chronic, slow wasting disease which can be asymptomatic for months or even years, and ultimately causes death if untreated. This is known as *sleeping sickness* and occurs in West and central Africa (Welburn *et al.*, 2001) and is transmitted by tsetse flies of the *Palpalis* group (see Figure 1.1). The term 'sleeping sickness' comes from the symptoms of the neurological phase of infection, which include confusion, reduced coordination, and disruption of the sleep cycle, fatigue punctuated with manic periods leading to daytime slumber and night-time insomnia. Humans are the principal reservoir of *T.b. gambiense*. For the infected human the disease develops first from a hemolymphatic stage with mild symptoms, including fever, headaches, joint pains and itching, into a second stage, where the parasites

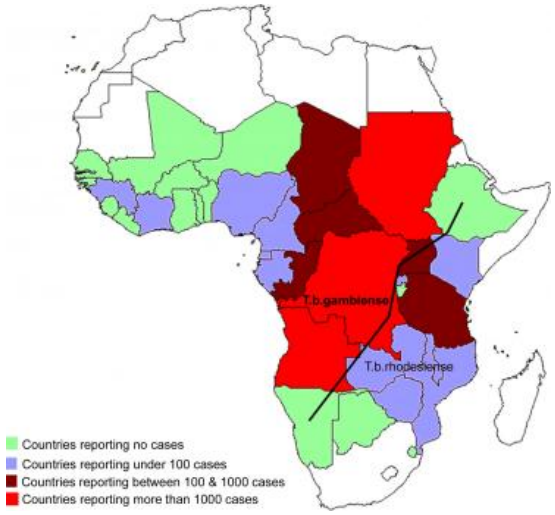


Figure 1.1: Geographical distribution of Trypanosomiasis in Africa showing the epidemiological status of countries considered endemic for the disease. Map obtained from Simarro *et al.* (2008). doi:info:doi/10.1371/journal.pmed.0050055.g003.

cross the blood brain barrier and establish a cerebral infection. However, there is uncertainty about the actual duration of both stage 1 and stage 2 infection, particularly with respect to how long a patient remains infectious (Checchi *et al.*, 2008). These differences can be exploited for disease control. A review of published evidence on the natural progression of gambiense HAT in the absence of treatment emphasizes the need for further studies of chronic carriage and human trypano-tolerance.

The evolution of a well-tuned system of immune evasion enables these parasites, similar to other vector-borne pathogens (Barbour & Restrepo, 2000), to establish infections that can persist within the human host for considerable periods of time. Antigenic variation allows the parasite to be one step in front of the hosts' immune defence mechanisms, therefore, all efforts to develop a vaccine against trypanosomiasis have been ineffective. Acquired immunity is rarely fully protective against reinfection and its efficacy seems to depend on the duration and intensity of past exposure to infection. In many instances, the mechanisms that facilitate parasite persistence and repeated host infection are not sufficiently understood.

T.b. rhodesiense is zoonotic, it causes a much more acute form of the disease which is fatal within weeks if untreated. *T.b. rhodesiense* occurs east of the Rift Valley

and its vectors are flies from the *Morsitans* group. Its predominant reservoir is cattle. Uganda is the only country where both forms of HAT are known to occur. Despite having separate foci, there has been concern that the two subspecies will soon overlap because of the transport of infected cattle to Western Uganda (Picozzi *et al.*, 2005).

Lastly, the subspecies *T.b. brucei* occurs throughout sub-Saharan Africa. It is not infective to humans because it is rapidly lysed in the human serum, but it infects most animal agriculture. Trypanosomiasis in domestic animals is termed 'nagana'. Nagana is of much economic importance because where it is prevalent, meat, milk, dung and draft power production are greatly reduced or lost altogether.

Current chemotherapy treatment for HAT is complex, since special drugs have to be used for the different development stages of the disease, as well as for the subspecies of parasite concerned (Brun *et al.*, 2010). Pentamidin is a drug efficient only against *T.b. gambiense* and applied only in the first stage of the disease. Its action against the parasite is mainly to inhibit the uptake of critical nutrients such as cholesterol and phospholipids. Melarsoprol is the only approved drug for effectively treating both subspecies of human African trypanosomiasis in its advanced stage. Its uptake by the trypanosome is highly toxic and leads to rapid lysis of the parasite. However, the drug's potency is constrained due to a severe side effect: encephalopathy, a series of permanent brain injuries which develop in 5% of treated patients. In addition to the deleterious treatment with Melarsoprol, the number of drug-resistant strains of *T. brucei* seems to be increasing. Mechanisms of drug resistance have been elucidated and the need for promising strategies for research into new anti-parasite compounds is urgent. However, the optimal ways in which to devise HAT control policies in light of the differing biology and epidemiology of the parasites depend on the wider aspects of control policy, including the responsibilities of individuals, governments and international organisations in control programmes (Fevre *et al.*, 2006).

1.2.2 Antigenic variation and the VSG archive

Trypanosomes have a coat composed of around 5.5×10^6 molecules of variant surface glycoprotein (*VSG*) covering their whole surface. The gene encoding this protein undergoes variation, through the sequential switching of different archived copies into expression sites (Berriman *et al.*, 2005). VSG genes are parasite antigens that stimulate

a specific immune response because each molecule contains one or more epitopes recognized by host antibodies. The parasites expressing a particular VSG gene represent a particular antigenic variant. Immune evasion occurs when a specific immune response against one antigenic molecule fails to recognize a variant antigenic molecule. The antigenic variants usually differ at one or more epitopes, the sites recognized by specific immunity. The sequential replacement of different variants through the course of an infection leads to an oscillating parasite load (Figure 1.2), characteristic of many antigenically-varying pathogens.

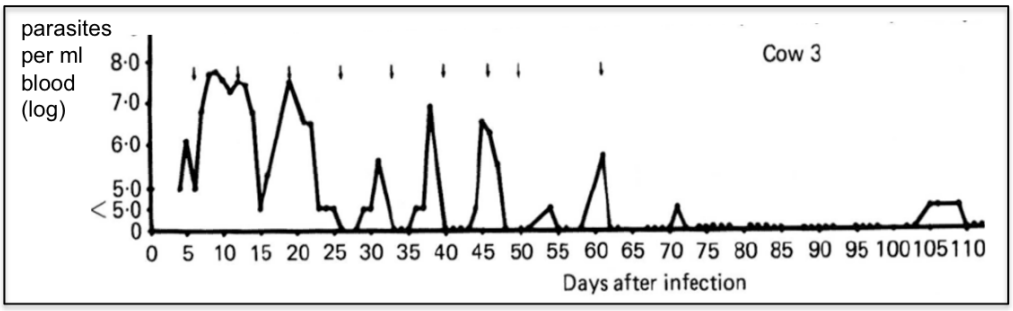


Figure 1.2: A typical trypanosome infection in a cow develops as a series of parasitaemia peaks. Each relapse corresponds to a new group of antigenic variants being expressed by the parasite. Image found in (Barry, 1986), where the particular host depicted eventually self-cured and cleared the infection. Longitudinal parasitaemia profiles such as this one are scarce in the literature, as they are hard to obtain experimentally. The majority of studies typically focuses on laboratory mice infected with *T.b. brucei* and analyzes only the first few weeks of an infection.

Different VSG genes are selected through a switching process from an archive of $\sim$2000 silent *VSG* genes encoding different forms of this protein. In contrast to viral antigenic variation obtained through continuous mutation, the antigenic variation of trypanosomes and other protozoa relies on expression of genes stored in the parasite genome. Switches in expression occur at a rate of up to 10^{-2} per cell division (Turner, 1997). The switch mechanism is mainly gene conversion of archival copies into a transcriptionally active bloodstream expression site (BES). Gene conversion is a type of intragenomic recombination that converts all or part of the target sequence without altering the donor sequence. Only one out of around 20 possible BES is active at any time. Thus, the parasite can also change expression by switching between different

expression sites. However, the mechanisms leading to switching between transcription sites is not fully understood (Morrison *et al.*, 2009).

Typical of antigenic variation, trypanosome infections show hierarchical variant expression, in which some variants appear early in infections and others appear progressively later (Barry, 1986; Capbern *et al.*, 1977; Gray, 1965; Morrison *et al.*, 2005). This hierarchical expression has been suggested to be mediated by genetic properties of different *VSG* genes, such as their location in the parasite genome (Barry, 1997; Marcello & Barry, 2007a; Morrison *et al.*, 2005; Turner, 1999). Generally during an infection, intact VSG genes, located in telomeric regions of minichromosomes are activated first, subsequently followed by expression of subtelomeric array VSG genes, and finally 'mosaic' VSG genes assembled from pseudogenes start to be expressed and dominate in the chronic phase Morrison *et al.* (2009); Thon *et al.* (1990).

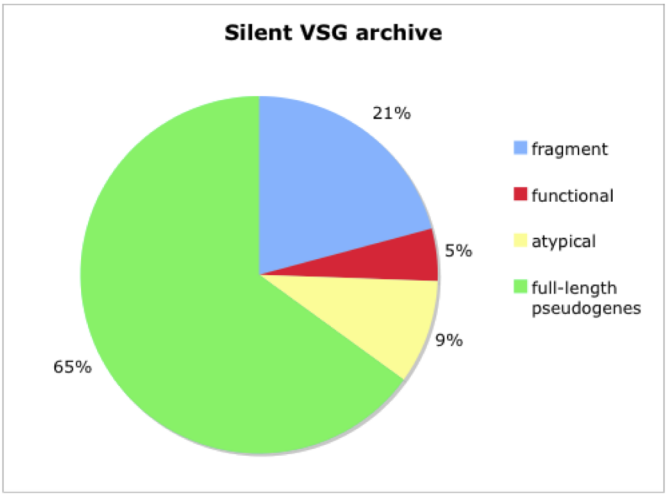


Figure 1.3: The structure and composition of the genomic reservoir of VSGs was determined as part of the genome project (Berriman *et al.*, 2005). There are between 1000 and 2000 potential VSG sequences, however among them, only 5% encode functional VSG open reading frames (ORFs) and of the remainder 9% encode an ORF for a VSG with atypical primary structure. The rest, $\approx$ 65 % consist mainly of disrupted VSG ORFs (pseudogenes) containing frame shifts and/or stop codons and the remainder are fragmentary genes (Marcello & Barry, 2007a).

In fact, most archive *VSG* genes ($\sim$ 95%) are pseudogenes (Fig.1.3), i.e. nonfunctional genes unable to code for protein, typically with omitted sections or premature

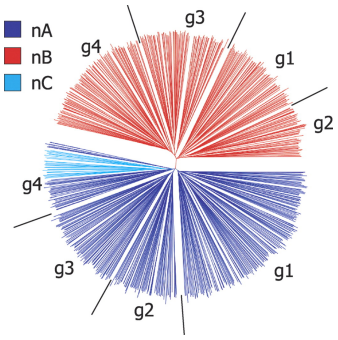
stop codons, and they can be expressed only after recombining multiple times with other members of the archive to produce an intact, mosaic gene, a set of events that occur with low probability. 'Mosaic' VSGs genes can be formed by replacing part of the expressed VSG with a structurally homologous region from the archive. The combinatorial nature of mosaic formation together with the vast silent VSG archive provides the parasite with a theoretically limitless antigenic variation potential, and is the primary barrier to vaccine development.

As shown by Marcello & Barry (2007a), about 60% of VSG genes are unique, distantly related to other VSGs, the rest occur in subfamilies of two to four close homologs ($>$ 50% peptide identity). This small subfamily structure of the archive seems to be fundamental in providing the interacting donors for mosaic formation (Marcello & Barry, 2007a), and thus for the maintenance of chronic infection.

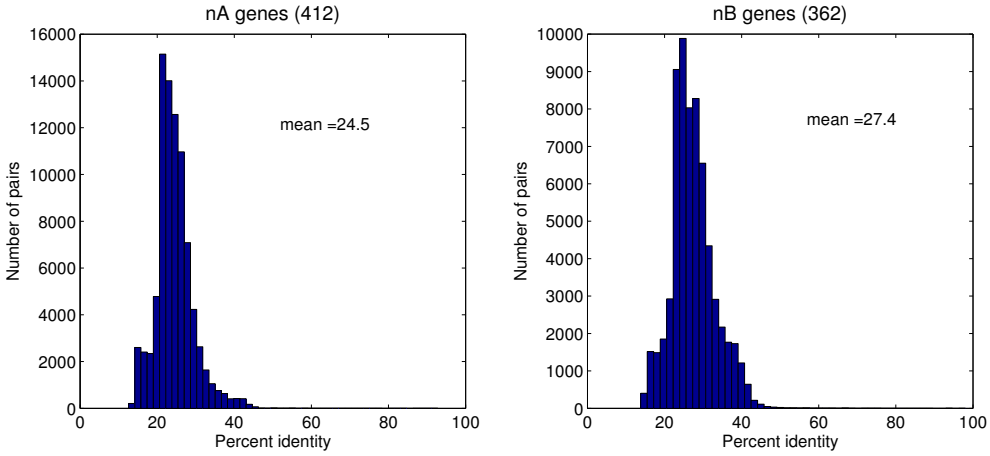
Within each locus type, there appear to be finer degrees of ordering, and recent studies by Frank (1999) and Lythgoe *et al.* (2007) have shown that antigen switch rates can range widely, in particular over discontinuous orders of magnitude. However, what exactly controls switch rates at the genetic level remains only partly understood. The general consensus seems to point so far towards intrinsic properties of the gene that is currently expressed, such as locus type, properties of the gene that is being activated such as number of repeats in the flanking regions, and/or mutual properties of the gene pair, such as pairwise homology of their two sequences.

1.2.3 The importance of VSG N-terminal domains

VSG genes of African trypanosomes consist of a hypervariable N-terminal domain of 350-400 residues and a more conserved C-terminal domain of 40-80 residues - encoding the portion of the VSG protein that is anchored to the plasma membrane (Carrington *et al.*, 1991). The N-terminal domains encode for the portion of the VSG protein that contain exposed surface loops thought to bear the variable epitopes (Hsia *et al.*, 1996; Miller *et al.*, 1984) and determine antigenicity. These domains can adopt a very similar alpha-helical coiled-coil higher order structure (Blum *et al.*, 1993), and have been categorized into three types A, B, and C, according to features of the primary structure: the location of the conserved cysteine residues and the presence of a heptad repeat in the region underlying the coiled coil (Figure 1.4(a)). As the two VSG



(a) VSG N-terminal domain



(b) Identity distribution

Figure 1.4: a) N-terminal VSG domains. Image from (Marcello & Barry, 2007b). Neighbor-joining tree based on ClustalX-generated multiple sequence alignment of predicted peptide sequences of 725 sequences. The three domain types A, B, and C, are colored individually (by HyperTree) as shown in the key. b) Empirical histograms of nucleotide pairwise identity across genes in the two major VSG N-terminal types, nA and nB domains, reflect a high variability among genes in this region.

domains differ in structure, function, and evolution and are subject to independent recombination, they are usually considered separately.

The VSG gene family is large and complex, displaying diversity in both domains (Carrington *et al.*, 1991). Genomic sequence determination of T. brucei strain TREU 927 has now greatly expanded the VSG sequence data set, allowing for more global analyses of sequence variability, which might shed light on the balance between structural constraints, the processes of epitope diversification and the antigenic variation potential of this parasite. It is thought that telomere-proximal genes such as those located on minichromosomes are activated first during the course of antigenic variation in African trypanosome infections, and that later, subtelomeric genes and mosaic genes assembled from pseudo-VSG genes are expressed. Amongst N- terminal domain types, only about one in three is theoretically "functional" (Marcello & Barry, 2007a), suggesting that the basis for a high proportion of the archive being degenerate is defective C termini, through presence of stop codon(s), frameshift(s), or disruption of cysteine pattern.

As a result, one third of N-terminal domains possibly can be utilized directly, by combining with a functional C-terminal domain from, for example, the VSG already present in the expression site and give rise to a mosaic gene. The genetic identity among VSG genes underlies the formation of expressed mosaic genes. High identity between the coding sequences of donor and recipient genes is believed to increase the chances of homologous recombination. This process is crucial during the chronic stages of African trypanosome infections, when the parasite has already expressed all functional antigens stored in its genome, and starts to express new mosaic genes. It is thus likely that the pairwise genetic identity configuration in the VSG archive (see Figure 1.4(b)) is subject to constant modulation by evolutionary forces that tend to optimize it, in order to facilitate mosaic gene formation, so important for the survival of this parasite within its vertebrate host. Despite the importance of processes such as point mutation and gene conversion on the general diversification of multi-gene families, their role on the evolutionary dynamics of VSG archive diversification in African trypanosomes has not yet been quantified.

Figure 1.5: Trypanosoma brucei live extracellularly in the blood of the vertebrate host. Image from http://www.pnas.org/content/100/3/F1.medium.gif.

1.2.4 Growth, differentiation and host immunity

This seemingly simple genetic basis for variation interacts at the within-host population level, however, with a number of parasite processes, including growth, death and differentiation, and host processes, most notably acquired immune responses against the antigens that are subject to antigenic variation. A considerable amount is known about parasite and host factors influencing antigenic variation in trypanosomes. In the mammalian host, the parasite lives extracellularly in the bloodstream (see Figure 1.5). There are two morphological forms of the parasite: the dividing "slender" form and the non-dividing "stumpy" form. It is only the stumpy form that can effectively infect the tsetse (Dean *et al.*, 2009), because it is pre-adapted to the vector environment and thus is able grow in the tsetse salivary glands.

In the mammalian host, slender cells differentiate to stumpy cells in a density-dependent manner (Reuner *et al.*, 1997). In addition to the parasite's ability to sense and respond to its local environment, another general feature of protozoan parasites is such a capacity to arrest their development. Frequently, protozoan parasites stop proliferation before undergoing differentiation that generates transmissible forms. This strategy is similar to the quiescence manifested in other eukaryotic cells, used to tolerate harsh environmental changes. Developmental arrest is thought to have the advantage of limiting the risk of lethal DNA replication errors that can occur during cell division (Mathews, 2011). Crucially, for parasites, it can also prevent uncontrolled proliferation in the host, which if unregulated could lead to host damage or death and so limit transmission potential. Consequently, the total parasite population reaches a

maximum carrying capacity. In trypanosomes, this has been measured in immunosuppressed hosts, and has been shown to equal the maximum height of infection peaks in normal infections (Balber, 1972; Hajduk & Vickerman, 1981; Luckins, 1972).

There is also symmetry in growth across different variants. Thus, there is maximally only a 10% variance in the growth rates of trypanosome clones isolated from the same infection, and any differences do not correlate with the *VSG* expressed (Seed, 1978). In addition, the kinetics of induction of immunity (Gray, 1965) and variant clearance (Hajduk & Vickerman, 1981) are common between variants. The general picture, at least for variants that appear early in infection, is that variant-specific responses arise rapidly and persist for prolonged periods (Gray, 1965; Morrison *et al.*, 2005), and that cross-protection among variants does not occur. Recently, however it has been shown that the capacity of a host to control an infection may be limited: in inbred mice, the supply of naive B-cells can decrease dramatically during a trypanosome infection, with the subsequent decrease in the rate of production of specific antibodies (Radwanska *et al.*, 2008).

Antibody (IgM and IgG) seems to be the most important part of the immune response against trypanosomes, which live in the bloodstream. Antibodies can damage parasites directly, enhance their clearance by phagocytosis, activate complement, block their entry into host cells and possibly neutralise parasite products. However, antibodies produced during infection have potentially pathogenic consequences if they bind to 'self' antigens, are of low specificity, block binding of protective antibodies or mediate complement depletion. Comparative studies of trypanotolerant and trypanosensitive breeds of cattle have revealed that there are significant differences in their antibody responses against trypanosomes and pathogenic effects experienced (Taylor, 1998).

So far, most of the empirical studies of trypanosomes have been based on laboratory experiments performed on mice and have considered only the first weeks of infection. Due to their low survival of trypanosome infections, mice may not be ideal hosts for studying the full scale and characteristics of antigenic variation. Experimentation on more natural hosts such as cattle or sheep might yield better insight into the true mechanisms involved.

A definitive analysis of the variant-specific antibody response developing throughout the chronic phase has however not yet been achieved. As argued by Taylor (1998), this would require the quantitative identification of all the variants expressed and the

associated variant-specific antibodies. Because antibodies are absorbed and cleared from circulation, serum antibody titres do not accurately reflect antibodies produced at the cellular level. Perhaps, antibody production rather than titres needs to be determined. Another difficulty comes from the observation that due to their capacity to attach to endothelial cells lining blood vessels, or to invade tissues, trypanosomes circulating in the blood may not be representative of the entire population. This means more accurate methods for sampling parasite populations are required. Thus, key in extracting all the necessary parameters for within-host dynamics of trypanosomes will be experiments that : 1) look far into the chronic infection stages; 2) analyze antigenic variation in more natural hosts, such as cattle; 3) measure the right quantities in terms of parasite numbers, antibody titres and antibody production; and 4) quantify pathogenesis and immunosuppression caused by trypanosome infections.

1.3 Models of antigenic variation

Several theoretical approaches have been used to investigate the within-host population dynamics of antigenically-varying pathogens. These can be broadly categorized as those which: (1) postulate that novel antigen types arise in a continuous manner (Antia *et al.*, 1996; Nowak *et al.*, 1990; Sasaki, 1994); (2) explicitly take into account the switch pathways between distinct antigen variants (Agur *et al.*, 1989; Frank, 1999; Kosinski, 1980; Lythgoe *et al.*, 2007). The first group of models are perhaps more appropriate to model antigenic drift in viruses, where new antigenic variants are produced by point mutations at various sites within the pathogen genome, such as in human immunodeficiency virus (*HIV*) infections, whereas the second group of models could be viewed as representing the situation of a fixed antigenic archive, where new variants appear through the sequential activation of existing silent genes. The latter models are more appropriate for protozoan pathogens such as *Trypanosoma brucei*, the causative agent of sleeping sickness, and *Plasmodium falciparum*, the malaria parasite, and bacteria. No single model or class of models is likely to explain entirely a phenomenon as complex as antigenic variation; however each of these types of model has improved our understanding of several important aspects of this process.

Additionally, there are models which neglect antigenic variation altogether, but focus rather on specific aspects of the kinetics of host-parasite interaction. These models

or 'sub-models' have attempted to isolate parasite growth and differentiation (Savill & Seed, 2004; Tyler *et al.*, 2001) or immune response activation and have explored these features of the dynamics generally over shorter time-scales, for example by modelling only the early stage of infection, in which antigenic variation plays only a minor part. These models have been largely motivated by empirical data, and thus have provided the foundations for the more complex modelling work by yielding concrete estimates of key parameters. Below we describe briefly some results of a few models on antigenic variation from the last 30 years, grouping them by their main motivating questions.

1.3.1 Variant order

One of the earliest attempts to model antigenic variation of African trypanosomes is the study by Kosinski (1980), who was interested in explaining the ordered sequential emergence of variants through an infection. The model simulated a trypanosome clone with 90 possible variants, with widely differing variant-specific growth rates, random variant origin and variant eradication by a persisting host immune response. Among the parameters varied were maximum parasitaemia and the growth rate differential between 'fast' and 'slow' variants. The study concluded that random generation and selection by growth rate alone could not produce the degree of variant orderliness reported in the literature.

In a subsequent study, Agur *et al.* (1989) modelled within-host antigenic variation dynamics of trypanosomes. Again, the focus of the model was to explain the order in the appearance of different variants. They considered the existence of switch-intermediates as a possible important factor, assuming that there were variants which could express simultaneously two VSGs in the process of switching from one major type to another. Hence, in an archive of N variants, their model allowed for N^2 possible variants to arise. The dynamics of each 'pure' variant n, on the diagonal, eliciting only one variant-specific immune response, were given by:

$$\frac{dv_n}{dt} = v_n \left[r_n \left(1 - \frac{V}{K} \right) - u a_n \right], \tag{1.1}$$

where u denoted the mortality coefficient from immune response a_n, r the intrinsic growth rate, and K the total carrying capacity. On the other hand, for the switch-

intermediates, off-diagonal, experiencing also cross-reactivity, the dynamics were:

$$\frac{dv_{n,m}}{dt} = v_{n,m}\left[r_{n,m}\left(1-\frac{V}{K}\right)-(u'_{n,m}a_n+u''_{n,m}a_m\right], \tag{1.2}$$

where $u'_{n,m}$ and $u''_{n,m}$ were the mortality coefficients due to anti-n and anti-m antibodies. The equations for the immune responses were given by:

$$\frac{da_n}{dt} = c_1 b_n\left[\frac{v_n}{v_n+C}\right]-c_2 a_n, \tag{1.3}$$

where c_1 was the antibody secretion coefficient, and c_2 the antibody removal coefficient. Agur *et al.* (1989) varied the switch probabilities from single expressor variants to double expressor variants and vice-versa, but the total product of the two resulting in the switch probability between two major variants was kept fixed. Their study concluded that it was not possible to explain the ordered appearance of variants by affecting the growth coefficients of single expressors or double expressors, or by affecting the antigen switch probabilities. Rather, a realistic parasitaemia could be obtained if the majority of switches had a double expressor switch-intermediate phase and if the double expressors had a differential susceptibility to the immune control.

In malaria, Recker *et al.* (2004) proposed a theoretical framework to account for antigenic variation among the variant surface antigens of *Plasmodium falciparum*, that relies entirely on suppression by transient partially cross-reactive immune responses. The dynamics of each variant i were given by:

$$\frac{dy_i}{dt} = \phi y_i - \alpha z_i y_i - \alpha' w_i y_i, \tag{1.4}$$

where each variant has net growth rate ϕ and is cleared at a rate α by the specific long-lasting immune response z_i, and at a rate α' by the transient immune responses w_i against minor shared epitopes. The dynamics of the specific immune response z_i against strain i was given by:

$$\frac{dz_i}{dt} = \beta y_i - \mu z_i, \tag{1.5}$$

where the parameter β denoted the per capita stimulation rate by antigen, and μ is the rate of decay of the immune response. The dynamics of the transient cross-reactive immune response against the minor shared epitopes was given by:

$$\frac{dw_i}{dt} = \beta' \sum_j y_j - \mu' w_i, \tag{1.6}$$

where j refers to all variants that share these epitopes with i. This framework intentionally excludes switching between variants as a means of structuring the appearance of antigenic variants. The model by Recker *et al.* (2004) showed that within a homogeneous host environment, variant-specific immunity alone is unable to promote immune evasion, but in combination with partially cross-reactive responses can give rise to sequential dominance of different antigenic variants. In particular, they showed that the relative efficacy of transient immune responses increased linearly the peak parasitaemia within the host and led to a decrease in the duration of infection.

After the early modelling work on the antigenic variation of African trypanosomes, the study by Lythgoe *et al.* (2007) was the first to consider explicitly the connection between the structure of the antigenic switch matrix and the genetic mechanisms giving rise to this structure. Their model was given by:

$$\frac{dv_i}{dt} = r_t v_i \left(1 - e^{-(V+M)/K}\right) - d a_i v_i + \sum_j (s_{ji} v_j - s_{ij} v_i), \tag{1.7}$$

$$\frac{dm_i}{dt} = r_t v_i e^{-(V+M)/K} - \delta a_i m_i, \tag{1.8}$$

$$\frac{da_i}{dt} = c_i (1 - a_i) \left(\frac{v_i(t-\tau) + m_i(t-\tau)}{C}\right)^x, \tag{1.9}$$

where besides growth at rate $r_t = re^{-t/\rho}$ and antigen switching at rate s_{ij} from i to j, the dynamics of each variant i was affected also by density-dependent differentiation to the stumpy non-dividing stage (m_i). Both parasite forms experience immune-mediated clearance at specific rates d and δ. The immune response, given by a_i was assumed to grow in a saturating manner from 0 to 1, following stimulation by antigen above a threshold C, with a characteristic delay τ, and general sensitivity x. By considering four different types of switch matrices: uniform, inter-dependent, differential and homology-dependent, Lythgoe *et al.* (2007) ruled out certain switching patterns and obtained realistic parasitaemias only for the differential and homology-based switch matrices, explaining variant order on the basis of a large variation, by orders of magnitude, in the variant differential switch rates.

1.3.2 Diversity threshold and immunity

Focusing on a different pathogen, in a model of HIV dynamics, Nowak *et al.* (1990) proposed the existence of a pathogen diversity threshold to be critical in antigenic

variation dynamics. Their model consisted of the following equations:

$$\frac{dv_i}{dt} = f_i(v_i,y) - v_i(s_i z + p_i x_i) \quad i = 1,2,..,n \tag{1.10}$$

$$\frac{dy}{dt} = K - dy - uvy \tag{1.11}$$

$$\frac{dx_i}{dt} = kv_i y - uvx_i, \quad i = 1,2,....n \tag{1.12}$$

$$\frac{dz}{dt} = k^{'} vy - uvz, \tag{1.13}$$

where v_i denoted the virus population of strain i, y denoted the total CD4+ cell count, x_i denoted the strain-specific CD4+ cells, and z the cross-reactive immune response. The reproductive per capita rate of each virus strain was defined as $f_i(v_i,y) = (r_i^{'} + r_i y)v_i$. In particular, one special case was also considered where the virus strains had the same characteristics $r_i = r, p_i = p, s_i = s$.

Nowak *et al.* (1990) showed, that when the number of antigenically distinct virus strains is below a diversity threshold in the host, given by $n_c = px/(r - sz)$, the immune system is able to regulate viral population growth, but when the diversity is above this threshold, the virus population induces the collapse of CD4+ lymphocyte cells. This model was the first to suggest that antigenic diversity was the cause and not the consequence of immunodeficiency disease.

The later theoretical study by Antia *et al.* (1996) showed that the interplay between antigen-specific and cross-reactive immune responses and pathogen diversity is crucial for the within-host control of the antigenically-varying pathogen. Their model consisted of equations for the dynamics of each variant, p_i, the variant-specific immune response, x_i, and the general immune response z:

$$\frac{dp_i}{dt} = rp_i(1 - P/c) - kp_i x_i - k^{'} p_i z, \tag{1.14}$$

$$\frac{dx_i}{dt} = \rho x_i \left(\frac{p_i}{p_i + \phi} \right) - \mu x_i, \tag{1.15}$$

$$\frac{dz}{dt} = \rho^{\prime} z \left(\frac{P}{P + \phi^{\prime}} \right) - \mu^{\prime} z. \tag{1.16}$$

The number of parasite variants changed over time as new antigenic types were produced at a constant rate m. Antia *et al.* (1996) argued that early in infection, when

pathogen diversity is low, variant-specific immune responses are responsible for controlling the pathogen. Later on in infection, after the number of variants surpasses a certain threshold $n_c = \left(\frac{\mu'\phi'}{\rho'-\mu'}\right) / \left(\frac{\mu\phi}{\rho-\mu}\right)$, cross-reactive immune control dominates and variant-specific immunity tends to zero. When the rate of production of novel antigenic variants is sufficiently small, the parasite population reaches within-host equilibrium before the next variant has appeared, yielding a parasitemia that increases in a stepwise manner, and since the rate of production of new variants in this case is proportional to the steady-state parasite density P, successive variants appear more and more rapidly. When the critical number of variants is reached, then the total parasitemia remains constant. In contrast, when the rate of new variant generation is high, Antia *et al.* (1996) argue that there will not be sufficient time for the equilibrium to be attained before a new variant is generated. In this case there is a high but fluctuating parasitemia, for example, as observed in trypanosome infections. If there is only variant-specific immunity, then with the passage of time the total parasitemia saturates at the carrying capacity; but when cross-reactive immunity is present, the total parasitemia is controlled at a much lower level by the cross-reactive immune response.

1.3.3 Evolution of antigenic variation

A new theoretical perspective on antigenic variation was brought by Sasaki (1994), who studied the evolution of antigen drift/switching, based on within-host interactions between the pathogen and the host immune system. In Sasaki's study, the evolutionarily stable mutation or switching rate is the value that maximizes the total number of transmissions from the infected host to other hosts, i.e. the total reproductive value of the pathogen (stationary pathogen density in a host). He considered three cases: (i) the case where the antigen types could be indexed on a one-dimensional lattice (stepping-stone model) and an antigen could mutate to either one of its two immediate neighbours with equal rate; (ii) the case of an infinite-allele model, assuming that any mutation at the antigen-determining sites of the pathogen genome produces a novel antigen type; (iii) the case of a finite antigen repertoire, where each antigen type could switch to one of the other types with equal probability.

The dynamics for the simplest (stepping-stone) model were given by:

$$\frac{dN_k(t)}{dt} = [r - \beta B_k(t)]N_k(t) + \mu \times \left[p \left(\frac{N_{k+1}(t) + N_{k-1}(t)}{2} \right) - N_k(t) \right] \quad (1.17)$$

$$\frac{dB_k(t)}{dt} = \alpha N_k(t) B_k(t) \quad (1.18)$$

where N_k denoted each population of antigen variants, and B_k the corresponding immune response in the host. The parameter μ represented the mutation rate per unit of time and p the fraction of mutations that contribute to the alteration of antigen. The parameters α and β represent the efficiency of the immune response. The infection started with just one pathogen type, which was allowed to mutate to neighbouring types with equal probability as the infection progressed. Approximating the space of antigenic types $\{k\}$ by a continuous variable, a partial differential equation was derived and analyzed.

Among their important findings, the models by Sasaki (1994) showed that in order to establish a persistent infection, where the growth rate of each new antigenic variant is positive, the pathogen should evolve an intermediate mutation rate, because of the inherent trade-off that the pathogen faces between immune evasion on one hand, and loss of progeny by lethal mutations on the other. There exists a critical level for the fraction of deleterious mutations, above which the pathogen cannot maintain itself in the host, whatever the mutation rate. In the finite repertoire model, Sasaki showed that there is a critical size of the repertoire below which the pathogen cannot succeed in giving rise to a persistent infection.

In another important paper, Frank (1999) was the first to use a within-host dynamics model to analyze the switch pathways between different antigens through an evolutionary approach. His antigenic variation model, comprising n variants ($i = 1, ..n$) was given by:

$$\frac{dp_i}{dt} = rp_i(1 - P/c) - kp_i a_i + \sum_j v_{ij} p_j - p_i \sum_j v_{ji}, \quad (1.19)$$

$$\frac{da_i}{dt} = \rho a_i m_i \left(\frac{p_i}{p_i + \phi} \right) - \mu a_i + \pi m_i, \quad (1.20)$$

$$\frac{dm_i}{dt} = \gamma m_i^2 (1 - m_i/\delta) + b m_i \left(\frac{p_i}{p_i + \phi} \right) - d m_i + \varepsilon, \quad (1.21)$$

where p_i denoted the level of parasite variant i, a_i the variant-specific immune response elicited in the host, and m_i the variant-specific immune memory. Key features of this model were the logistic growth of the parasite up to carrying capacity c, the switching dynamics given by the v_{ij} terms, the saturating nature of antibody production as a result of antigen stimulation, and the contribution of immune memory to faster clearance for secondary appearances of the same variant over infection.

The analysis and simulations performed by Frank (1999), varying the switch matrix S through evolutionary steps, whereby each entry was randomly determined by $M \times 10^{-E}$, suggested that wide variations in the magnitudes of switch rates could be favoured by natural selection on the basis of their beneficial effect to the parasite, as they prolonged infections and maximized total parasitaemia within host.

As discussed above, each modelling approach is different and motivated by different theoretical questions, often suited to a particular pathogen. The challenge remains to exploit existing findings and refine antigenic variation models to consider pathogen and immune dynamics in greater levels of detail and biological realism, informed and validated by new empirical findings. In particular, the wealth of genetic data coming from the sequencing of pathogen genomes must be exploited to achieve a higher resolution of immune evasion at the molecular level and use this knowledge to make more quantitative and predictive models.

Furthermore, availability of longitudinal trypanosome infection data and a better quantification of host immune responses could bring substantial improvements to the study of dynamical processes occurring between this pathogen and its host. The final aim is to reach a thorough understanding, both qualitative and quantitative, of all mechanisms involved in host-parasite interaction and use this understanding to predict infection outcomes, and work towards concrete control measures against the antigenically-varying pathogen.

1.4 Trypanosome epidemiology

At the population level, the persistence of a disease as complex as human trypanosomiasis depends on the interplay between many factors, among which are the vertebrate

host, the parasite and the vector responsible for transmission (Figure 1.6). Epidemiology tries to study the entire complex, by modelling the interactions between these populations and their consequences for disease prevalence, especially in humans (Simarro *et al.*, 2008). Of the three interactions, probably the greatest importance for explaining epidemiological findings has been attributed to the understanding of host-vector relations, and more recently to social, political and economic factors (WHO, 2006).

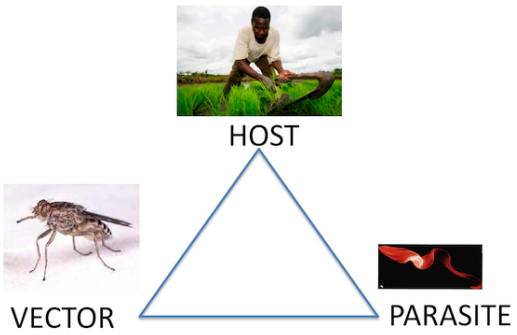


Figure 1.6: The triangle of epidemiological interactions mediating persistence of Human Trypanosomiasis in Africa. Successful control of this infectious disease at the human population level depends on the integration of measures targeting the human and non-human hosts, the tsetse population and the parasite interactions with them.

In the early part of the twentieth century, human African trypanosomiasis (HAT), also known as sleeping sickness, decimated the population in many parts of sub-Saharan Africa. As a result, in the 1930s, the colonial administrations, aware of the negative impact of the disease on its territories, established disease control programmes. Systematic screening, treatment, and follow-up of millions of individuals in the whole continent led to transmission coming to a near halt by the 1960s.

However, with the advent of independence in most HAT-endemic countries, the newly independent authorities had other priorities to consider. The rarity of HAT cases, and a decline in awareness of how the disease might rebounce, led to a lack of interest in disease surveillance. Over time the disease slowly returned (Figure 1.7), and some thirty years later, flare-ups were observed throughout past endemic areas (Simarro *et al.*, 2008). This led to the WHO beginning intensive control measures from 1995 onwards. Between 1997 and 2006, the gambiense form responded well to

intensive control activities mainly focused on the human reservoir (the animal reservoir was considered to play only a minor role in transmission). The number of people under active surveillance increased, and the number of new cases decreased (WHO, 2006), as illustrated in Figure 1.8. However, control activities focusing on the human *T. b. rhodesiense* reservoir were found insufficient to control the disease, probably due to the impact of the animal reservoir on transmission. Thus, *T. b. rhodesiense* showed only a small decrease in the number of cases.

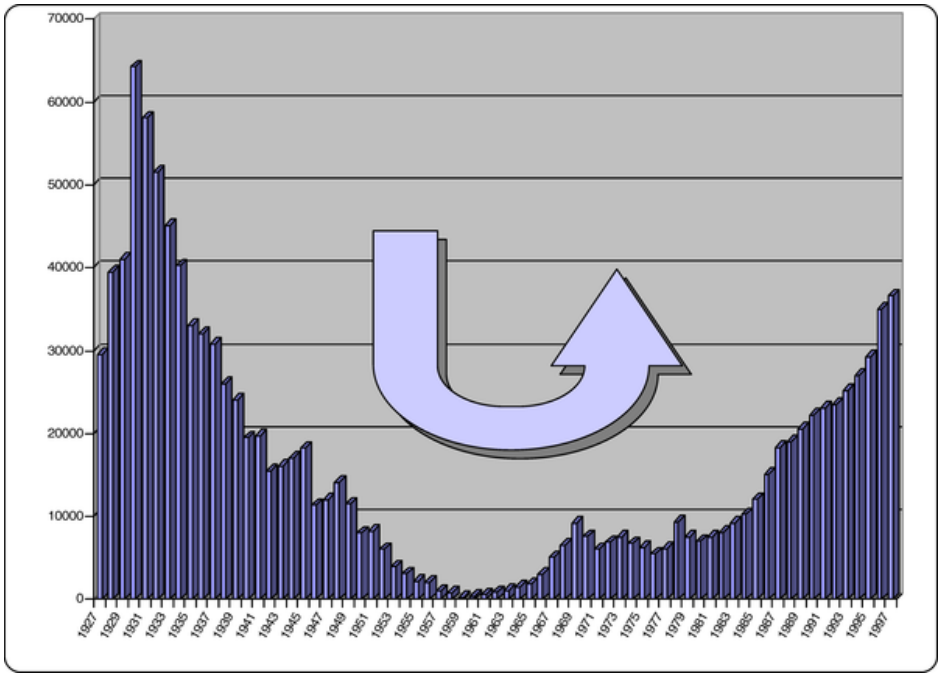


Figure 1.7: The number of new cases of human African trypanosomiasis in Africa in the period 1927-1997. Intensified control efforts resulted in a sharp decrease in the 1960s, but failure to maintain them over longer periods led to a disease rebounce in the 1990s. Image from (Simarro *et al.*, 2008).

1.4.1 Host-vector interaction

Trypanosomes can infect a wide range of host species, but often only one or a few species act as a reservoir (Rogers, 1988). For *T.b. gambiense*, which consists of the 90% of HAT reported cases, the human population is thought to serve as the main

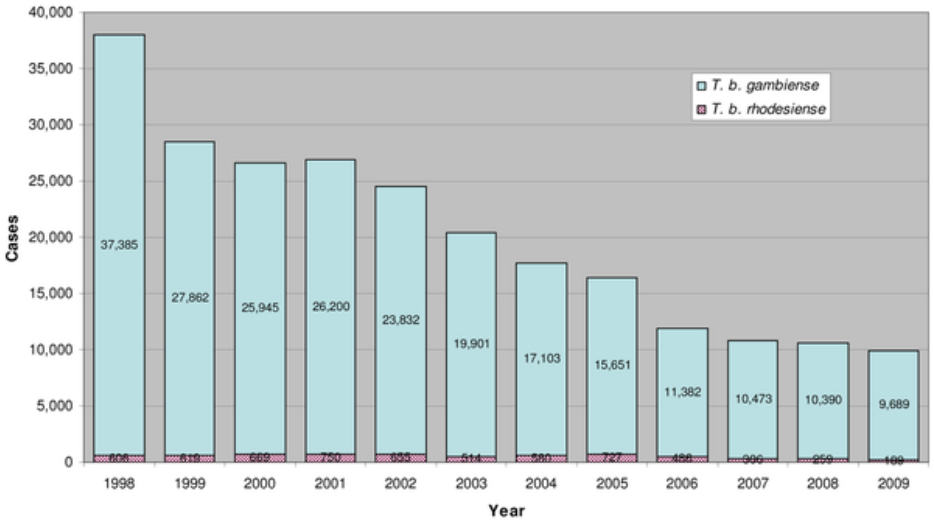


Figure 1.8: A steady decrease in the number of new cases of human African trypanosomiasis occurred in Africa after the 1990s, when new control and surveillance measures were implemented by the WHO. *T.b. gambiense* remains the prevalent parasite species causing disease in 95% of the cases. Image from (Simarro *et al.*, 2011).

reservoir, whereas for *T.b rhodesiense*, representing about 10% of the reported cases, the reservoir is thought to consist of natural wild populations. African trypanosomiasis is essentially a disease affecting agrarian populations and its incidence is greatly influenced by the level or type of human activities, through their effect on the degree of contact with the fly. The proximity of humans to vector habitat, sustained human-fly contact, the feeding habits and frequency of vectors, their relative abundance, seasonal and climatic effects, etc. are probably all important determinants of human-fly-human transmission. On the one hand, human infections reduce labour resources, while on the other, livestock infections limit availability of meat and milk and deprive African farmers of draught animal power, substantially minimising crop production. Thus, both human and animal trypanosomiasis contribute to the underdevelopment of the African continent, and are considered a major obstacle in the establishment of a flourishing agriculture to provide food security and to lead to sustainable economic growth and healthy populations.

At high human population densities, if the land is cleared for agricultural use to such an extent that the tsetse habitat is substantially reduced, the disease is inevitably

eliminated. At lower population densities, however, the type, rather than the extent of human activities and their relation to the ecology of the fly, seems to largely determine disease incidence. For example, activities such as collecting water and firewood, washing, fishing, hunting, and cultivation, when practiced near tsetse habitat, especially in more severe climatic conditions such as draught, are thought to greatly enhance the chance of disease transmission (Willett, 1963).

Current interventions aimed at vector control involve the use of insecticides (through the sequential aerosol spraying technique, insecticide-treated targets (Vale *et al.*, 1988), or insecticide-treated insects (Bauer *et al.*, 1995; Hargrove *et al.*, 2000); the use of traps (Brigtwell *et al.*, 1991); and the sterile insect technique (SIT) (Vreysen *et al.*, 2000).

The sequential aerosol technique uses extremely low concentration of insecticide through several consecutive aerial sprayings and can effectively clear large areas of tsetse flies in a relatively short time, but it is expensive and requires substantial economic and infrastructure support. Selective spraying application of insecticides to animals on which tsetse feed are another effective means of vector control. Odour-baited targets or traps have been used in many countries to effectively suppress tsetse populations. The relative low cost and simplicity of this approach recommends it for use by local communities, but its scale of application is so small that control efforts are bound to be challenged by re-invasion. While effective baits have been developed for savannah tsetse, to date no such baits exist for riverine tsetse, which are major vectors of HAT, promoting continued research efforts in this direction. The SIT, which involves the release of laboratory-reared and sterilised males to compete with wild males so that females inseminated by them produce no offspring, has been effectively used for eradication of tsetse (G. austeni), for example, in Unguja Island in Zanzibar (Vreysen *et al.*, 2000). However, due to the exorbitant cost of SIT, the feasibility of this approach in areas where multiple species are present remains doubtful. Despite the considerable progress made in controlling the vector, an ideal methodology easily accessible to the population at risk is still missing.

1.4.2 Host-parasite interaction

One of the main ways in which the characteristics of trypanosome strains may influence the epidemiology of sleeping-sickness and the efficacy of control measures is

through the effect of the mildness or severity of the infection produced on transmission. If the infection is mild, the chance of human-fly-human transmission is enhanced, both because of the prolonged infectiousness period and because infected hosts are commonly not so incapacitated as to be confined to their huts and thus out of contact with flies (Simarro *et al.*, 2008). If the trypanosome strain is more similar to *T.b. rhodesiense* instead, producing a much more acute infection, the infectious period is shorter, and any infected host is more likely to be confined inside and out of contact with the vector for much of that time.

However, even if under these assumptions one would expect the human reservoir to remain unchanged in the presence of mild strains and decrease in the presence of more virulent strains, the existence of animal reservoirs is likely to favour the maintenance of more virulent strains in the population. Virulent strains are likely to be capable of infecting and being transmitted from many hosts- for example wild animals, with their greater resistance to trypanosomiasis, or domestic animals such as cattle, pigs, goats and sheep, which are also likely to display low susceptibility to infection by any but the more virulent strains of the *T. brucei* group. Indeed, recent surveys of trypanosomes in cattle, have revealed a high prevalence of infections, suggesting that domestic animals are potentially far more probable to act as pathogen reservoirs than previously thought (Taylor, 1998).

A fuller understanding of within-host parasite dynamics for each important host will be crucial in quantifying its relative contribution to global disease transmission. Integrating health systemsbased approaches across different countries will be crucial in areas affected by *T. b. gambiense*, while in areas affected by *T. b. rhodesiense*, disease control cannot rely exclusively on human health services and will probably have to involve veterinary and entomological interventions as well.

The human disease takes two forms, chronic and acute, depending on the parasite involved. A person infected with *Trypanosoma brucei gambiense* can be infected for months or even years without major signs or symptoms of the disease. When symptoms do emerge - such as severe headaches, sustained fever, sleep disturbances, alteration of mental state, and neurological disorders - the patient is often already in an advanced disease stage where the central nervous system is affected. *Trypanosoma brucei rhodesiense*, found in eastern and southern Africa ($<$10% of reported cases), causes an acute infection. The first signs and symptoms- such as chancre, occasional

headaches, irregular fevers, pruritus, and the development of adenopathies - are observed after a few weeks or months. Following this first stage, when the parasite has invaded the blood and lymph subsequent to the infective bite from the fly, the disease develops rapidly into a second stage when parasites cross the blood-brain barrier, invading the central nervous system.

Better diagnostic tools are needed to establish more accurately the prevalence of the disease in populations at risk and its different stages. Attempts to identify new antigens should result in more specific and sensitive tests for serodiagnosis of the disease, while changes in test format (i.e., the development of non-invasive saliva tests) should result in more user-friendly tests (Simarro *et al.*, 2008). Much progress has been made in the development of molecular tools. Specific genes for both *T. b. gambiense* and *T. b. rhodesiense* have been identified for PCR-based detection of infection, however further research in this direction to improve accuracy is needed.

1.4.3 Vector-parasite interaction

Being a vector-transmitted parasite, the chance of trypanosome transmission depends on the ability of tsetse flies to transmit and the insect life-cycle and behaviour. Besides genetic differences across potential vectors, factors such as temperature, humidity, habitat type (bush, riverine, forest, savannah, etc.) can play an important role in vector survival, reproduction, feeding habits, movement and dispersal, which in turn affect the host-vector contact rate, vector abundance and vector spatial distribution (Krafsur, 2009; Rogers & Randolph, 1991; Willett, 1963).

So far, data characterizing vector-parasite interactions for trypanosomes are scarce, because large field studies are intrinsically very complex and expensive, but obviously, collecting such data, at different spatial scales remains key for understanding disease dynamics and for envisioning control strategies which reduce prevalence, such as successful targeting of the vector, or protection of the environment (e.g. avoid habitat fragmentation, degradation), crucial for maintaining the vector population at a balance. One promising prospect has come from the suggestion that data from meteorological satellites could be used to predict both the mortality rate and the abundance (key determinants of disease transmission potential) of tsetse over very large areas of

the African continent, and to produce maps of high risk zones of disease transmission for the African trypanosomiases and, by implication, for many other vector-borne diseases (Rogers & Randolph, 1991).

For appropriate sustainable control, it is desirable to quantify more subtle parasite-tsetse interactions characteristics (Aksoy *et al.*, 2003), such as infection prevalence in the vectors, its age distribution profiles (Woolhouse *et al.*, 1993), feeding preferences of tsetse (Farikou *et al.*, 2010; Kohagne *et al.*, 2010), and the diversity of the parasite strains harboured by insects, influencing typanosome recombination (Peacock *et al.*, 2011) and strain repertoire divergence (Hutchison *et al.*, 2007). Important areas of recent interest are the management of insect resistance to insecticides (Hargrove *et al.*, 2000), the manipulation of vector behaviour such as host choice, and the investigation of vector symbionts that could interfere with trypanosome infection and onward transmission to new hosts (Aksoy *et al.*, 2005).

1.4.4 Mathematical models and perspectives

Models of transmission of trypanosomes by tsetse flies have been derived to some extent from models for malaria transmission by mosquitoes. Many of the principles and applications are relevant to descriptions of transmission of general vector-borne parasitic diseases. In general, there are important mechanisms through which the demography of the host population(s) influences the persistence of the infective agent, and vice-versa, mechanisms through which the parasite affects host population growth and persistence.

Several ordinary differential equation models (ODEs) have been developed for analysing the dynamics of trypanosome transmission between cattle and tsetse populations (Baker, 1992; Milligan & Baker, 1988; Rogers, 1988), the economic impact of trypanosomiasis (Habetemariam *et al.*, 1983), and the effects of vaccination (McDermott & Coleman, 2001). They are based on sets of inter-related classes of the host (S-I-R) and the vector (S-I). Numerical simulations as well as analytical frameworks have been used. In particular, the basic reproduction number of the pathogen (Diekmann *et al.*, 1990), defined as the expected number of secondary cases per primary case in a totally susceptible population, has been calculated for such models and used as an index of the capacity of the parasite to cause an epidemic. Within the idealized

deterministic description, when $R_0 > 1$, the newly introduced parasite starts to spread exponentially in the host population, while going extinct when $R_0 < 1$. Different types of intervention can be conceived, aimed at reducing R_0 below 1. These can range from insecticide spraying (increase tsetse mortality) to host vaccination (reduce host susceptibility) or administration of drugs, and the models can indicate which one among potential strategies has the highest impact on R_0.

Although the mathematical approaches used in these models on trypanosomes are standard and not new in epidemiology, the challenge remains to correctly parametrize them, and develop the appropriate functional forms that determine the qualitative nature of the links between the different subpopulations in consideration. Often, at this level of modelling, the biological details of host-parasite interaction are overlooked, and one deals with average characteristics such as average infection duration, average parasite virulence, average host susceptibility and so on. Depending on the type of question one is interested in answering, and the time scale of consideration, these details become more or less important and the models have to be refined accordingly.

As emphasized in the book by Diekmann & Heesterbeek (2000), the distinction between different host types can have a profound influence on epidemiological characteristics such as R_0, especially when it involves a correlation between infectivity and susceptibility. In that case, to model host heterogeneity, a structured population approach is more appropriate, and the analysis within such framework is far from trivial.

Additionally, a promising modelling approach to deal with the intrinsic dependences between the within-host and the between-host level is the nested modelling approach, as illustrated by Gilchrist & Sasaki (2002), which can go beyond the demographic time scale and make evolutionary considerations about host-parasite interactions, e.g. the evolution of virulence. In fact, starting with the early work by Anderson & May (1982); Ewald (1983); Levin & Pimental (1981), the whole theory of adaptive dynamics (or, evolutionary dynamics, when genetics is taken into account) aims to understand how evolution by natural selection has shaped the various ways in which pathogens exploit their hosts and are transmitted (Diekmann *et al.*, 1996).

Despite the strong and robust mathematical theory that has been developed to analyze various aspects of pathogen adaptive dynamics, these results have not yet been translated to the specific case of trypanosomes, and their significance has not yet had an impact on our understanding of trypanosome dynamics as a whole. As usual, the

modelling approaches undertaken so far reflect the questions that have motivated researchers in their study of trypanosome infections, and when focused on a particular timescale (e.g. demographic) the insights obtained can be limited. However, with increasing availability of theoretical frameworks that can incorporate pathogen and host evolution, a much deeper understanding of infectious disease dynamics and its ecological repercussions can be achieved. The growing body of genetic data, coming from recent advances in throughput technologies and genome sequencing, can serve as a major frontline for the validation and testing of theoretical evolutionary predictions.

1.5 Diversity at the genetic level

While ecological theory can tell us *why* certain pathogen life-history traits should evolve in certain directions that optimize transmission or pathogen persistence in the field, the answer to *how* these optima are achieved, lies in the genetics. It is by investigating genetic processes at the level of the parasite genome, that we can begin to map the structure at the micro scale to function at the macro scale, and unveil the long evolutionary trajectories that brought each particular pathogen to the present day. It is an enormous challenge, but the potential benefits are also high.

As illustrated by Grenfell *et al.* (2004), the phylodynamic framework provides one such potentially promising template for integrating the comparative population dynamics, population genetics and phylogenetics of microparasites. Phylogenies of pathogens are determined by a combination of immune selection, changes in pathogen population size and spatial dynamics. By studying phylogenetic data, one can investigate the influence of selective and nonselective processes acting on parasite populations through time. Pathogen population dynamics and genetics influence each other in a variety of ways, and integrated frameworks that mechanistically bridge between the two are needed to quantify these interdependencies. While some progress has been made in this direction for some viruses such as intra-host HIV and human influenza A, availability of data showing long-term intra-host phylogenies of protozoan infections are scarcer or missing altogether.

One of the yet unsolved puzzles for trypanosomes is understanding the sophisticated antigenic variation system, whose molecular basis is increasingly being exposed

by genome sequencing (Morrison *et al.*, 2009). The functional links between the various interacting components at the VSG gene level, such as genetic recombination, gene duplication and gene diversification, which mediate antigenic variation, have only partially been understood (Marcello & Barry, 2007a). The large number of genes involved in *T.brucei* antigenic variation provides a unique opportunity for studies which can disentagle the role of genetic processes such as gene conversion and point mutation in the diversification of the silent VSG archive (Morrison *et al.*, 2005), and for understanding the modular structure of antigenic switch rates (Frank & Barbour, 2006; Lythgoe *et al.*, 2007).

However, sequence diversity does not necessarily imply antigenic dissimilarity. A key priority remains determining the genotype-phenotype relation (Marcello & Barry, 2007b), and the extent of antigenic overlap between different VSGs. The relative specificity versus cross-reactivity between different variants is very important both for within-host dynamics, - as high levels of cross-reactivity generally induce larger competition among parasite subpopulations, - and between-host dynamics, with implications on the selective pressure on parasite diversity. VSG genetic data could shed light on the precise interference among trypanosome antigenic variants, which could inform vaccine design. It would be interesting to learn whether there are variants which display interference or facilitation. Particular variants could potentially induce a memory response that interferes with the host's ability to generate a specific response to other variants, or could stimulate a very general response which would be detrimental to many other variants. Thus, the antigenic archive may be shaped both by the requirement to avoid cross-reaction and by the degree to which variants suppress the immune response to other variants (Frank, 2002; Turner, 1999).

Furthermore, parasite surface antigens often play a role in attachment and entry into host cells or attachment to certain types of host tissue, as it has been shown for the malaria parasite (Reeder & Brown, 1996). Varying these attachment properties allows attack of different cell types or adhesion to different tissues. Such variability, in addition to antigenic variability can provide the parasite with additional resources or protection from host defences, thus increasing the benefits of antigenic variation. Do such phenotypic correlations occur also in trypanosomes? At present, the answer is not clear.

Concerning the antigenic switch rates, Frank (1999) has hypothesized that wide variation in switch rates between antigenic variants confers a fitness advantage to the parasite, and that for this reason it might be selected as an optimal configuration. The rate at which switches occur is likely to affect the ability of the parasite to extend an infection. Switching too quickly might lead to archive exhaustion, whereas switching too slowly might lead to parasite clearance before the novel antigenic types are expressed. Thus, natural selection can strongly influence the molecular machinery of antigenic variation, in order to adjust for the appropriate rate of switching between archival variants.

With regards to molecular findings, two main types of switching have been observed in trypanosome antigenic variation: hierarchical switching of intact genes ($\sim$ 10% of the VSG archive), mediated by the 70bp repeats in their flanking regions, and homology-based switching for pseudogenes ($\sim$ 90% of the archive), when expressed as mosaics (Morrison *et al.*, 2009).

Short repeated nucleotide sequences often lead to high error rates during replication. Short repeats have recurring units typically 1-5 bases per unit. Errors seem to arise when a DNA polymerase either skips forward a repeat unit, causing a deletion of one unit, or slips back one unit, producing a one-unit insertion. These insertions or deletions can turn on and off gene expression, by causing frameshift mutations that hamper translation and production of a full protein (Frank, 2002). As yet, this switch mechanism has not yet been quantified for trypanosomes, although the prevailing view is that certain repeats in the flanking regions lead to higher baseline activation rates of VSG genes, and that hierarchical expression may be due to hierarchical differences in repeat length and homology across intact genes (McCulloch & Barry, 1999).

The mosaic VSG gene switches instead apparently rely on pairwise nucleotide sequence identity: the more similar two genes, the higher their chances of interacting via gene conversion to form a mosaic (Marcello & Barry, 2007a). Again, the precise quantification of this process is lacking, due to the highly stochastic and complex nature of mosaicism, but one could potentially envision a situation whereby capturing the *sequence identity-gene conversion* interplay could lead to a crucial understanding of the process that 'fuels' the chronic stage of infection. If one could somehow block or limit the mosaic formation capacity of the parasite, the parasite load could be controlled in the host, and the severe neurological phase of chronic infection perhaps

avoided. Furthermore, the requirement that the VSG archive should harbour a good number of high-identity potential interacting donors, and maintain a suitable global identity configuration that favours mosaic formation can be translated to the genetic processes responsible for shaping identity between genes.

All together, a series of open questions remain in the field of VSG diversity at the genetic level, especially with regards to the molecular mechanisms determining the switch rates and the extent to which they are adaptable. How is the repeat number in the flanking regions of intact VSG genes kept at an optimum via contraction and expansion processes? How is the pairwise genetic identity configuration in the VSG archive kept at an optimum for mosaic gene formation, through the opposing forces of mutation and gene conversion? A crucial challenge remains to determine the extent to which switch rates can evolve to enhance parasite fitness and the factors that might shape this evolution at the molecular level (Frank, 2002). Modelling approaches run in parallel with experiments can help in this. Switch rate evolution can be tested both by selecting *in vitro* for faster or slower switch rates, and *in vivo*, comparing scenarios that imposed different immune pressures on rates of switching, and on particular variant expression orders. Such studies would allow one to link the molecular mechanisms of switching to the adaptive significance of switching.

Finally, zooming out to the large ecological context, and investigating trypanosome diversity across geographical areas and different host species in the field should bring additional insight into the evolutionary dynamics of this pathogen. An illustration comes from field isolates collected from Uganda revealing many differences both in the size and composition of their antigenic repertoires (Hutchison *et al.*, 2007). A natural question is: does the history of variation within and between strains reflect host bias, geographical barriers or simply neutral evolution? Modern genetics can shed light on the precise signatures discernible from cross-species genetic comparison, of parasite-host adaptation and the mechanisms that enable successful infection of a wide host range.

1.6 Outline of this thesis

In this thesis, we use a series of mathematical and computational models to investigate first the links between within-host antigenic variation processes and trypanosome

transmission between hosts, and secondly the impact of genetic processes on the evolutionary dynamics of trypanosome VSG archive diversification. One goal is to move towards a deeper and more quantitative understanding of the central role played by antigenic variation in parasite fitness across biological scales.

We start in **Chapter 2** with a model describing the within-host dynamics of trypanosomes, and analyze the various ways in which the antigenic archive structure, mirrored in the switch rates impacts on infection outcome. The novelty of our approach lies in the stochasticity of new variant generation and the modular structure of the VSG archive captured by three parameters: the number of variants in one block, the number of blocks and the between-block switch rate, whose effects on the dynamics we analyze. The results of this chapter have appeared in (Gjini *et al.*, 2010). To assess the implications of within-host dynamics on field transmission, we use a nested modelling approach in **Chapter 3**, where we show how the interplay between antigenic archive structure and host ecological characteristics modulates parasite within-host fitness and further the basic reproduction number R_0. Our modelling reveals that different host traits can act as evolutionary drivers for different aspects of the antigenic archive. In **Chapter 4** we focus on the genetic identity within VSG gene subfamilies, a crucial factor in mosaic gene expression in chronic infections. We describe a hidden Markov model approach to estimate the evolutionary processes generating clustered patterns of genetic diversity between subfamily members. In addition to quantifying the respective local rates of gene conversion and point mutation, our approach yields estimates for the gene conversion tract length distribution and the average diversity contributed by conversion events. In **Chapter 5** we derive a general diffusion approximation equation to model the evolutionary dynamics of pairwise identity in the VSG archive. Here, the balance between diversification and homogenization processes leads to a theoretical global stationary identity distribution. We fit empirical VSG gene identity data to this theoretical distribution to obtain global estimates for mutation and gene conversion rates.

In **Chapter 6**, we bring the results of the different chapters together and discuss our findings on quantitative aspects of the antigenic variation dynamics of African trypanosomes and the impact of various genetic factors shaping their successful immune evasion machinery. With the emerging knowledge of the structure of their antigenic archive, we argue that the development of cross-scale syntheses, as the one attempted

in this thesis, should enable us to generate fresh insights about their function, and the deeper evolutionary ecology of pathogens in general. The potential to control infectious disease by reducing the extent to which pathogens access antigenic diversity is discussed as a promising prospect for the future.

Chapter 2

Modelling within-host trypanosome dynamics

2.1 Introduction

Basic parasite dynamics models have been essential in understanding quantitative issues of pathogen replication within hosts in the study of infectious diseases, including malaria, HIV infections, influenza, hepatitis, African trypanosomes, etc. (Anderson & May, 1991). Since the 1980s, pathogen dynamics modelling has developed following the initial population dynamics framework proposed in the 1920s by Volterra (1926) and Kermack & McKendrick (1927), and it has proven crucial for studying several aspects of infectious disease. Together with the advanced understanding of the vertebrate immune system in the late 1970s, the first conceptual models of the immune system were developed (Bell *et al.*, 1978), describing the production, proliferation and differentiation of cellular compartments of the immune system such as T-cells, CD8+ cells and antibodies in response to stimulation by foreign antigen. The field of immune system modelling has progressed substantially in the last decade, owing also to the development of computational and visualization tools that have helped to capture even the subtlest features of immune activation (Rapin *et al.*, 2006).

Such developments have increasingly enabled the theoretical study of pathogen dynamics within hosts. The temporal parasite turnover in infected patients, the dynamics during drug therapy, the depletion of host resources, the evolution of drug-resistant variants, the immune-suppression mechanisms employed by infectious agents etc. are

but a few among the many infectious disease characteristics modeled in the literature (Antia & Lipsitch, 1997; De Boer & Perelson, 1994; Ganusov *et al.*, 2002; Nowak *et al.*, 1990). More recently, within-host models have begun to provide insights into stochastic and population genetic effects, addressing questions about the role of recombination and mutation in pathogens (Althaus & Bonhoeffer, 2005), and their evolutionary dynamics (Luciani & Alizon, 2009).

Connecting the dynamics of pathogen replication at the level of single cells, to parasite population level processes and the interaction with the immune system of the host remains challenging. Despite the progress made, crucial questions are still open, even for very intensely studied infectious diseases such as HIV/AIDS. The factors determining clearance or control of infections are not completely understood. Similarly, the factors mediating pathogen immune evasion and chronic infection are still a matter of debate as to what extent host or parasite-intrinsic factors are involved. It is necessary to integrate new findings on the molecular and cellular processes of the life cycles of pathogens and immune response kinetics within modelling frameworks that mechanistically capture their mutual interactions.

The interest in modelling within-host trypanosome dynamics originates in the1980s. The majority of antigenic variation studies on trypanosomes (Agur *et al.*, 1989; Antia *et al.*, 1996; Frank, 1999; Kosinski, 1980; Lythgoe *et al.*, 2007; Seed, 1978; Turner & Barry, 1989) have so far focused on isolated aspects of the within-host parasite dynamics, such as variant order, switching pathways or host immunity. In this chapter, we aim to combine ideas from previous models and formulate a more general framework to investigate chronic infection dynamics in more detail. Our aim is to dissect mechanisms operating at two important temporal scales: within a single parasitaemia peak and across many peaks. We expect different processes to be critical at each scale.

We propose a mathematical model to study the mechanisms driving trypanosome antigenic variation over the course of an infection, and those maintaining chronicity. In particular, we focus on the interplay between the structure of the antigenic archive, mirrored in the switch rates between antigenic variants, and within-host processes, such as parasite differentiation to a non-dividing form and the specific immune responses. A better understanding of the interplay between antigenic archive structure and within-host processes could be instrumental in the design of control strategies and further experimental research.

Our goal is to address the potential and limitations of antigenic variation as a mechanism for sustaining chronic infection. The balance between general control of the infection mediated by parasite differentiation, and variant-specific control mediated by host immunity is first studied, and two crucial infection characteristics are examined: infection duration and peak parasite load, which have direct implications for parasite transmission and host survival.

We study other aspects of chronic infection dynamics and several possible infection scenarios, such as oscillating parasite loads and the antigenic composition of each peak. Our analysis of model outcomes is divided in two parts: first focusing on the expression and dynamics of blocks of variants, and then on the switching between subsequent blocks. By looking at a range of feasible parameter combinations, and further considering two simple model modifications that include immune suppression and cross-reactive immunity, we stress the role of the balance between host and parasite factors. Our aim is to offer additional insight into and improve understanding of the features driving trypanosome infections, but also to highlight more general aspects of host–pathogen interaction.

2.2 Model

Our model is formulated to capture the primary features of within-host trypanosome dynamics and advances the one proposed by Lythgoe *et al.* (2007). The model is composed of a deterministic and a stochastic part, the deterministic being responsible for the dynamics of variant growth and clearance, the stochastic part representing the randomness in first arrival times for each variant.

2.2.1 Variant dynamics

We distinguish two within-host forms of the parasite: slender (dividing) and stumpy (non-dividing, transmissible to the tsetse fly, a vector for the parasite). We denote the parasite number in the i-th variant subpopulation $i = 1,...,N$, by v_i and m_i, for slender and stumpy cells respectively, and the specific antibody response by a_i for variant i. N is the size of the parasite antigen gene archive.

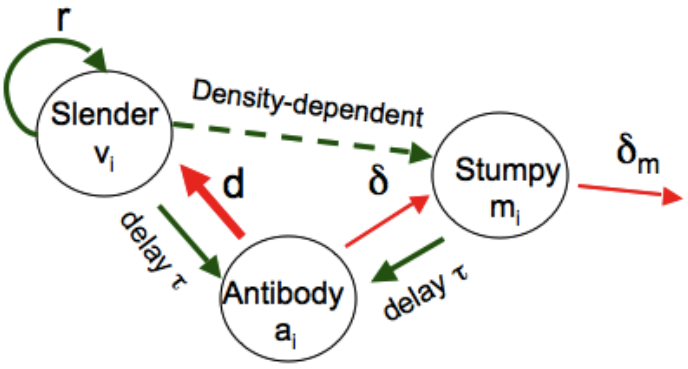


Figure 2.1: Diagram illustrating the deterministic dynamics of an arbitrary variant, described by Eqs. 2.1-2.3. The arrows indicate the kinetic interactions between the different parasite subpopulations and specific antibody responses. The red arrows indicate clearance, the green arrows indicate growth or stimulation, and the dashed arrow indicates the slender-to-stumpy differentiation.

The dynamics for each variant and antibody response (see Figure 2.1) are given by:

$$\frac{dv_i}{dt} = rv_i\left(1-\frac{V+M}{K}\right) - da_iv_i, \tag{2.1}$$

$$\frac{dm_i}{dt} = rv_i\frac{V+M}{K} - \delta a_i m_i - \delta_M m_i, \tag{2.2}$$

$$\frac{da_i}{dt} = c\left(\frac{v_i(t-\tau)+m_i(t-\tau)}{C}\right)^x (1-a_i), \tag{2.3}$$

where $V=\sum_i v_i$ and $M=\sum_i m_i$ denote the total number of slender and stumpy parasites. Each slender cell divides with an intrinsic per capita division rate, r, which is a constant throughout the infection. Unlike the Lythgoe model (Lythgoe *et al.*, 2007), r does not decay exponentially. An exponentially decaying r as a function of time gradually blocks parasite replication independently of the immune response. Besides there being no clear biological motivation for this assumption, its inclusion in the model would confound the mechanisms responsible for long-term clearance of the parasite.

A key feature of parasite dynamics is the density-dependent differentiation of each slender cell to the stumpy form, where K is the total within-host parasite carrying capacity. Our choice of the differentiation rate $rv_i(V+M)/K$ describes the linear way in which the total parasite population size influences the rate of production of stumpy forms. This linear differentiation function differs from the exponential function used

by Lythgoe *et al.* (2007), because it allows the slender cell population to decrease as a result of differentiation, thus inducing oscillations in the intrinsic (immune-free) slender-stumpy dynamics and it gives rise to a clear stable parasite persistence equilibrium, not found in the previous model. While other more complicated functions of density-dependence are possible, there is not yet a clear biological basis for one form over another in the differentiation literature. The simpler linear form has the advantages of correctly mimicking infection dynamics in the absence of immune control, both in terms of an oscillatory approach to a clear carrying capacity (Hajduk & Vickerman, 1981), where slender replication fully stops and there is a balance between the two forms of the parasite, and the benefit of being more amenable to analytic treatment.

Each stumpy cell has a natural mortality δ_M. This parameter was absent in the model by Lythgoe *et al.* (2007), and as a consequence the total parasite population, due to the persistence of stumpy cells, continued to increase exponentially in the absence of immune control, while the replication of the slender cells tended to zero. We find the inclusion of stumpy cell death is very important for controlling parasite population growth and is supported by experimental evidence (McLintock *et al.*, 1993; Tyler *et al.*, 2001). There is not an explicit death rate of the slender cells in the model because it can be implicitly combined in the parameters r and K. Notice that the density-dependent differentiation process and the natural mortality of the stumpy cells together act as a form of general control of the parasite within the host, directed against all variants indiscriminately.

For the kinetics of the immune response we follow the same approach of Lythgoe *et al.* (2007). Due to lack of a detailed empirical quantitative resolution of anti-trypanosome immune responses in the host, we adopt a more phenomenological approach for modelling the immune response dynamics. Thus, we assume there is only variant-specific immunity and the $i-$th variant is removed by a specific antibody response a_i at a rate d for slender cells, and δ for stumpy cells. Given the evidence that slender cells are killed more rapidly by the immune response, we assume $d > \delta$ (McLintock *et al.*, 1993). Variant specific antibody responses grow as a result of antigen stimulation, up to a maximum of 1, starting initially from 0. The maximum rate at which the immune response can increase against any variant i is given by c, meaning that all variants are equally immunogenic as supported by observations *in vivo* (Gray, 1965). The time that it takes for a variant to stimulate the specific immune response

in the host (Tyler *et al.*, 2001), gives the delay τ in Eq. (2.3), which is typically of the order of 100 h, a value consistent with the general literature on the build-up of specific immunity (Janeway *et al.*, 2005).

The sensitivity of the specific immune response to low antigen stimulation is denoted by x, which slows down the saturation of a_i when the $i-$th parasite variant subpopulaton is below the threshold C, and accelerates it when the variant population exceeds C. This baseline formulation assumes that high and prolonged parasitaemias or antigenic diversity do not significantly impair the anti-*VSG* immune response, as observed in infections (Capbern *et al.*, 1977; Gray, 1965; Morrison *et al.*, 2005; Robinson *et al.*, 1999). Lastly, under the generality that specific memory B-cells develop in parallel with the primary immune response in trypanosome infections, we factor that specific immune responses persist, as observed empirically (Morrison *et al.*, 2005; Robinson *et al.*, 1999). This irreversibility of the immune response prohibits second or further outbreaks by the same antigen variant in the same infection, making chronicity an exclusive consequence of parasite antigenic variation. Model parameters are summarized in Table 2.1.

2.2.2 Variant emergence

Our main departure from the approach of Lythgoe *et al.* (2007) is in how we describe the emergence of new antigenic variants. While in their model, the switching was included through two additional terms in the dynamic equations: a source and a sink term for each variant, counting the contributions of all incoming switches and outgoing switches, in our formulation, switching is stochastic. When a new parasite variant arises during infection, its number is very small and may be prone to extinction. For this reason, the emergence of new variants is more appropriately modelled as a stochastic process. The stochastic component in the model consists of the random first arrival times for each variant, t_i. Under this new framework, the slender variant dynamics (Eq. 2.1) becomes:

$$\frac{dv_i}{dt} = \left[rv_i\left(1 - \frac{V+M}{K}\right) - da_iv_i \right] H(t-t_i) + D(t-t_i), \tag{2.4}$$

where $H(t-t_i)$ and $D(t-t_i)$ are the Heaviside and Dirac-delta functions respectively. $H(t-t_i)$ is zero for $t < t_i$ and 1 for $t \geq t_i$, whereas $D(t-t_i)$ is zero everywhere except

at $t = t_i$, where it takes the value of 1. The stochastic arrival times t_i are determined through a Markov process, similar to the approach of Kepler & Perelson (1995). First, we denote by s_{ij} the rate of antigenic switching from variant i to variant j. The probability of an antigenic switch per parasite division is fixed and is equal to σ (Turner & Barry, 1989). Considering that one division time equals $\ln(2)/r$, the total switch rate per unit of time is given by $r\sigma/\ln(2)$. This implies that the switch rates between antigenic variants, $S = (s_{ij})_{N\times N}$ satisfy:

$$\sum_j s_{ij} = \frac{r\sigma}{\ln(2)}, \tag{2.5}$$

placing a constraint on the row sums of the switch matrix. Then, to compute first arrival times, we consider $P_i(t)$, the probability that variant i has not yet emerged by time t. Consider a very small time interval Δt. By the Markov assumption, we have:

$$P(t+\Delta t) = P(t)\left(1 - \sum_{j\neq i} s_{ji} v_j(t)\Delta t\right). \tag{2.6}$$

Subtracting $P(t)$ from both sides and dividing by Δt we get:

$$\frac{P(t+\Delta t) - P(t)}{\Delta t} = -\sum_{j\neq i} s_{ji} v_j(t) P(t). \tag{2.7}$$

Taking the limit $\Delta t \to 0$, the dynamics of $P_i(t)$ are governed by:

$$\frac{dP_i}{dt} = -P_i \sum_{j\neq i} s_{ji} v_j, \tag{2.8}$$

collecting the switching contributions from all other variants into variant i. Note that $P_i(0) = 1$ for all variants i, except the inoculating variant. To calculate t_i, a random number is drawn from the uniform distribution $[0,1]$. At $t = t_i$, P_i reaches this random number, and $v_i(t_i) = 1$. From this point onwards v_i, m_i, a_i, so far inactive, begin the deterministic dynamics given by Eqs.2.1-2.3. Thus the switching process in this formulation matters only as a mechanism for generating previously absent antigenic variants and its contribution to parasite growth is negligible. Finally, the full model is given by Eqs.2.2-2.4 and 2.8, with parameters listed in Table 2.1. The delay differential equations are solved numerically in Matlab, using solver dde23. The initial conditions are: $v_1(0) = V_0$, $v_{i\neq 1}(0) = 0$, $m_i(0) = 0$, and $a_i(t) = 0$, for $t \in [-\tau, 0]$ and for all i. We

use the Events option of dde23 to pause and resume integration at the first arrival times of new variants. An illustration of the infection dynamics obtained with the model is given in Figure 2.2.

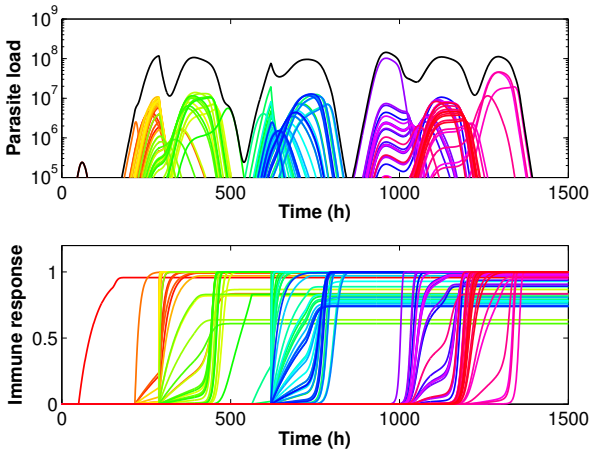


Figure 2.2: Example of full model dynamics with $N = 120$, $\eta = 12$, $x = 2$, $K = 10^8$, $C = 10^7$ and other parameters as in Table 2.1. The solid black line shows the total parasite load $V + M$, the coloured lines indicate individual variants $v_i + m_i$, coloured arbitrarily using an automatic colormap in the order 1 to N.

2.2.3 The switch matrix

Motivated by the subfamily structure of the trypanosome antigen gene archive (Marcello & Barry, 2007a; Morrison *et al.*, 2005), we assume the switch matrix, S, is characterized by a block structure (Figure 2.3). Antigenic variants are organized in blocks of η variants each, denoted by $B_1, B_2, ...$, where switching happens fast. Between-block switch rates are of lower orders of magnitudes, which represent the difficulty in accessing diversity from distant antigenic clusters. The ratio of between-block switch rates (off-diagonal blocks) and within-block switching (blocks on the diagonal) is denoted by $\varepsilon \ll 1$. A full description of the algorithm used to build the switch matrix based on 3 parameters ($\eta, N_{blocks}, \varepsilon$) is given in section 2.2.4. There are two main types of switching: 1) the *non-hierarchical* mode, which assumes only the distance between blocks determines their between-block switch rate (power law distribution);

Table 2.1: Within-host model parameters and interpretation

Parameter	Interpretation (units)	Value	Reference/Comment
r	Intrinsic growth rate of slender cells (hour^{-1})	0.1	Turner *et al.* (1995)
d	Maximal killing efficiency of slender cells by the immune response (hour^{-1})	0.5	McLintock *et al.* (1993)
δ	Maximal killing efficiency of stumpy cells by the immune response (hour^{-1})	0.1	McLintock *et al.* (1993)
c	Rate of growth of specific immune response (hour^{-1})	100	Lythgoe *et al.* (2007)
K	Within-host carrying capacity for the total parasite population	$1\times10^8 - 1\times10^{12}$	Reuner *et al.* (1997), varied
C	Threshold variant population level leading to maximal growth of specific immune response	$1\times10^8 - 1\times10^{12}$	Lythgoe *et al.* (2007), varied
x	Sensitivity of immune responses to small parasite concentration	1-3	Lythgoe *et al.* (2007), varied
τ	Delay in the stimulation of specific immunity (hours)	100	Tyler *et al.* (2001), varied
δ_M	Stumpy cell mortality rate (hour^{-1})	0.025	Tyler *et al.* (2001), Savill & Seed (2004)
σ	Switch probability per division	0.01	Turner & Barry (1989)
s_{ji}	Switch rate from j to i (hour^{-1})	varied	$\sum_j s_{ij} = \frac{\sigma r}{\ln 2}$ (Appendix 2.2.4)
N	Total number of variants	$O(10^3)$	Berriman *et al.* (2005)
η	Number of variants in one block	$1-100$	varied
B	Number of blocks in S	$B = N/\eta$	varied

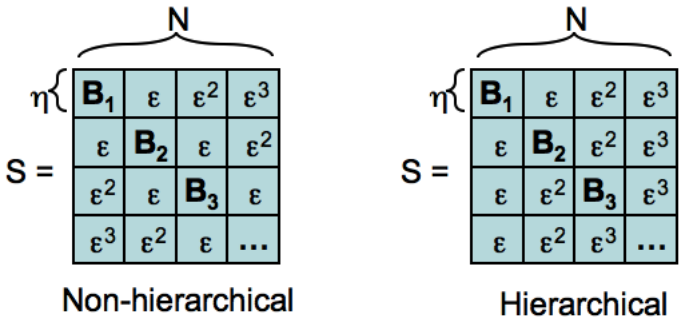


Figure 2.3: Schematic representation of two types of switch matrices S, which describe the rates of switching between parasite antigenic variants (see section 2.2.4 for full details).

and 2) the *hierarchical* mode, which assumes each block has an intrinsic activation rate governing their order of appearance, with later blocks being harder to switch to. In the simulations we mainly use the non-hierarchical switch matrix, but most results, unless otherwise stated, apply to both types. Ultimately, the real trypanosome archive may contain a combination of both hierarchical and non-hierarchical variant organization. Crucially, in the limit of very small between-block switch rates, variants of different blocks grow independently.

2.2.4 Switch matrix construction

For simplicity we assume that all antigenic blocks are of the same size. Input parameters: η, N, ε, and σ. The switch matrix S has N variants, distributed in $B = N/\eta$ blocks, each consisting of η variants. Blocks lie on the diagonal of the matrix. The off-diagonal blocks contain between-block switch rates. The non-hierarchical switch matrix is constructed as follows:

1. For each block, both on and off the diagonal, we generate $\eta \times \eta$ random uniform numbers in $[0,1]$.

2. Then we normalize them so that each row sums up to 1. We compute the within-block average switch rate for blocks on the diagonal from the constraint $q + \varepsilon q + \varepsilon^2 q + ... \varepsilon^B q = r\sigma/\ln(2)$, giving: $q = r\sigma(\varepsilon - 1)/(\ln(2)(\varepsilon^{N/\eta} - 1))$.

3. We then apply the magnitude structure according to Fig. 2.1 (non-hierarchical). The ε^n indicates that the between-block switch rate is ε^n times the average within-block switch rate q.

4. Finally S is normalized globally so that each row sums up to $r\sigma/\ln(2)$, so that S entries become rates.

The hierarchical switch matrix is constructed in an analogous manner.

2.2.5 Data to support the modelling

Longitudinal infection data with antigen variant characterization for trypanosomes are very scarce in the literature, despite extensive experimental studies in the last decades. An important early study by Capbern *et al.* (1977) has shown that more than 100 variants can be expressed sequentially over the same infection. Although the times of variant detection across different chronic infections in rabbits varied more or less continuously, it is likely that the slow build-up of the immune response masks any fine differences in intrinsic variant arrival times due to switching. A later study by Barry (1986) analyzed chronic trypanosome infections in a set of different host species, and found that the predominant group, comprising VATs which apparently developed within the first 3 weeks, varied in size according to the total number of trypanosomes in the bloodstream within that period, suggesting a clear link between parasite switch rate on one hand, and the total parasite load that a host can support on the other. If one assumes that parasite intrinsic switch rates remain constant across hosts, than any variation in appearance times should be due to the coupling of switching with growth factors in the host's bloodstream.

There has been converging evidence on the order of variant appearance within a host, ranking the telomeric intact VSG on minichromosomes as the genes most prone to be activated first, followed by subtelomeric array VSGs, and finally subtelomeric pseudogenes, which may have undergone multiple gene conversion events to create mosaics (Morrison *et al.*, 2005, 2009). Although these define just three broad VSG groupings in the archive, it is conceivable that within such groupings, finer degrees of ordering and structuring occur. For example, a mosaic generated from 2 donors

may appear earlier than a more complex mosaic, generated from 3 donors, or a mosaic formed by VSG within the same subfamily may precede mosaics generated from VSG belonging to different subfamilies (Marcello & Barry, 2007b). Similarly, intact genes depend for switching on the flanking sequence homologies with the expression site. Similar levels of such homologies could lead to subgroups of intact genes being activated at the same time and growing simultaneously as a block.

To really capture these finer degrees of structuring in switching, it is necessary to consider not only the presence (yes/no) of a particular variant at a given point in time, but its abundance, relative to the total parasite population. Unfortunately, such comprehensive data and analysis is still lacking, with majority of studies having focused on single relapses. More recently Hall et al. (pers. comm.) have conducted a series of antigenic variation experiments on mice, following infections for more than 30 days, where there is direct evidence for varying levels of antigen diversity over an infection with a limited number of variants in each parasitaemia peak, and for mosaic gene expression already within the first few relapses.

These findings, taken together suggest that there may be defined genetic factors governing switch rates of different VSG, despite the intrinsic stochasticity in the switching process. Considering thus a block structure for the antigen switch matrix, where switches within variants of the same block/group, facilitated by similar genetic features (identity, number of repeats, switching mechanism, genome location etc.), are more frequent than across blocks seems a natural step for exploring the antigenic variation potential of the parasite and understanding its function in terms of infection characteristics sensitive to it. Moreover, a block structure of the antigen repertoire is likely to be relevant for other pathogens, including bacteria such as Anaplasma (Foley *et al.*, 2009; Futse *et al.*, 2008), or other protozoa, a prominent example being the malaria parasite (Bull *et al.*, 2008; Recker *et al.*, 2011).

2.3 Model behaviour

Once a new antigenic block of variants arises stochastically, its dynamics follow deterministic dynamics, making it easy to calculate the asymptotic steady states for one variant. To summarize the main results of the asymptotic analysis, there are two types of steady states for each variant in a block, namely, a continuum of parasite clearance

states: $v_i^* = 0, \quad m_i^* = 0, \quad 0 \leq a_i^* \leq 1$, which are stable for $a_i^* > r/d$, and an infection persistence state: $v_i^* = \frac{K(r-d)(\delta+\delta_M)}{\eta r(r-d+\delta+\delta_M)}, \quad m_i^* = \frac{K(d-r)^2}{\eta r(r-d+\delta+\delta_M)}, \quad a_i^* = 1$, which exists only for $r > d$ and is always stable, as shown below.

2.3.1 Asymptotic behaviour

In one block variants undergo identical dynamics, thus it is sufficient to analyze the asymptotic behaviour of a single variant in a block, e.g. variant i, and the result applies to all other variants $j = 1, ..., \eta$. Besides the trivial steady-state $(v_i^*, m_i^*, a_i^*) = (0,0,0)$, which is unstable, system (1) has: a continuum of semi-trivial steady states where infection is cleared: $v_i^* = 0, m_i^* = 0, a_i^* \neq 0, \quad 0 < a_i^* \leq 1$, and a disease persistence steady-state: $v_i^* = \frac{K(r-d)(\delta+\delta_M)}{\eta r(r-d+\delta+\delta_M)}, m_i^* = \frac{K(d-r)^2}{\eta r(r-d+\delta+\delta_M)}, a_i^* = 1$, which is biologically realistic only for $r > d$. The Jacobian Matrix of the linearized system evaluated at (v_i^*, m_i^*, a_i^*) is

$$J_0 = \begin{pmatrix} -a_i^* d - \frac{\eta r v_i^*}{K} + r(1 - \frac{\eta(v_i^*+m_i^*)}{K}) & -\frac{\eta r v_i^*}{K} & -d v_i^* \\ \frac{r v_i^*}{K} + \frac{r(v_i^*+m_i^*)}{K} & \frac{r v_i^*}{K} - \delta a_i^* - \delta_M & -\delta m_i^* \\ 0 & 0 & -c(\frac{v_i^*(t-\tau)+m_i^*(t-\tau)}{C})^x \end{pmatrix}, \tag{2.9}$$

and the matrix associated with the delay is

$$J_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ x\frac{c(1-a_i^*)}{C}(\frac{v_i^*+m_i^*}{C})^{x-1} & x\frac{c(1-a_i^*)}{C}(\frac{v_i^*+m_i^*}{C})^{x-1} & 0 \end{pmatrix}. \tag{2.10}$$

Defining the matrix Δ by $\Delta = \lambda I - J_0|_{(v_i^*, m_i^*, a_i^*)} - e^{-\lambda\tau} J_1|_{(v_i^*, m_i^*, a_i^*)}$, the characteristic equation determining stability is given by $\det(\Delta(\lambda)) = 0$.
At the infection–persistence steady-state $(v_i^*, m_i^*, 1)$ the matrix J_1 vanishes, and we have the eigenvalues:

$$\begin{aligned} \lambda_1 &= -c(\frac{v_i^* + m_i^*}{C})^x, \\ \lambda_{2,3} &= \frac{1}{2}\left(-\delta - \delta_M \pm \sqrt{(\delta+\delta_M)^2 - 4(r-d)(\delta+\delta_M)}\right). \end{aligned} \tag{2.11}$$

All of these eigenvalues have negative real part, and this implies that when the steady-state is biologically realistic, it is always stable.

Concerning the infection clearance steady state, in the case of $x = 1$, the matrix Δ at steady state $(0,0,a_i^*)$ is:

$$\Delta = \begin{pmatrix} \lambda + a_i^* d - r & 0 & 0 \\ 0 & \lambda + \delta a_i * + \delta_M & 0 \\ \frac{c}{C}(1-a_i^*)e^{-\lambda\tau} & \frac{c}{C}(1-a_i^*)e^{-\lambda\tau} & \lambda \end{pmatrix}, \tag{2.12}$$

whose determinant has solutions independent of the delay τ: $\lambda_1 = r - a_i^* d$, $\lambda_2 = -\delta_M - \delta a_i^*$ and $\lambda_3 = 0$. Notice that if $a_i^* < r/d$ the infection clearance steady state is unstable. If $a_i^* \geq r/d$, λ_1 and λ_2 are both negative, or $\lambda_1 = 0$ and $\lambda_2 < 0$. Given that we find at least one zero eigenvalue (λ_3), the stability of the infection clearance steady state has to be determined via the centre manifold.

Similarly, when $x > 1$, J_1 vanishes at $(0,0,a_i^*)$. We find that the disease-clearance steady-states $(0,0,a_i^*)$ are degenerate, forming a line of steady states. Linearization near such a steady state has again the eigenvalues: $\lambda_1 = r - d a_i^*$, $\lambda_2 = -\delta_M - \delta a_i^*$ and $\lambda_3 = 0$. As long as $a_i^* < r/d$, the corresponding steady-state is unstable. If $a_i^* > r/d$, both λ_1 and λ_2 are negative and calculation of the centre manifold is necessary to determine stability.

Center manifold reduction

To calculate the centre manifold, we linearise the system around the disease-clearance steady state $(0,0,a^*)$. We drop the index i for simplicity. Thus, we can write $a(t) = w(t) + a^*$. The system corresponding to a single variant can then be rewritten as:

$$\frac{dv}{dt} = rv(1 - \eta\frac{v+m}{K}) - d(w+a^*)v, \tag{2.13}$$

$$\frac{dm}{dt} = rv\eta\frac{v+m}{K} - \delta_M m - \delta(w+a^*)m, \tag{2.14}$$

$$\frac{dw}{dt} = c(1-w-a^*)(\frac{v+m}{C})^x, \tag{2.15}$$

whose $(0,0,0)$ steady state corresponds to the $(0,0,a^*)$ steady state of the original system. We look for a solution of the type

$$w = h(v,m) = a_1 v + a_2 m + b_1 v^2 + b_2 m^2 + c_1 vm + H.O.T. \tag{2.16}$$

Then we have:

$$\frac{dw}{dt} = \frac{dh}{dv}\frac{dv}{dt} + \frac{dh}{dm}\frac{dm}{dt}, \tag{2.17}$$

and $dh/dv = a_1 + 2b_1v + c_1m$, and similarly for dh/dm. In this way, we have transformed the original differential equation for w (Eq. 2.15) into a polynomial in v and m. The degree of this polynomial will depend on x. Notice that for the centre manifold calculation, we need x to be an integer, i.e. $x = 1, 2, 3...$ Furthermore, depending on the value of x, we need to adjust the centre manifold expansion. Usually, in $h(v,m)$ we must retain terms up to order $x+1$.

Then we substitute the expression $w = h(v,m)$ into the three differential equations for v, m, and w. Thus we obtain another polynomial expression for dw/dt which has to equate to the expression obtained previously in Eq.2.17. By equating powers of v and m across the two equations, we can get the precise values of the coefficients in the center manifold $h(v,m)$. After $h(v,m)$ is made explicit, we consider the reduced system $dv/dt, dm/dt$, which consists of polynomials in v and m only. The stability of the $(0,0)$ state in the reduced system is the same as the stability of the $(0,0,0)$ state in the 3-dimensional system, and consequently as the $(0,0,a^*)$ state in the original system. The two eigenvalues of the reduced system are: $\lambda_1 = -\delta_M - \delta a^*$, and $\lambda_2 = -da^* + r$. We find that for any value of x, if $a^* > r/d$, both eigenvalues are negative and thus the center manifold calculation gives that the infection clearance equilibrium is stable. If $a^* = r/d$, one of the eigenvalues of the Jacobian of the reduced system, evaluated at the $(0,0)$ steady state, namely, λ_2 is again zero, thus the stability of this steady state has to be studied in the context of a further reduced 1-dimensional system. Substituting $a^* = r/d$, and expressing $m = g(v) = \alpha_0 + \alpha_1 v + \alpha_2 v^2 + H.O.T.$, after some calculations, we get that the dynamics of the reduced system is given by $dv/dt = -r\eta/K \times v^2 - dr^2\eta^2/K(d\delta_M + \delta Kr) \times v^3 + O(v^4)$. For this last equation, $v = 0$ is stable. We conclude that $(0,0,a^*)$ is asymptotically stable when $a^* \geq r/d$.

2.3.2 Transient dynamics

Since we always assume $r < d$, the only steady state seen in the model simulations is the parasite clearance state for each variant. Thus the long term asymptotic behaviour of the system is infection clearance. However, what is interesting is the time it takes for clearance to be established and the features of the infection profiles before clearance. These are part of the transient dynamics of host-parasite interaction, whose main characteristics we briefly summarize below:

i) Within-host dynamics are characterized by oscillating total parasite load (Figure 2.2), where each peak is composed of different antigenic variants (Barry & McCulloch, 2001). Although infection may be initiated by a single variant, stochastic antigenic variation quickly gives rise to new variants thus prolonging the infection;

ii) Consistent with empirical observations (Miller & Turner, 1981; Robinson *et al.*, 1999), an infection peak can be composed of one or more variants, depending on the switch matrix;

iii) Independently of the switch matrix, the dynamics of each variant are dominated by slender cells in the growth phase and then by stumpy cells at the peak and throughout the decline phase (Figure 2.4);

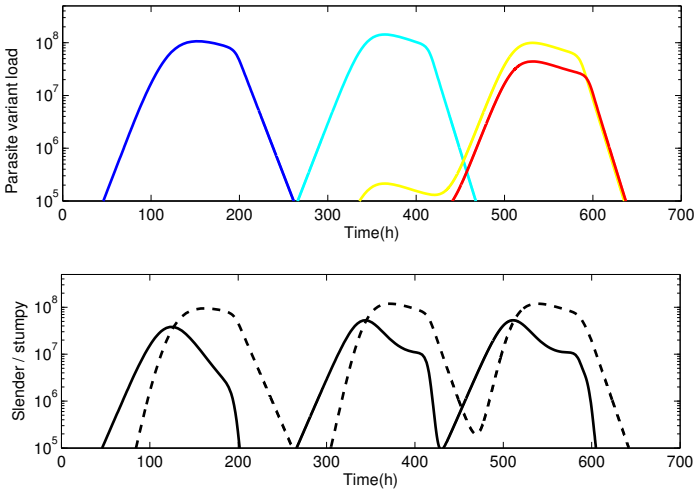


Figure 2.4: Antigenic variation dynamics and the consequences of differentiation. The top panel shows an example of typical dynamics of individual variants during an infection, given in different colors. The bottom panel shows the composition of the total parasitaemia in terms of slender and stumpy cells. As the total parasite population grows, slender cell dominance (solid black line) is gradually replaced by dominance of stumpy cells (dotted line), crucial for parasite transmission to the vector.

iv) Numerical simulations confirm that variant first arrival times highly correlate with variant mean activation rates from the switch matrix. The higher the variant activation rate, the earlier a particular variant appears during infection, as reported also

in (Marcello & Barry, 2007a). Early variants are also associated with a smaller variability in first arrival times across stochastic simulations (Figure 2.11), consistent with empirical findings (Marcello & Barry, 2007a; Morrison *et al.*, 2005). Since the switch rates within a block are larger than between blocks, all the variants in a block emerge at similar times.

v) When the switch matrix follows the non-hierarchical mode (Figure 2.3) consecutive parasite peaks have generally equal spacing, and the order of variant appearance is determined by the position of the inoculating variant in the matrix and the sequence of its neighbouring blocks. Whereas if the switch matrix is hierarchical, consecutive parasite peaks occur at increasingly greater temporal distances and the order of variant appearance is independent of the inoculating variant.

The behaviour of the model is complex, resulting from the interplay of many parameters. The results we present here are robust even when dropping the perfect symmetry across variants in the model, i.e. allowing variability within the same order of magnitude, in growth rates, rates of differentiation, immunogenicity and delay in immune stimulation. These changes do not change the qualitative and, to a large extent, quantitative behaviour of the model. To understand the mechanisms behind trypanosome within-host dynamics, it is helpful to begin by first examining what governs a single infection peak and then what drives the full infection profile within a host.

2.4 Single block dynamics

The variants in an infection peak, usually correspond to variants of the same block in the switch matrix S, and we notice that their first arrival times are clustered (see Figure 2.2). Thus, to study a single infection peak we can study a single block of variants, by assuming the variants of a block emerge around the same time: $t_i \approx t_j$, for each variant i and j within a block.

2.4.1 Block size, $\eta = 1$

The simplest case is to assume a block is composed of only one variant. The dynamics of a single variant can be broken down into 3 phases: I) *the growth*, II) *non-growth* and III) *decline phase*, as illustrated in Figure 2.5. These 3 phases also appear in the

full infection profile. In Phase I the variant grows at an exponential rate, mediated by fast proliferation of slender cells, v_i. As the total parasite population increases, cells begin to transform to the stumpy form, m_i, in a density-dependent manner, thus the effective growth rate of the parasite population is reduced. However, specific antibody responses continue to increase as $v_i + m_i > 0$. Once $V + M \approx K$, we enter Phase II, where $V + M \approx K$. Throughout Phase II, the total parasite load is maintained at a balance due to general control by parasite differentiation and the natural mortality of the stumpy cells. Phase II lasts until $a_i = r/d$, for any i, signalling the start of specific control of the parasite mediated by host immune responses. Asymptotic analysis of the model reveals that a necessary requirement for the stability of the infection clearance equilibrium is $a_i > r/d$. If $a_i = r/d$ before $V + M \approx K$, then Phase II is not initiated. In fact, as soon as a_i surpasses this threshold, Phase III of the dynamics begins, leading to the decline of each variant.

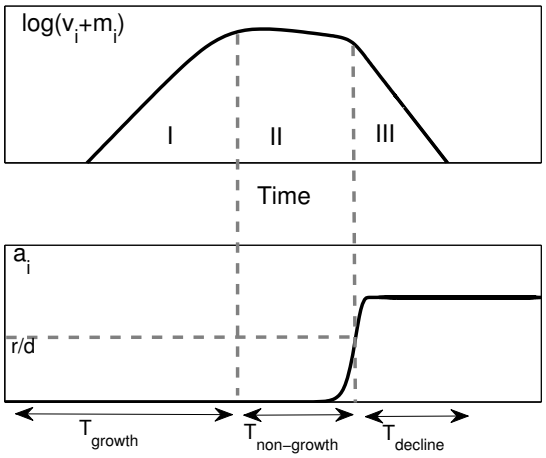


Figure 2.5: Schematic of individual variant infection dynamics illustrating the specific immune response (bottom) and variant growth (top) with three phases: I) growth, II) non–growth and III) decline. The total duration of these three phases defines a block wave.

Notice that in general Phase I and III always happen because there is no prior immunity against any variant, permitting growth, and hence the build-up of the immune response due to antigen stimulation, which leads to eventual decline. Phase II instead is more complex and it is key in determining the dynamics. Depending on whether Phase

II is long or short, a variant may take longer or shorter to be cleared. The duration of Phase II is inextricably linked to the strength of specific immunity.

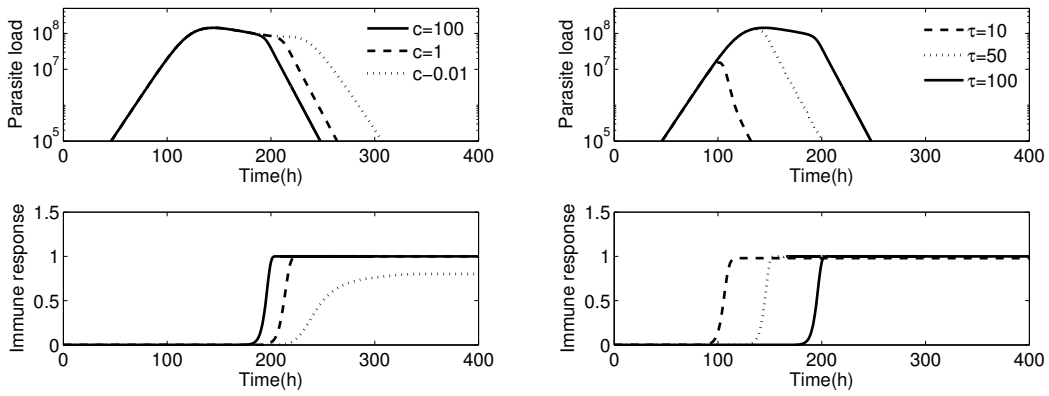
There are many parameters controlling the activation of specific immunity in our model, each of which has a subtle effect on a particular aspect of immune kinetics (Figure 2.6). The growth rate c for example, controls only the beginning of the decline (Phase III) in the parasite population: the larger c is, the longer the duration of Phase II and viceversa, but c has no effect on the magnitude of the peak or the slope of the decline.

The immune response delay τ, on the other hand can influence the magnitude of the peak, especially when it is small. Because it represents the level of decoupling between parasite numbers and immune response growth, if τ is small, the negative feedback exerted by immune control on the parasite population is initiated earlier and the ultimate peak reached by the parasite population is low. This implies the absence of Phase II and a short decline phase. However, as τ increases, it starts to control only the duration of Phase II in the dynamics of the parasite variant: τ large implying long persistence of Phase II, τ small, instead, implying shorter non-growth phase and fast decline.

In contrast, the parameter x, which controls the responsiveness of antibody production to parasite numbers relative to the threshold C has a much bigger encompassing effect than τ or c. It modulates the nonlinear relationship that exists between parasite numbers and the build-up of host immunity. A large x slows down further antibody growth when the antigen stimulation is below C and accelerates it when the antigen stimulation is above C, leading to a sharp increase in a_i. For this reason, unlike c and τ, the parameter x can also influence the slope of the growth phase, in addition to the magnitude of the peak and the duration of Phase II.

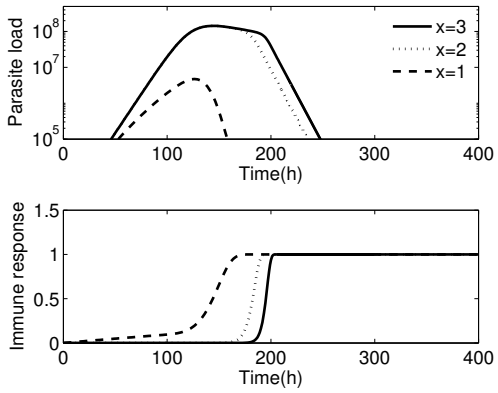
Finally, the duration of Phase III is controlled by the immune response killing rates d and δ, in particular δ, since it denotes the killing rate of the stumpy cells, which is usually lower than that of slender cells. When this killing rate is reduced, the decline of the parasite population is very slow, however there is no effect on the growth phase or the peak parasite load.

In general, the weaker specific immunity is during Phase I (e.g. large τ, x, small c), the more the parasite can grow, and the longer it takes for a_i to reach r/d, hence the longer Phase II. In contrast, the stronger specific immunity, the less the parasite can

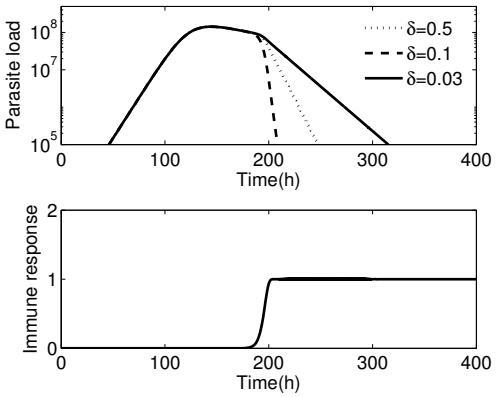


(a) The effect of c

(b) The effect of τ



(c) The effect of x



(d) The effect of δ

Figure 2.6: Illustration of the effects of different immune response parameters where parasite load corresponds to $V+M$ in the model. $N=1, C=K=10^8$, and all other parameters are as in Table 2.1.

grow and the faster it is cleared, shortening Phase II or removing it altogether.

2.4.2 Block size, $\eta > 1$

We consider now larger blocks of variants, allowing for more than one variant to emerge at the same time. This is a more realistic scenario as empirical studies show that one infection peak is generally composed of many variants (Morrison *et al.*, 2005). We assume there is an initial parasite load V_0, which is divided equally among the η variants of the block, and there is no prior host exposure to any of the antigens.

The same way that the dynamics of each variant are composed of 3 phases, the dynamics of a block is composed of the *growth*, *non-growth* and *decline* phase. But now, the number of variants growing together matters. Ultimately, specific immune stimulation depends on the availability of specific antigen, hence on the magnitude of $v_i + m_i$. This goes down as the number of variants, η, sharing the carrying capacity K within the host increases. Model simulations show that as η increases, the duration of Phase II of the dynamics tends to infinity, because the numbers of any individual variant are low, approximately K/η, resulting in weak immune stimulation and so a_i struggles to reach r/d (see Figure 2.7). It is then natural to ask: what is the critical value of η separating these distinct dynamical regimes, of rapid antigenic clearance and long term antigenic persistence?

2.4.3 The block size threshold (η_{crit})

To distinguish between fast and slow infection clearance, we analyze the duration of Phase II, $T_{non-growth}$, via quasi-steady state arguments. As illustrated in Figure 2.6(d), it is reasonable to assume that specific immune responses initially change on a slower time scale than slender and stumpy cells, so that the effect of immunity–mediated clearance during the growth phase (Phase I) can be neglected. Let T_{growth} denote the time it takes for the parasite population to reach the nontrivial quasi–steady state given by solving Eqs.2.1 and 2.2 with $a_i = 0$.

As a result of the symmetry between variants, $v_i + m_i = K/\eta$ at $t = T_{growth}$. We assume the parasite population stays at quasi-steady state during $T_{non-growth}$ because specific immunity has not begun to act yet. As a consequence of the delay τ in the

stimulation of immune responses, a_i starts to change at time $t = T_{growth} + \tau$, obeying the following equation:

$$\frac{da_i}{dt} \approx c(1-a_i)\left(\frac{K}{\eta C}\right)^x, \tag{2.18}$$

which implies

$$0 < a_i(t) \leq 1 - \exp\left[-ct(K/\eta C)^x\right], \tag{2.19}$$

for $t \geq T_{growth} + \tau$. This is the maximal rate of change for a_i, thus the time ($T_{r/d}$) it would take for a_i to reach the required r/d threshold for the initiation of infection clearance, can be approximated by:

$$a_i(T_{r/d}) \approx 1 - \exp\left[-cT_{r/d}(K/\eta C)^x\right] = r/d, \tag{2.20}$$

giving

$$T_{r/d} \approx -(K/\eta C)^{-x}\ln(1-r/d)/c. \tag{2.21}$$

Notice that Phase II, where parasite load is at its peak, lasts until the immune response reaches the level r/d, thus we have: $T_{non-growth} \approx \tau + T_{r/d}$. We define phase II as being long if $T_{non-growth} = \tau + T_{r/d} > 2\tau$, and short if $\tau + T_{r/d} \leq 2\tau$. In the first case, parasite decline (phase III) takes longer to begin.

The choice of the interval 2τ to divide these regimes may seem arbitrary at first, but some justification for it comes from our observation that once in phase II, 2τ is the longest interval during which the stimulation of a_i by the particular variant has the magnitude K/η. Thus, within the quasi-steady state framework, we take this interval as a useful reference point to define $T_{non-growth} < 2\tau$ as fast clearance and $T_{non-growth} > 2\tau$ as slow clearance.

Assuming all other parameters are fixed, the value of η for which $T_{r/d} = \tau$ yields the critical number of variants η_{crit} dividing the two regimes of fast clearance and long persistence. Simple algebra reveals that

$$\eta_{crit} = \frac{K}{C}\left(\frac{-\ln(1-r/d)}{c\tau}\right)^{-1/x}, \tag{2.22}$$

such that when the block size is relatively small, i.e. $\eta < \eta_{crit}$ specific immunity rapidly clears all variants of that block, whereas for $\eta \geq \eta_{crit}$ differentiation is the main controlling force and so Phase II is very long.

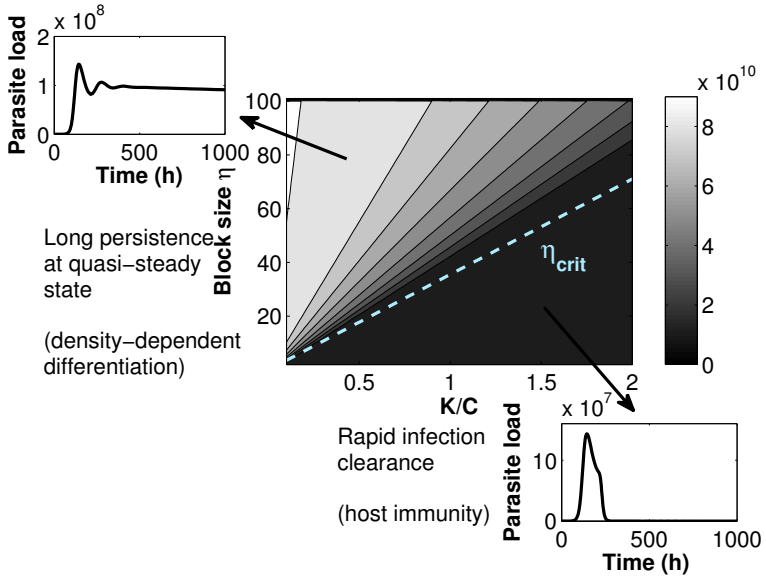


Figure 2.7: Contour plot of total parasite load $\int_0^{1000} V(t) + M(t)dt$ as a function of block size η and K/C. The lighter areas indicate a higher value for the area under the curve $V + M$, whereas the darker areas indicate a smaller value. The analytical approximation to η_{crit} by our quasi-steady state approximation (dashed line) is in good agreement with the model numerical simulation results. Parameter combinations where $\eta < \eta_{crit}$ (dark regions) lead to rapid variant clearance and generally lower peak parasite load. Parameter combinations where $\eta > \eta_{crit}$ (lighter regions) lead to extended persistence of variants, and peak parasite load close to carrying capacity. The insets show i) rapid clearance ($\eta = 20, C = 2K/3$) and ii) long persistence ($\eta = 80, C = 2K$). Parameters as in Table 2.1 and $K = 10^8$, $x = 3$.

When Phase II is long, individual variants of a block are restricted to low levels, insufficient to optimally stimulate specific immune responses. This allows the variants to persist at low densities, until sufficient host immunity is mounted. Similar thresholds have been found previously in antigenic variation models (Nowak *et al.*, 1990; Sasaki, 1994). In contrast to those results, our result refers to a quasi–steady state, and applies only to variants whose first arrival times are the same. For variants arising in the host at different times, the same threshold does not hold. Sufficient decoupling may allow them to grow independently and stimulate specific immunity.

We observe that η_{crit} depends linearly on K/C, the ratio between the within-host carrying capacity and the immune response threshold, expressing the competition be-

tween density-dependent differentiation and host immunity (Figure 2.7). When K/C increases, η_{crit} increases, meaning that the immune stimulation to individual variants is stronger, hence there is sufficient specific control to clear larger blocks of variants. When K/C decreases, η_{crit} is reduced, implying dominance of general control, hence even small blocks can easily persist without clearance.

The significance of this threshold for within-host dynamics points toward the interaction between specific and general control of the parasite. This interplay occurs in many pathogen infections besides trypanosomes. In trypanosomes, in particular, properties of their antigenic archive, such as the size of an antigenic block, appear crucial and may tip the balance toward one mechanism of parasite control or the other. We notice how the ultimate duration of Phases I+ II+III, defining a 'block wave', depends on the block size η. This affects the total parasite load, $\int(V+M)dt$, contained in a block wave (Figure 2.7), which may have implications for transmission.

Considering that differentiation seems to play such an important role in within-host dynamics, a natural question is: how sensitive is our result to the choice of the differentiation function? In our model we used a linear differentiation rate $f(V+M) = rv_i(V+M)/K$. In general, the η_{crit} phenomenon holds whenever the differentiation rate $f(V+M)$ leads to an apparent carrying capacity within the host (stable nontrivial equilibrium when $a_i = 0$), even when the specific immune response is saturating with respect to $v_i + m_i$ (e.g. Holling type II). In contrast, if $f(V+M) \to 0$ as $V+M \to \infty$, the notion of η_{crit} is no longer valid, as the total parasite population continues to grow exponentially, albeit at a reduced rate. In particular, when host immunity is non-saturating with respect to $v_i + m_i$, control of the parasite is always possible if antigen stimulation keeps increasing. In the absence of density-dependence in the differentiation function, for example if each slender cell becomes stumpy at a fixed rate, the two parasite forms grow exponentially in the host when immune control is absent, and are always cleared rapidly whenever $a_i > 0$, so again no η_{crit} phenomenon arises.

Given the clear importance of antigenic block size, in the following, we explore the sensitivity of other infection characteristics such as the peak parasite number, the size of variant subpeaks, and slender/stumpy ratio to η.

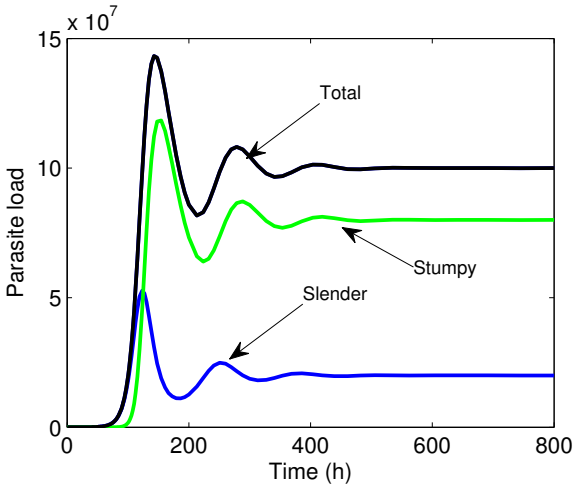


Figure 2.8: Parasite dynamics in the differentiation-only case. The total parasite population, $V+M$, settles at the resulting carrying capacity K. The ratio V^*/M^* is given by δ_M/r.

2.4.4 Why does the block size (η) matter?

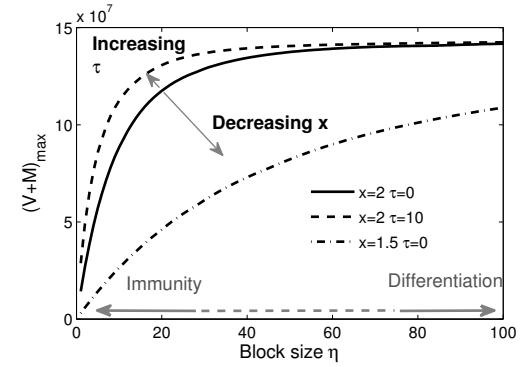
Our model captures a continuum of dynamic scenarios from the limit of C very large relative to K, to the limit of K very large relative to C. In the former case, differentiation dominates, determining the peak at carrying capacity and extending Phase II. In the latter case, specific host immunity dominates, and all variants of the block are quickly controlled by the action of specific antibodies, yielding a short Phase II.

Let us consider the first limit: C tends to infinity. We now have only general control through density-dependent differentiation. This implies that the variants in the block are completely coupled in their dynamics (see Appendix A.1 for the mathematical analysis). In this case, the peak parasite load $\max_t[V(t)+M(t)] = (V+M)_{max}$ is determined by the within-host carrying capacity K, independent of block size. Conversely, individual variant peaks, $(v_i+m_i)_{max}$ decrease linearly with η, because variants now share the carrying capacity. The ratio of slender-to-stumpy numbers initially favours the slender forms during Phase I, and the ratio gradually tends towards δ_M/r during Phase II (see Figure 2.8) as the dynamics gradually approach the non-trivial steady state. Because $\delta_M < r$, differentiation-only control favours prevalence of the stumpy transmission forms of the parasite.

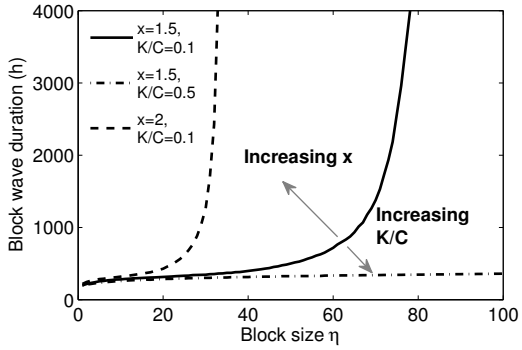
Now consider the second limit: K tends to infinity. In this case, only slender cells are present ($m_i(t) = 0$, for all t), because density-dependence, mediating stumpy cell production, is absent. So, parasite numbers are only controlled by specific immunity, and the dynamics of each variant are decoupled from each other. Assuming for a moment the delay $\tau = 0$ and $x = 1$, we focus on the case where the sensitivity of the immune response to parasite numbers is at its strongest. As a consequence, the total peak V_{max} increases with the number of variants present in the block, while the size of individual variant peaks $(v_i)_{max}$ and block wave duration remain unaffected (see Appendix A.2 for the analysis).

Finally, the case of moderate C and K lies in between the two scenarios discussed so far. By varying η, the dynamics approaches smoothly one extreme or the other (Figure 2.9). At the start of the infection $a_i(0) = 0$, the dynamics of stumpy cells initially depends entirely on that of slender cells, with $dm_i/dt \approx r\eta v_i^2/K$. When $r\eta/K$ is sufficiently small, i.e. $r\eta/K \ll c/C$, then the number of stumpy cells is also small. This implies that the slender cell– immune system interaction governs the parasite population dynamics, giving predominance of slender forms, $(v_i)_{max} \gg (m_i)_{max}$, and hence $V_{max} \gg M_{max}$. Due to the weakness of general control via differentiation, the coupling between variants is weak, and V_{max}, M_{max} and $(V+M)_{max}$, all increase with the block size η, albeit the absolute magnitude of parasite load and block wave duration is low. When $r\eta/K \gg c/C$, differentiation becomes stronger and this results in higher stumpy cell production. As η increases, the numbers of stumpy cells catch up with the slenders and even reach a higher peak. Thus, the slender-to-stumpy ratio V_{max}/M_{max} decreases with η tending towards the value determined by differentiation dominance (Figure 2.9(c)).

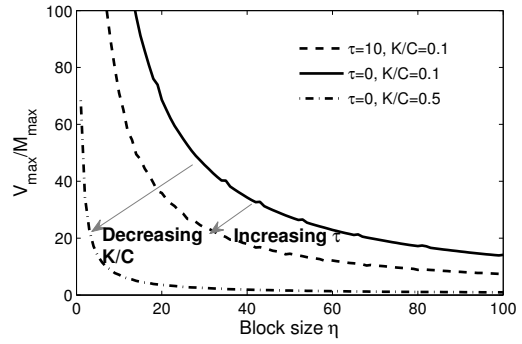
In general, the relative dominance of specific over general parasite control can be counterbalanced by changes in opposing parameters. Block wave duration, for example, increases as a function of η faster when x is large, but more slowly when the ratio K/C is large (Figure 2.9(b)). When returning to cases where $\tau > 0$ and $x \geq 1$, numerical simulations reveal that differentiation dominance occurs at even smaller block sizes than the $\tau = 0$ and $x = 1$ case (see Figure 2.9(a)). A large delay τ or large x favours the decoupling between current levels of parasite load and immune-response. This decoupling generally results in the individual variant dynamics being controlled via density-dependent differentiation.



(a) Peak parasite load



(b) Block wave duration



(c) Slender/stumpy ratio

Figure 2.9: Infection characteristics as a function of η. **a)** Peak parasite load increases with η when host immunity dominates in parasite control. **b)** Block wave duration increases with η when differentiation is dominant for large block sizes. **c)** Slender-to-stumpy ratio decreases with η towards the value r/δ_m mediated by differentiation dominance. The dashed and dash-dotted lines illustrate that the effects of differentiation are accelerated by larger immune delay τ and small K/C, whereas immunity dominance is favoured by small x and small τ. Parameter values are: $K = 10^8, C = 10^9$ (a); $\tau = 0, K = 10^8$ (b); $x = 1.7, K = 10^8$ (c). All other parameters are as in Table 2.1. Duration is calculated as the time it takes for $V + M$ to fall below its initial value $V_0 = 10^3$.

In summary, in immunity-dominant scenarios ($r\eta/K \ll c/C$) the peak parasite load increases with η, whereas in differentiation-dominant scenarios ($r\eta/K \gg c/C$), block wave duration increases with η. These two infection characteristics at the scale of single peaks: peak parasite load and 'block wave' duration, together with stumpy-to-slender dominance have clearly an important bearing on parasite transmission, with consequences for the selective pressure on the block size η.

2.5 Multiple-block dynamics $N = B\eta$

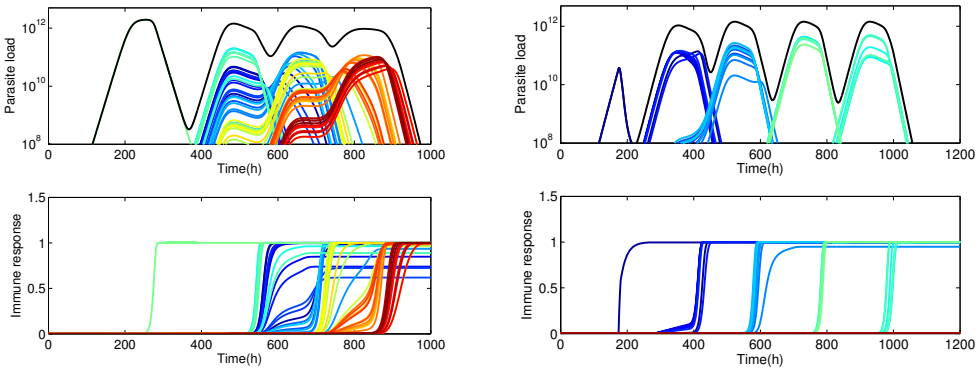
So far, we have only focused on a subset of the antigenic archive, namely variants that emerge at the same time within the host, neglecting their switching. In the following, we analyze stochastic switching between blocks leading to the dynamics of the full model (Eqs.2.2-2.4,2.8), where new variants arise at different times. In order to obtain infection dynamics exhibiting multiple peaks, the switch matrix must be composed of many blocks, i.e. $B > 1$, each block in isolation must be relatively small, i.e. $\eta < \eta_{crit}$, and the overlap between consecutive block waves in an infection must also be small, i.e. between-block switching must be small ($\varepsilon \ll 1$). When the between-block switching is high, the separation between consecutive peaks is low, whereas, if switching between blocks happens at higher rates, the consecutive block waves appear more separated over an infection (Figure 2.10).

2.5.1 The critical variant activation rate

A unique feature of the stochastic model, also observed in experiments, is that some variants never arise during infection. To understand this phenomenon, we go back to the entries of the switch matrix, S. We define the mean activation rate of a new variant i in the system as:

$$\bar{s_i} = \frac{1}{\eta} \sum_{j \in CurrentBlock} s_{ji}, \tag{2.23}$$

under the assumption that the major contribution in switching comes from the block of variants currently growing in the host. Through $\bar{s_i}$ the variants of any new block can be ranked, from high activation rate ($\bar{s_i}$ large) to low activation rate variants ($\bar{s_i}$ small). Besides the higher-level organization of variants in units of blocks, there is thus another



(a) Low separation (b) High separation

Figure 2.10: Between-block connectivity controls the separation between antigen block waves and archive turnover rate. a) $\varepsilon = 0.1$. The parasite load develops as a series of highly overlapping block waves. b) $\varepsilon = 0.001$. Antigenic variation proceeds more slowly and subsequent blocks of variants appear more separated from each other.

layer of hierarchy in terms of mean activation rates, which fine-tunes sequential variant appearance within a block. Notice from the construction of the switch matrix that these mean activation rates all belong to the same order of magnitude εq, where q is the within-block average switch rate and $\varepsilon \ll 1$ (see Section 2.2.4).

Recall that $P_i(t)$ is the probability that variant i has not yet emerged in the host by time t. Substituting Eq. (2.23) into the original equation for P_i (Eq. 2.8), we get the simpler equation

$$\frac{dP_i}{dt} = -P_i \bar{s}_i V, \tag{2.24}$$

where $V(t)$ denotes the sum of all slender cell populations from the current block of variants. Solving Eq. (2.24) gives:

$$P_i(t) = \exp\left(-\bar{s}_i \int_0^t V(s)ds\right). \tag{2.25}$$

In order to better understand stochastic variant generation, we simplify the analysis by assuming the same stochastic generation threshold for each variant, following the deterministic approximation approach used by Kepler & Perelson (1995). If a variant is never generated, this means that the probability $P_i(t)$ never reaches the required

generation threshold. The explanation lies in the interaction between variant mean activation rate $\bar{s_i}$ and the total slender cell number within a block $\int_0^t V(s)ds$. Notice that when a block contains a few variants, the integral $\int_0^t V(s)ds$ is bounded, as all phases of the dynamics are short. This leads to a critical lower bound for $\bar{s_i}$ as shown in the following.

Mathematical derivation

Here we estimate analytically the critical threshold for the switch rate between antigenic blocks, which determines whether a subsequent block will be generated or not. Adopting the deterministic threshold approximation of Kepler & Perelson (1995), we assume the same generation threshold $1/e$ for all new variants. Stochastic arrival times t_i (see Figure 2.11) are then replaced by discrete arrival times T_i. When $P_i(t)$ crosses the given threshold, $t = T_i$ and variant i is generated. The probability that variant i has not yet been generated by time t is rewritten as $P_i(t) = \exp\left(-\bar{s_i}\int_0^t V(s)ds\right)$ (see Eq. 2.23).

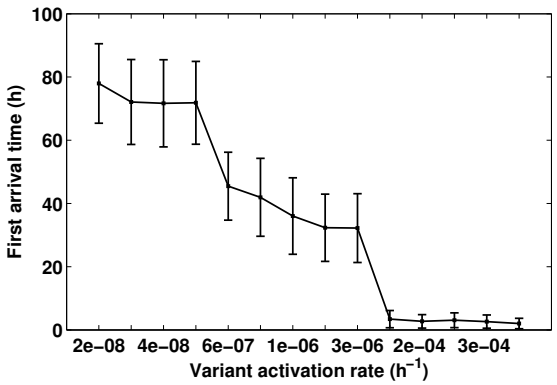


Figure 2.11: Illustration of first arrival times of 14 variants, as a function of variant activation rate. Mean values ± s.d. are given for 500 simulations of the hybrid model for 800 h. The parameters used are as default. In particular: $C = K = 10^8, x = 3, \tau = 100, N = 15, \eta = 5, v_1(0) = 10^3$. Notice that all variants of this particular archive arise within the first 100 hours of infection. Variants of the same block arise around the same time, and high activation rate variants show less variability in arrival times within the host than low activation rate variants.

Considering a moderate block size $\eta < \eta_{crit}$, such that $V = \eta v$, we proceed by

bounding the integral $\int_0^t V(s)ds$ from above. Because $V(t) \geq 0$, for all t, we have $\int_0^t V(s)ds < \int_0^\infty V(s)ds$. It is easy to see that $V_{max} < V_{max} + M < K$. So V_{max} never reaches K exactly. Thus, we can bound the growth phase of the slender population by an exponential growth. Denote by τ_K the time it would take for the slender population V to reach K growing exponentially at rate r from an initial population V_0.

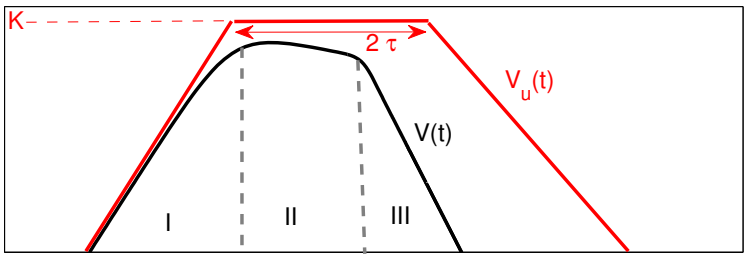


Figure 2.12: Illustration of the upper bound for $V(t)$ in deriving the critical switch rate threshold.

Next, recall that we defined $T_{non-growth} \leq 2\tau$, for $\eta < \eta_{crit}$ (see Section 2.4.3). We can thus bound the non-growth phase of the slender population by a persistence of V at carrying capacity starting from $t = \tau_K$ until $t = \tau_K + 2\tau$. Finally we can bound from above the slender decline phase by a weaker density-dependence in the differentiation function and a weaker immune control. This way we can construct a piecewise upper bound for dV/dt corresponding to each phase of the dynamics illustrated in Figure 2.12, using Heaviside functions H, as follows:

$$\begin{aligned}
\frac{dV}{dt} &= r(1-\frac{V+M}{K})V - daV \\
&< H(\tau_K - t)rV + H(t-\tau_K-2\tau)\left[rV - r\frac{V^2}{K} - daV\right] \\
&< H(\tau_K - t)rV - H(t-\tau_K-2\tau)r\frac{V^2}{K} \\
&:= \frac{dV_u}{dt}.
\end{aligned} \tag{2.26}$$

We have used the fact that $a > r/d$ throughout Phase III mediated by host-immunity, i.e. for time $t \geq \tau_K + 2\tau$ and also that in the decline phase, stochastic extinction of the parasite population occurs if V falls below some critical value, $V_{ext} = 1$. This happens

at time $T_{ext} = 1/r(K-1)$. Thus,

$$\int_0^\infty V_u(s)ds = \int_0^{T_{ext}} V_u(s)ds = K\left[\frac{1}{r} + 2\tau + \frac{1}{r}\log\left(K\right)\right]. \tag{2.27}$$

Notice the terms in the last expression depend on K, r and τ. The relative magnitudes of the different terms will depend on the biological parameter values of the system. Upon the condition that the average parasite lifespan, given by $1/r$ is much smaller than double the immune response delay, $(1/r \ll 2\tau)$, the contribution of both terms $1/r$ and $1/r\log(K)$ inside the bracket is $o(\tau)$. In our system (see Table 2.1), $1/r = 10, \log(K) \leq 27$, thus satisfying the requirement for the approximation: $\int_0^{T_{ext}} V_u(s)ds \approx 2\tau K$. Thus, $\int_0^\infty V(s)ds < \int_0^\infty V_u(s)ds \approx 2K\tau$, and we define the critical activation rate by

$$s_{crit} = \frac{1}{2K\tau}, \tag{2.28}$$

as a lower bound, such that if $\bar{s}_i < s_{crit}$, we have:

$$P_i(t) > \exp\left[-\bar{s}_i \int_0^\infty V(s)ds\right] > \exp\left[-s_{crit}\int_0^\infty V(s)ds\right] > 1/e, \tag{2.29}$$

for all t, which means variant i will never be generated. On the other hand, for $\bar{s}_i > s_{crit}$ variant i may be generated (Figure 2.13).

Unsurprisingly, s_{crit} is determined by K, the within-host carrying capacity, and τ, the delay in immune response activation. When K and τ are large, each block of variants grows to a higher level and persists longer within the host before being suppressed by the host immune system, as discussed in Section 2.4. The s_{crit} threshold, being low, in this case favors the parasite by allowing rare variants to play their part in the dynamics, thus prolonging infection. In general, any changes in parameter values that increase the total replicative potential, $\int_0^t V(s)ds$, of the current antigenic block lower s_{crit} and facilitate stochastic emergence for future blocks.

We observe that so far, the carrying capacity, K appears to have two very important roles in the full dynamics of the model: it controls the duration of Phase II in each block wave, and at the same time it affects switching between consecutive blocks. Also the immune response delay appears be involved in both critical thresholds, η_{crit} and s_{crit}, consistent with previous studies showing that delays in immune activation considerably affect the window of opportunity for the pathogen to exploit its host before being suppressed by the immune system (Fenton *et al.*, 2006).

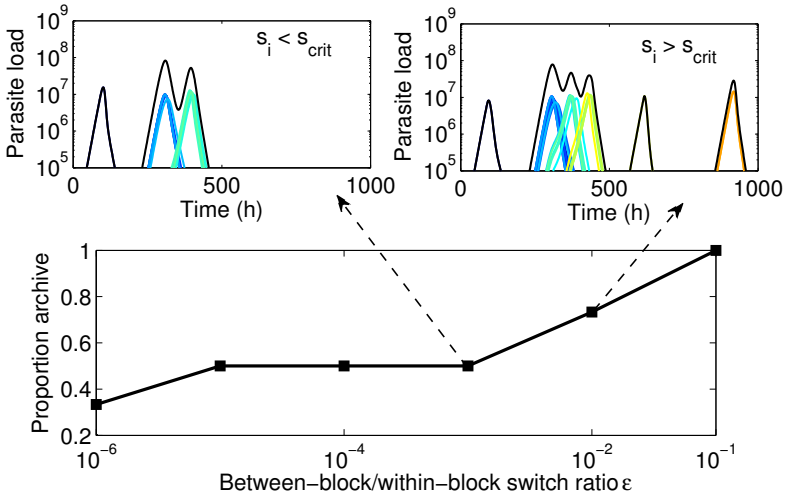


Figure 2.13: The role of ε on infection dynamics when switching is hierarchical. As between-block switch rates are reduced, subsequent peaks occur further and further apart and a smaller proportion of the archive can be generated during infection, because more variants have mean activation rates below $s_{crit} = 1/(2K\tau) = 5 \times 10^{-8}$. Parameter values: $K = C = 10^8, x = 3, \tau = 10, N = 30, \eta = 5$. All other parameters are as in Table 2.1.

2.5.2 How does the antigenic variation dynamics scale with K?

So far, we have considered changes in the switch rates, which describe the parasite antigenic archive. Now, we fix the archive and examine the role of the within-host carrying capacity, which via differentiation controls the maximal total number of parasites that can grow during infection and is likely to vary across different hosts. Some hosts may induce the differentiation process at lower parasite loads, while other hosts may induce it at higher parasite loads. Notice that body size is one constraint at the host level, determining the total blood volume of the host, thus potentially limiting the growth capacity of the pathogen.

To investigate how within-host dynamics scales with K, we assume K-invariant kinetics for the immune response rate (e.g. $K/C = constant$) and K-invariant parasite intrinsic growth rate. Notice that the initial inoculum size $V(0) = V_0$ is host-invariant as it generally corresponds to the number of parasites injected by the tsetse. The model shows that keeping the archive (switch matrix), and other kinetic parameters fixed, the average peak total parasite load increases with K, especially when density-dependent

differentiation of the parasite into stumpy form acts faster than specific host immune responses (e.g. $K/C \leq 1$).

Another effect of the carrying capacity K is the role it plays in the initial rate of growth of the parasite. This initial growth rate depends on the proportion of the total carrying capacity occupied by the initial parasite population: $dv/dt \approx rV_0(1 - V_0/K)$. When K is larger, the initial growth rate of the parasite within the host is also larger, leading to a faster increase in the number of variants, and hence to stronger density-dependent limitation of each variant. As a result, the first peaks that appear are relatively high, sharp and narrow when K is small, while appearing lower, rounder in shape and wider when containing more variants, when K is large (Figure 2.14).

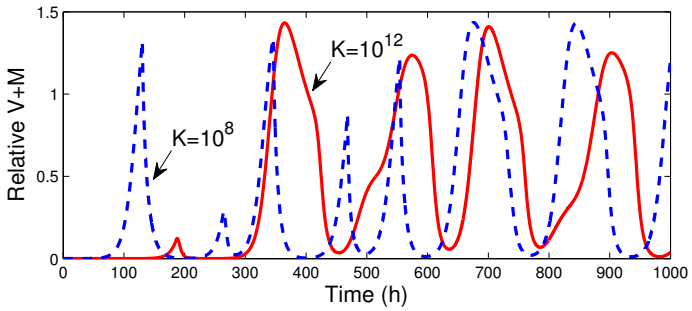


Figure 2.14: Relative parasite dynamics $(V+M)/K$ varies with K. Simulation parameters as in Table 2.1, with $C = K, \eta = 3, N_{blocks} = 5, \varepsilon = 0.001$.

Thus, when the archive is fixed, K effects the rate of antigen turnover. Because the stochastic generation of variants depends directly on the magnitude of the slender cell population ($\sum_j v_j$ in Eq. 2.25), which in turn depends on the carrying capacity, new variants are generated faster when the carrying capacity is larger (see Figure 2.15). Such differences have been observed empirically (Morrison, 2007), and in the model, the effect is stronger for variants with low activation rates and a hierarchical switch matrix. When switching to new blocks occurs with increasingly lower rates ($\varepsilon, \varepsilon^2, \varepsilon^3$ and so on), there is a high chance for stochastic extinction of the parasite in hosts where K is small. The total parasite population is simply too low to allow for these rare events to appear.

Hence, in the limit as ε tends to 0, we expect more differences in the rate of antigenic variation and the overall proportion of the archive expressed over the course of

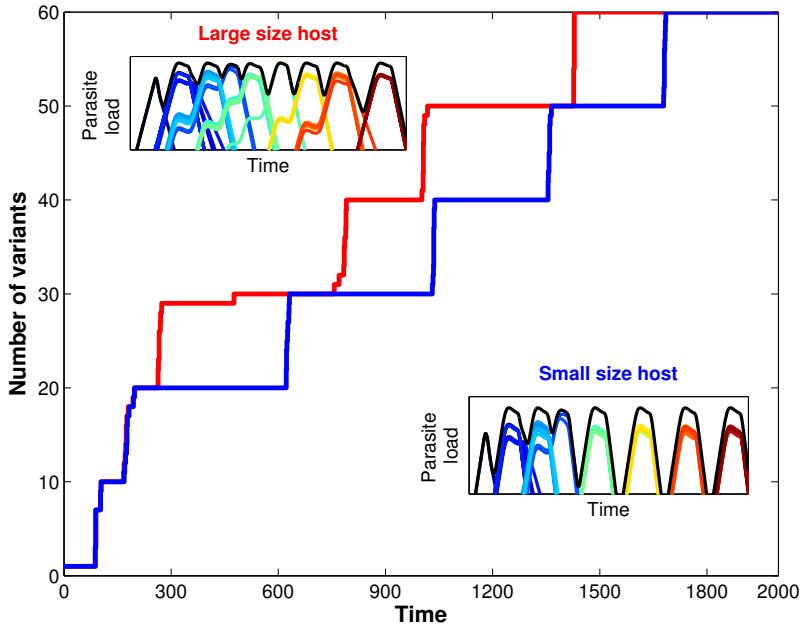


Figure 2.15: The antigen turnover rate varies with K. The lines show the cumulative number of variants that have been generated up to time t over infection ($t_i > t$). If K increases with host body size, in larger hosts ($K = 10^{12}$) we expect the same variants to be expressed earlier than in small hosts ($K = 10^9$). Simulation parameters as in Table 2.1, with: $C = K$. The switch matrix used is non-hierarchical, with $\eta = 10, N_{blocks} = 6, \varepsilon = 0.0001$.

an infection between hosts with different K. In addition, when comparing parasite population diversity during infection (as given by Shannon entropy (Shannon, 1951)) between hosts with different K, we observe that less parasite variants circulate in small K-hosts relative to large K-hosts per unit of time, even when the parasite antigenic archive is the same.

2.5.3 Effects of the switch matrix

To summarize, there are three main descriptors of the antigenic archive reflected in the switch matrix of our model: the size of a single block η, the number of blocks B, and between-block switch rates determined by ε. We find that when single blocks get larger, infection duration increases but if η is too large, infection persists indefinitely, independently of the number of blocks and their between-block connectivity.

If the number of blocks, B increases, infection duration and the number of peaks increase almost linearly, without affecting the nature of the dynamics of each peak. We observe that archive modularity, i.e. larger B, enhances the overall oscillatory behaviour and the sharpness of individual peaks in infection, even when the total number of variants, N, is kept constant. We expect that a sharp parasite peak in a real infection may well indicate a small antigenic block and strong variant-specific immune control. If the shape of a peak appears rounder, this may suggest a larger number of variants contained therein and a weak immune control. An important feature of infection dynamics then, the variability of peak densities, provides a further clue about dynamical mechanisms that may be responsible for parasite population cycles within the host. This structural property found in infection profiles occurs also in other more general predator-prey contexts (see (Turchin *et al.*, 2000) for an example from lemming population cycles), and may potentially be exploited to decipher top-down and bottom-up control from experimental observations of a dynamical system.

Finally, when between-block switch rates are lower (small ε), the separation between block waves increases, giving longer infection duration overall but less antigen diversity within each peak (Figure 2.10). However, there is a trade-off: if between-block switch rates are too low, variants of subsequent blocks may not be generated at all (Figure 2.13). This is related to the s_{crit} phenomenon discussed in Section 2.5.1. So if switching between blocks is low, the benefit from longer duration must offset the risk of stochastic parasite extinction.

2.6 Extensions of the standard model

The current model may be extended to capture a more realistic host immune response. Arguably, our formulation so far neglects decay in antibody responses but this assumption is not an over-simplification for trypanosome infections, where immune memory plays a central role. Undoubtedly, when more data becomes available, a mechanistic description of antibody and memory B-cell dynamics would be preferable to the phenomenological approach adopted here. However, we explore three other immune scenarios that might prove important: 1) the existence of prior immunity against some variants; 2) cross-reactivity between different variants, and 3) immunosuppression as

a result of active infection in the host. Clearly, these extensions affect the dynamics mostly when host immunity dominates in controlling an infection.

2.6.1 Prior immunity

The existence of prior immunity results in partially immune hosts either via prior exposure to infection, or via vaccination, having provided some initial level of strain-specific immunity. We find that the size of a single block becomes especially important when there is prior immunity against some variants: the larger the block, the higher the chance for the parasite to establish an infection by switching sufficiently fast to a variant unseen before. Notice that Phase I of the dynamics of any variant occurs only if the level of pre-existing immunity against this variant is low, more precisely if $a_i(0) < r/d$ and if $v_i(0) < K(1 - da_i(0)/r)$. If the size of the first block is too small, then there are only a few new variants available for expression, and the parasite risks immediate extinction in the partially immune host. This suggests another selective pressure on the size of single blocks in the antigenic archive of trypanosomes, namely the requirement to seed an infection in partially immune hosts.

2.6.2 Cross-reactivity

There is evidence that trypanosome mosaic *VSG* variants emerging in the chronic stages of infection, with at least 75% sequence identity, exhibit high levels of cross-reactivity (Marcello & Barry, 2007b). Variants within an antigenic block generally have a high sequence identity, whereas variants across blocks have a low sequence identity. In this context, by adding cross-reactivity into the model, we introduce positive coupling between specific immune responses, allowing existing immune responses against particular variants to partially clear other variants. The model is modified by replacing da_i and δa_i in Eqs. (2.1) and (2.2) by $d\sum_{j=1}^{N} a_j\gamma_{ji}$ and $\delta\sum_{j=1}^{N} a_j\gamma_{ji}$, where γ_{ji} reflects the probability that antibodies raised against variant j can clear variant i. In other systems, including malaria, cross-reactivity has been proposed, at least on a theoretical level, to play a significant role on the order of variant expression (Recker *et al.*, 2004). To analyze cross-reactivity in the context of trypanosomes, we focus only on naive hosts, and only on the uniform cross-reactivity case, thus this feature of the immune response plays no role in the order of variant appearance.

2.6.2.1 Within a block

We find that uniform cross-reactivity within a block acts as a general background immunity, increasing the net clearance rate of each variant of that block. When the block size is large, this results in a lower peak parasite load, the opposite of what was seen in the immunity-dominant scenarios of Section 2.4.4. However, when differentiation dominates the parasite control, cross-reactivity has no effect on the dynamics, as the growth inhibition of the parasite due to differentiation is generally much stronger and more significant than the growth inhibition due to cross – reactivity. Thus variant peaks still decrease with η and the block wave duration increases with η, only with cross-reactivity, their sensitivity to η is higher.

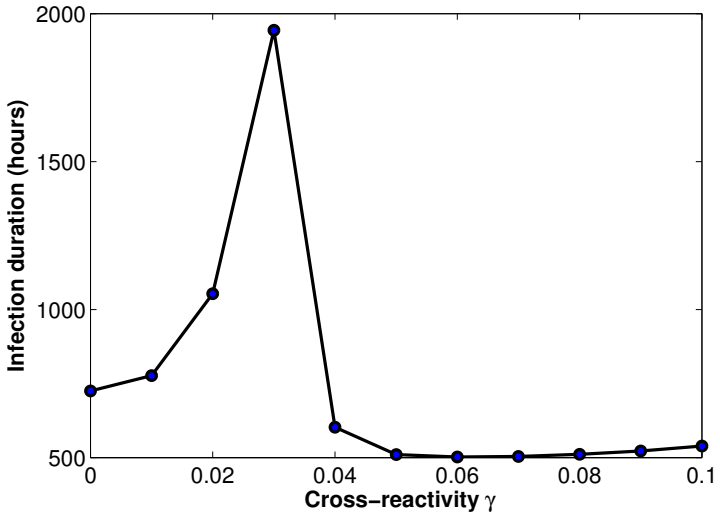


Figure 2.16: Infection duration as a function of between-block cross-reactivity γ. Intermediate values of cross-reactivity are good from the pathogen perspective as they allow longer persistence and enhance transmission. Parameters are as in Table 2.1, with $N = 15, \eta = 5, K = C = 10^8, x = 2$, and S is non-hierarchical. Duration is calculated here as the time it takes for $V + M$ to fall below its initial value $V_0 = 10^3$.

2.6.2.2 Between blocks

Cross-reactivity between archive blocks adds another factor of relatedness between parasite variants, in addition to the switch matrix. Unlike the switch matrix that de-

termines how fast the parasite moves in antigenic space, cross-reactivity marks how advantageous that movement really is.

Paradoxically, for intermediate values of uniform cross-reactivity between blocks, the parasite is able to prolong infection (see Figure 2.16), because persistent immunity raised against the early variants is sufficient to suppress later variants, resulting in their specific antibody responses not quickly reaching their maximum. This leads to lower levels of future parasite peaks within the host, as more variants emerge, but spread over a longer time-scale, an effect shown previously for Plasmodium (Recker & Gupta, 2006). This is reminiscent of what has been observed in some trypanosome infections of cattle that eventually self-cured (Figure 1.2). The decrease in parasite load over the course of infection could be due to the competence of the host in raising cross-reactive antibodies that made it more and more difficult for the parasite to express antigenically different variants or mosaics, eventually leading to archive exhaustion. For higher between-block cross-reactivity, however, unsurprisingly, clearance of future variants happens more rapidly, decreasing both the parasite load and infection duration (Figure 2.17).

Furthermore, depending on the immune response sensitivity x, the level of cross-reactivity between consecutive antigenic blocks may play a role in determining the relative critical threshold for the new block size as a function of the size of the previous block. Assuming the cross–reactivity between all variants of the previous block (η^{old}) and all variants of the new block (η^{new}) equals γ, we can apply similar time-scale and quasi-steady-state arguments as in Section 2.4.3, to derive the critical block size threshold for the new block, depending on the degree of cross-protection from antibodies raised against the previous block (see Appendix A.3 for full details):

$$\eta_{crit}^{new} = \frac{K}{C}\left(1 - \frac{d\gamma}{r}\eta^{old}\right)\left(\frac{-\log(1 - r/d + \gamma\eta^{old})}{c\tau}\right)^{-1/x}. \tag{2.30}$$

Interestingly two opposing effects can be observed: η_{crit}^{new} decreases with the number of variants in the previous antigenic block (η_{old}) when the immune responses are less sensitive to antigen stimulation ($x \geq 2$), implying that cross-reactivity between blocks may enhance the possibility of long persistence of future variants; but if immune responses are very sensitive ($x = 1$), the critical block size of the new block, η_{crit}^{new}, is

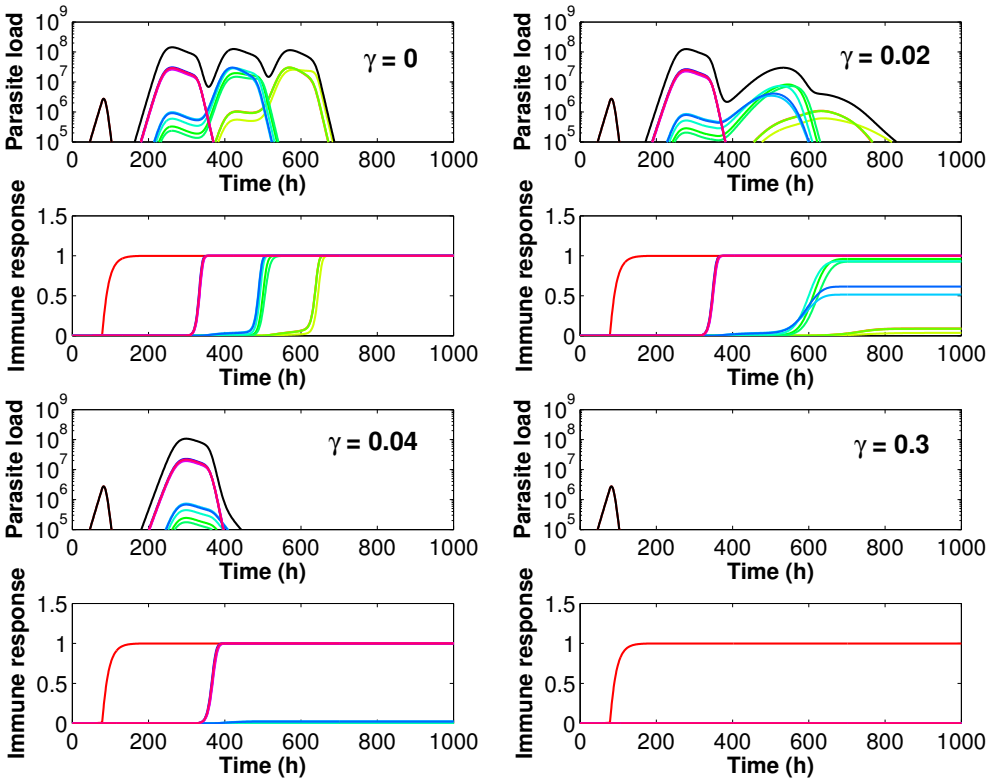


Figure 2.17: Infection dynamics with cross-reactivity between blocks of antigenic variants. As γ increases, not only is infection duration gradually reduced, but also stochastic emergence of new variants is made impossible and only a small repertoire of variants is ever seen. The black lines indicates $V + M$, the colored lines $v_i + m_i$. Parameters as in Figure 2.16.

an increasing function of η_{old}, implying that cross-reactivity between blocks may enhance future variant clearance if immune action is fast and tightly coupled to parasite numbers.

In other words, depending on the general sensitivity of immune responses, the level of cross-reactivity between consecutively activated antigenic blocks may play a role in selecting for their optimal sizes. In order to ensure host survival, the parasite must avoid overwhelming the host with long persistence of all variants. In general, this requires that η_{crit} be large. When immune responses across blocks of variants are cross-reactive, in order for antigenic variation to give rise to oscillating parasite loads (as opposed to overwhelming infection persistence), the antigenic archive of the parasite should be structured in such a way that: i) later-appearing antigen blocks are small if early antigen blocks are large, in slow host-immunity scenarios (x large), and ii) late antigen blocks are large if early blocks are large in fast host-immunity scenarios (x small).

This analysis suggests that immune cross-reactivity across parasite variants may be important in within-host antigenic variation, affecting not only the magnitude of infection peaks, and overall infection duration, with potential implications for host health and parasite transmission, but also affecting the evolutionary selective pressures on the block structure of the antigenic archive.

2.6.3 Immune suppression

It has been shown that the host's capacity to control an infection may be limited: in inbred mice, the supply of naive B-cells can decrease dramatically during a trypanosome infection, with the consequent decrease in the rate of production of specific antibodies (Radwanska *et al.*, 2008). While the exact nature of immune suppression may be determined by a series of different factors, we considered only the number of antigenic variants as a factor, an index of parasite diversity within the host, which represents the most conservative case.

2.6.3.1 Instantaneous immune suppression

When immune suppression depends on the number (η) of variants within the same block, i.e. from diversity currently circulating within the host, Eq.2.3 can be replaced

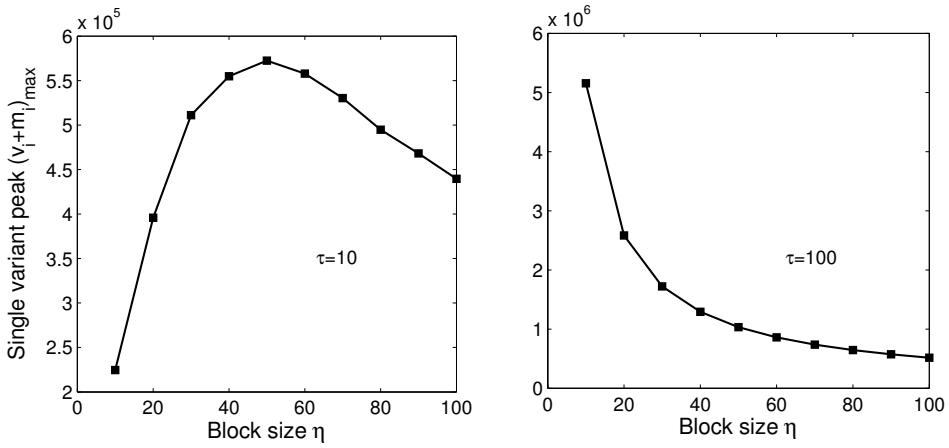


Figure 2.18: Effects of immune suppression from parasite diversity within a block, η. When host immunity is intrinsically fast ($\tau = 10$, left panel), $(V+M)_{max}$ increases more than linearly (for a while) with increasing η, by the additional growth experienced by each individual variant $v_i + m_i$. Instead, when host immunity is slower ($\tau = 100$, right panel), the immunosuppression positive effects by increasing diversity are counterbalanced faster by the negative effect on each variant exerted by density-dependent differentiation. As a result, the total peak is fixed and each individual variant gets a smaller share of $(V+M)_{max}$ when η is large. Parameter values as default, except: $C = 10^7, K = 10^8, c = 10, x = 1$.

by

$$\frac{da_i}{dt} = \frac{c}{A(\eta)} \left(\frac{v_i(t-\tau) + m_i(t-\tau)}{C} \right)^x (1 - a_i), \tag{2.31}$$

where $A(\eta) = \phi e^{\alpha\eta}/(\phi + e^{\alpha\eta} - 1)$ (see Appendix A.4 for details and analysis). We find that in the presence of immune suppression, increases in the block size η produce even higher increases in the peak total parasite load, than previously shown. However, as differentiation becomes stronger with increasing η, effects from within-block immune suppression become weaker, and the density-dependent effects from differentiation, which lower variant peaks, become dominant (Figure 2.18). So in general, for moderate immune suppression, the qualitative behaviour of the system is very similar to the inexhaustible immune response case analyzed throughout Section 2.4.

2.6.3.2 Cumulative immunosuppression

When immune suppression acts as a function of the cumulative number of variants seen by the host, the effects can be more dramatic, especially if many variants are generated early in infection. Eq. (2.3) is replaced by

$$\frac{da_i}{dt} = \frac{c}{A(N_t)} \left(\frac{v_i(t-\tau) + m_i(t-\tau)}{C} \right)^x (1 - a_i), \tag{2.32}$$

where N_t refers to the total number of variants generated up to time t. Cumulative immune suppression has the consequence that the intrinsic host capacity to mount an immune response against new variants (given originally by c in the model) is progressively reduced towards the limiting value c/ϕ. As a result, variants that emerge later in infection face a weaker immune system, grow more rapidly and persist for longer times. Depending on the precise value of the limiting value ϕ, and the speed of immune deterioration α, this effect will be weak or strong and may even lead to overwhelming long term parasite persistence. In contrast to the instantaneous immune suppression case, the benefits from cumulative immune suppression are experienced by the parasite only later in infection.

2.7 Discussion

Although a large number of studies have addressed trypanosome within-host infections, focusing on isolated aspects of the dynamics like variant order (Agur *et al.*, 1989; Kosinski, 1980; Seed, 1978; Turner & Barry, 1989), the interaction of specific and cross-reactive host immunity, (Antia *et al.*, 1996), the connectivity pathways between variants (Frank, 1999; Lythgoe *et al.*, 2007) and parasite density–dependent differentiation into the stumpy parasite form (Savill & Seed, 2004; Tyler *et al.*, 2001), none of the models therein have analyzed all those aspects together in one context. The current model binds all these elements in a common framework, offering deeper insight into trypanosome dynamics and closer integration of these dynamics with switching mechanics revealed by the increasing availability of parasite genetic data.

The model illustrates how the structure of the parasite antigenic archive dictates critical thresholds at the within-host level, which have important implications for infection. The balance between specific immunity and parasite differentiation, while

determining many features of chronic infection, such as peak parasite load, duration, and slender/stumpy ratio, emerges to be very sensitive to the size of an antigenic block, η. Variance in the size of different blocks in the switch matrix can make this balance dynamic over the course of an infection, giving rise to variability in infection profiles. Another finding of the model is that demographic stochasticity, as shown also by Sasaki & Haraguchi (2000), can matter in within-host antigenic variation dynamics, in our case especially, since the switch rates between variants range over several orders of magnitude. The effects of stochastic variant emergence are particularly important early on, if the parasite infects partially immune hosts, and generally in the chronic phase, where the maintenance of infection depends on rapid jumps between antigenic blocks. The minimum switch rate, s_{crit}, required for stochastic generation of a new variant is an emergent property of the model and is not dependent on the details of the switch matrix, whether hierarchical or not.

Two different chronic infection scenarios arise from the model. A stationary chronic infection, where Phase II lasts indefinitely, requires a single block to be large enough to allow differentiation to dominate parasite control within the host. In this case, the same variants persist throughout the infection, limited only by the carrying capacity. An oscillatory infection with multiple peaks (characteristic of trypansome infection), requires many small blocks, so that immune-mediated clearance of each block is possible, and sufficiently high between-block switch rates to enable exploration of other antigen blocks. The latter scenario is characterized by sequential emergence of new variants. These two requirements correspond to the two critical thresholds derived in this chapter, η_{crit} and s_{crit}. Essentially the parasite faces a trade-off for maximal use of its archive. Jumping to consecutive blocks should occur neither too fast, to avoid overwhelming the host, nor too slowly, to avoid premature infection clearance.

For a fixed archive instead, these two chronic scenarios (stationary and oscillatory) are attained for different levels of host immune-competence. We expect that in immune-compromised hosts, persistence of a stationary infection is established more easily, whereas in immune-competent hosts, multiple-peak chronicity is established more easily. This model prediction is confirmed by some existing empirical studies (Hajduk & Vickerman, 1981), but can be tested further experimentally. The relative differences between these two scenarios, link also to the general differences between

acute and chronic infection (Alizon & van Baalen, 2008a) and may have an important evolutionary significance for the parasite.

2.7.1 How genomic data can inform the model

The precise values of antigenic switch rates, lying at the interface between parasite genetics and within-host dynamics, may be very difficult to extract empirically. In a top-down approach, they would require a longitudinal study of several parallel infections and multiple screenings for variant identification at each peak. In this way, an antigenic network could be inferred statistically from co-occurrences of variants, as illustrated recently for Plasmodium (Recker *et al.*, 2011). A few studies over the first 3-5 weeks of infection do show that individual variants are predictable in time of appearance (Marcello & Barry, 2007a; Robinson *et al.*, 1999; Timmers *et al.*, 1987), but more high-throughput approaches could be applied over longer infection periods. The block size η would then correspond to the size of clusters in this network. In a bottom-up approach, relative switch rates might be inferred from sequence analysis of *VSG* gene flanks, as was achieved by Barbour *et al.* (2006), who showed that gene flank characteristics dictate fine timing of expression of variants in the bacterium *Borrelia hermsii*.

If the switch rates between trypanosome *VSG* variants can be mechanistically derived from genetic processes in the antigenic archive, and if the rates of these genetic processes are measured, then parameters of the switch matrix such as η, N, ε can be properly quantified, and moreover, the global structure of the switch matrix, whether hierarchical or non-hierarchical, or a combination of both, can be elucidated. Clearly, an integration of bottom-up and top-down approaches provides an ideal framework, where the model can meet experiments.

2.7.2 Future work and perspectives

Whether a stationary or an oscillatory chronic infection confers a higher fitness to the parasite remains unclear. An oscillatory parasite load may be of selective advantange over a stationary one, because its cumulative negative effect over time on host survival might be smaller, thereby extending the time-window for transmission of the parasite. Alternatively, the two strategies may endow the parasite with equal fitness, and hence

trypanosome strains adopting either archive strategy should be possible to find in the field. If such different strains are not observed, this could indicate that only certain archive configurations that are easy to attain genetically can evolve.

Next comes the question of how the critical thresholds at the level of a single host affect parasite fitness at the population level. We can now begin to ask: at which antigenic block size do the virulence effects from long parasite persistence start to outweigh the transmission benefits? How does this depend on the balance between parasite differentiation and host immunity? The model suggests that the continuum of differentiation-immunity scenarios may favour different antigenic block sizes, in order to maximize parasite load within the host. However, because trypanosomes are vector-borne parasites, after some threshold, increases in the number of variants under one peak (η) and in the total parasitaemia, may only serve to increase virulence of the infection and not transmission. In this sense, we would expect very virulent strains, e.g. strains with large blocks of VSG genes in their VSG repertoire, to be counterselected for.

Although we have not embedded our within-host model into an epidemiological context, some critical links with transmission and virulence already emerge through parameters such as the within-host carrying capacity K. Related to the rate of parasite differentiation, K plays a critical role as a ceiling for parasite population growth, affecting ultimately the replication and transmission potential of the parasite. Clearly using the model to address infection across different host species would involve some perturbation of K. Presumably the size of the host may influence K, serving as an ecological upper bound, and some experimental data supporting this idea already exists (Barry, 1986). However, more research is needed to properly connect K to trypanosome infection and antigenic variation. Possible avenues for elucidating K from experiments, just to mention a few, could be the empirical exploration of trypanosome quorum sensing mechanisms, a better understanding of the kinetics of the SIF factor triggering differentiation (Reuner *et al.*, 1997), and the identification of host mechanisms that might be involved therein.

From the fact that the critical between-block switch rate depends on K, it seems plausible that parasite genetic processes, responsible for the VSG archive subfamily structure, must ultimately evolve towards some K-dependent optimum, ensuring sufficient separation between blocks but also a minimum degree of connectivity. Any

host factor that reduces the parasite replication potential must be counterbalanced by increases in between-block switch rates (e.g. genetic identity between variant subfamilies) to ensure stochastic generation of new blocks. Again one can ask how large are the parasite transmission benefits from expressing a higher number of blocks within a host, compared to virulence costs? How does this depend on the particular type of host and its ecology?

Further research in this direction, linking selection pressure at the within-host and epidemiological level with the genetic processes operating on the parasite antigenic archive could have important implications for our understanding of trypanosome evolutionary dynamics.

We will address some of these issues in the next chapter, where we aim to uncover the wider ecological context of a trypanosome infection and explicitly consider parasite fitness as a function of within-host infection kinetics.

Chapter 3

Understanding trypanosome fitness: how to optimize infection profiles

3.1 Introduction

Many pathogens cause chronic infections by varying their antigenic properties within hosts (Barbour & Restrepo, 2000; Barbour *et al.*, 2006). Antigenic variation has important implications for disease progression within an individual and for transmission of the pathogen across many generations of hosts (Lipsitch & O'Hagan, 2007). Persistent infections like malaria (*Plasmodium falciparum*) or sleeping sickness (African trypanosome) rely on parasite antigenic variation, characterized by oscillating pathogen load, where each peak is dominated by a distinct wave of antigenic variants of the parasite.

So far, we have focused on the within-host implications of antigenic variation for African trypanosomes. In Chapter 2, we examined the direct and indirect effects of antigenic archive structure on within-host trypanosome dynamics and showed that different structures of the antigenic archive can have an effect on infection duration, peak parasite number, the ratio between replicative and transmissive parasite life-stages and oscillatory parasitaemia. However, there are larger-scale consequences of antigenic variation for infectious disease epidemiology that are tightly coupled to the within-host level. Onward transmission of the parasite to the vector and subsequently to new hosts depends on the infectivity of the current host, which in turn depends on the duration of infection and peak parasite load. If the infection is short, the probability to

transmit to the vector is small and viceversa. Similarly, if the peak parasite load is low, the density of parasites in a vector bloodmeal may be insufficient to successfully establish an infection in the vector. In this context, it becomes natural to extend the within-host framework to a larger ecological framework, where additional host and vector characteristics must be taken into account.

The aim of this chapter is to take an integrative cross-scale approach to the study of tryapanosome antigenic variation. For this, we use the relatively new modelling approach, of nested models, proposed by Gilchrist & Sasaki (2002), but see (Mideo *et al.*, 2008) for a recent review. By nesting the within-host model of Chapter 2 within an epidemiological framework, we investigate the larger-scale consequences of antigenic variation for parasite fitness within a host. Subsequently, we discuss the selection pressure on the archive structure, arising from host epidemiological and ecological factors and routes of transmission.

The exact factors involved in the evolutionary dynamics of pathogens depend on the type of pathogen, the particular host, and the particular trait in question, however there are many parallels across different systems, which can often be explained through general trade-offs that emerge between transmission and virulence. Some empirical studies have found trade-offs between transmission and virulence in host-parasite systems (Mackinnon & Read, 1999). The predictions coming from theoretical models have long postulated that as a result of such trade-offs, parasites should evolve intermediate levels of virulence (Anderson & May, 1982; Ewald, 1983; Levin & Pimental, 1981). Similarly, nested models of pathogens with different stages in their life cycle have shown that such parasites should evolve optimal rates of switching between reproduction and transmission life-stages to maximize their fitness (Alizon & van Baalen, 2008b; Koella & Antia, 1995; Sasaki & Iwasa, 1991).

Applying the nested framework to a novel context (African trypanosomes) allows us to revisit the emerging trade-offs between parasite transmissibility and virulence, and explore the role played by specific antigenic archive parameters in determining these trade-offs. Our approach is to assume that the structure of the antigenic archive, reflected in the block size, the number of blocks and between-/within-block switch ratio is a conserved and heritable parasite trait to a large extent, and as such is subject to genetic control and adaptation over long time scales. A natural question to ask is: do transmission-virulence trade-offs lead to similar intermediate magnitudes for

antigenic archive parameters? If yes, how do these intermediate optima depend on host characteristics? Furthermore, how are these optimal archive configurations achieved genetically by the parasite?

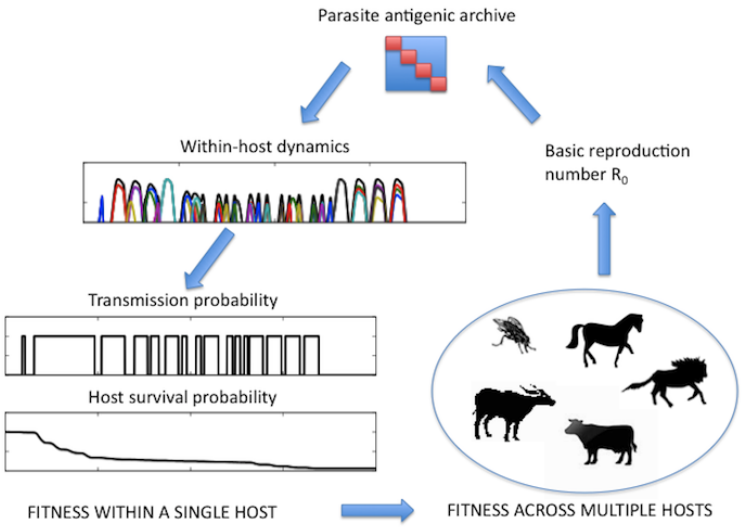


Figure 3.1: Illustration of the complex ecological feedbacks arising in the evolution of antigenic archives of pathogens such as the African Trypanosome.

Life-history evolution of multi-host parasites, such as the African trypanosome, raises many questions about the importance of epidemiology and the constraints emerging at the within- and between host level (Gandon, 2004). One of the most interesting questions has to do with the evolution of generalist versus specialist survival strategies. It has been shown that trypanosome hosts often differ in their ability to fight parasitic infections, or in the pathogenesis they experience (Barry, 1986; Morrison & Murray, 1985; Taylor, 1998). Thus, another issue of particular relevance addressed in this chapter is how the multiplicity of potential trypanosome hosts (each with its own immune competence, body size, etc.) can impact on the evolution of antigenic archive traits and parasite speciation. As illustrated in Figure 3.1, our aim is to achieve a more comprehensive understanding of the role played by antigenic variation in the life cycle of the African trypanosome, and ultimately how its long-term evolution is coupled to its host range in the field.

3.2 Pathogen success across biological scales

3.2.1 Pathogen fitness in the host community

We begin by defining parasite success at the population level. Because trypanosomes can infect many hosts, and alternate between hosts and vectors, the expression of their overall fitness in the field must include these transitions. The mathematical theory for the basic reproduction number of multihost pathogens has been developed by Diekmann *et al.* (1990). The formal derivation for R_0 via basic principles is straightforward. Intuitively, it is the sum of the contribution to R_0 by each host type and can be derived using the next-generation-matrix, which is defined as follows for vector-transmitted pathogens that infect multiple host species,

$$G = \begin{pmatrix} 0 & 0 & 0 & \dots & g_{1N} \\ 0 & 0 & 0 & \dots & g_{2N} \\ 0 & 0 & 0 & \dots & g_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{N1} & g_{N2} & g_{N3} & \cdots & g_{NN} \end{pmatrix}.$$

In the next-generation-matrix, each entry g_{ij} denotes the expected number of new cases of type i individuals, caused by one infected individual of type j during its entire period of infectivity, in a population consisting entirely of susceptibles. The last row of G corresponds to transmission from host to vector, whereas the last column corresponds to transmission from vector to host. Here we directly obtain g_{ij} from the definition, assuming each host population is in demographic steady state prior to invasion by the pathogen and in the initial phase of the invasion. This applies also to the vector population, given by Z. We then have:

$$g_{iN} = b_i \alpha_H \frac{1}{\gamma}, \quad \text{(vector-to-host)} \tag{3.1}$$

$$g_{Ni} = b_i \frac{Z}{H_i} F_i, \quad \text{(host-to-vector)} \tag{3.2}$$

where b_i is the per capita vector bite rate on host i, α_H is the probability of the host becoming infected following a bite from an infected fly, $1/\gamma$ is the vector life expectancy, H_i is the abundance of host i, and F_i is the average duration of host's infectiousness in the ith host species.

Diekmann *et al.* (1990) show that the dominant eigenvalue of G is equivalent to the basic reproductive number R_0. The eigenvalue with the largest real part has precisely the properties of R_0. We define

$$R_0 = \frac{Z\alpha_H}{\gamma} \sum_{i \in Hosts} \frac{b_i^2 F_i}{H_i}, \tag{3.3}$$

as the basic reproductive number for the multi-host pathogen. Implicitly, it is assumed that epidemiological parameters in the vector (virulence, transmission and recovery) are constant, and that the vector population reaches equilibrium more rapidly than the host population. If $R_0 > 1$, the parasite can maintain itself in the host population, otherwise the infectious disease goes extinct with probability 1. From Equation 3.3 we can see that a strong contribution to R_0 comes from F_i, coupled with the abundance of each host species H_i, and the host-specific contact rates b_i.

The contribution that each host type gives to R_0, namely F_i can be taken as an index of average parasite success in that particular host, in other words as 'parasite within-host fitness'. Notice that F_i can be calculated in 2 ways: 1) from the population level, using population level estimates for infection duration and transmissibility; or 2) from the within-host level, by computing the integral over time of the host probability to transmit the infection, weighted by the probability that the host is alive. The latter approach corresponds to the nested model that we present below.

3.2.2 Pathogen fitness in a single host

The net duration of host infectiousness, denoted by F, is equal to its instantaneous ability to pass on the infection $\beta(t)$, weighted by the probability $\varphi(t)$ that the host is alive, and integrated over the entire infection period. These depend in many ways on parasite within-host dynamics. For example, for trypanosomes, the total number of transmission life-stages (i.e. stumpy cells) present in the host at a particular time will affect the probability to successfully infect the vector. Similarly, the total number of parasite cells circulating in the bloodstream will harm the host directly, by using crucial resources, or indirectly, by stimulating the host's immune system, with the consequence of ultimately reducing host survival. Following the nested modeling approach originally proposed by Gilchrist & Sasaki (2002), one can use a within-host model and knowledge about the biology of the system to precisely quantify these relationships.

To describe the within-host dynamics of the trypanosome during a chronic infection, we use the hybrid model given by Eqs. (2.1)-(2.8) in Chapter 2. Recall that for each variant, slender (replicating) cells are denoted by v, stumpy (non-dividing, tsetse infective) cells by m, and the specific antibody response of the host is denoted by a. The total number of slender and stumpy cells respectively is given by V and M. Using our within-host model, we mathematically define net within-host trypanosome fitness in relation to a particular host type as:

$$\begin{aligned} F &= \int_0^\infty \beta(t)\varphi(t)dt \\ &= \int_0^\infty \beta(t)\exp\left[-\int_0^t \left(u+\mu\frac{V(s)+M(s)}{K_{max}}\right)ds\right]dt, \end{aligned} \tag{3.4}$$

where $\beta(t)$ is the instantaneous transmission probability, and the decreasing exponential function is the host survival probability. As with other vector-borne parasites, the transmission probability β for trypanosomes is a saturating function of parasite population within the host, in particular of the tsetse-infective stumpy subpopulation $M(t)$. As M increases, β tends to 1. Many complicated S-shaped sigmoid functions could be used, but the simplest interpretation for β is a threshold number of stumpy cells, z, in a tsetse bloodmeal, needed to infect the fly (Van den Bossche *et al.*, 2005).

We denote the volume of a tsetse bloodmeal by k, and the blood volume of the host by Ω. Assuming a homogeneous spatial distribution of the parasite population in the bloodstream of the host, the number of stumpy cells taken by a fly upon bite is given by $M_{fly}(t) = kM(t)/\Omega$. If $M_{fly} < z$ then $\beta = 0$, otherwise, if $M_{fly} \geq z$, the fly gets infected with probability $\beta = 1$.

For the probability that the host is alive over time, $\varphi(t)$, we assume it depends on natural host's mortality rate u (per unit of time) and the additional parasite-induced mortality rate, μ (per unit of time per parasite occupying one unit of the maximal carrying capacity K_{max}). Notice that μ reflects both properties of the parasite and of the host. In the following, we mostly assume that μ is parasite-intrinsic. However, it will be argued later that pathogen virulence can vary across hosts, leading to host-intrinsic pathogenesis rates.

For simplicity, it is assumed that both life stages of the parasite harm the host at equal rates. Consider that being an extracellular parasite, within-host growth of the trypanosome is ultimately limited by the blood volume of its host, which provides the

pathogen the resources needed for replication and survival. For this reason, the introduced parameter of K_{max}, denoting the maximal carrying capacity, can be thought to be directly proportional to host blood volume Ω. We expect the ratio K_{max}/Ω to be constant across different hosts, whereas the actual carrying capacity related to the differentiation process, K (and consequently $K/K_{max} \leq 1$) may vary, depending on possible host-specific factors that could be involved in parasite differentiation.

Let us now return to parasite fitness in the field (Eq.3.3). Clearly, the overall sensitivity of R_0 to parasite life-history traits is given by the sum of the sensitivities of the individual F_i, weighted by b_i^2/H_i. It is generally accepted that parasites evolve to maximize their basic reproduction number R_0. Such maximization however may lead the parasite to specialize on a single host species, resulting in a life-history trait that generates an optimal infection profile in the preferred host, or an antigenic archive selected to maximize R_0 across a community of different host species. The situation we consider here deals with how the parasite can maximize within-host fitness, F[1], using different within-host strategies. Conceivably, such maximization will inevitably require the balance between two requirements: maximization of the transmission potential of the parasite from the given host, and minimization of the harm done to the host (see Figure 3.2 for an illustration of the nested model).

First, in Section 3.3 we consider how the differentiation process impacts parasite fitness in a single host and what are the transmission requirements imposed upon the conversion from replicating to transmissive life-stages of the parasite. Then in Section 3.4, considering the structure of the antigenic archive as the evolving parasite trait, we investigate how this structure, represented by the switch matrix S, can change to optimize the infection profile in a given host, and which host and parasite factors play a role.

To obtain a deeper understanding of the implications of archive structure for trypanosome within-host fitness, in Section 3.5 we examine more in detail particular infection scenarios mediated by the antigenic archive, such as stationary and oscillatory parasite loads, and acute and chronic infections. In Section 3.6 we apply our analysis

[1]We drop the subscript i, assuming the host is given, in order to simplify our notation from this point onwards.

to realistic hosts, where changes in within-host parameters such as the carrying capacity are accompanied by changes in physiological and epidemiological parameters. The consequences of such an assumption for pathogen fitness are discussed in relation to the trypanosome antigenic archive.

In Section 3.7, we raise the questions of archive global optimality in the field versus archive plasticity in each host, and finally, in Section 3.8, we propose a mechanistic framework, through which the antigenic archive and the switch rates can be linked to genetic processes acting at the level of the parasite genome. We conclude with a summary of main results and general discussion.

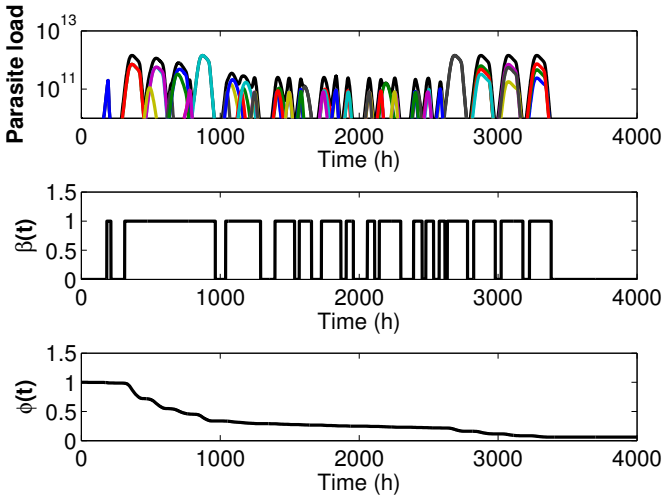


Figure 3.2: Illustration of within-host parasite dynamics integrated within an epidemiological framework. In this particular case, the within-host parasite fitness is $F = \int \beta(t)\varphi(t)dt = 757.24$. Parameter values as in Table 2.1 with: $K_{max} = K = C = 10^{12}, T = 4000, \eta = 3, N_{blocks} = 22, \varepsilon = 0.001, \mu = 8.8 \times 10^{-3.5}$. Different variants are given in arbitrary colour in the top graph.

3.3 Differentiation and within-host parasite fitness

In parallel with antigenic variation, there is another important within-host infection process: parasite differentiation from slender replicating form to stumpy non-replicating form. This process influences both the ratio between the two parasite life-stages and

their sum, and depends on the total parasite load within a host via $(V+M)/K$ in Eq. 2.1. Notice that the lower K is, the faster differentiation happens, and the higher K is, the more the parasite can grow and the slower the differentiation. Ultimately, K acts as a within-host carrying capacity: either a ceiling for the total parasite load $V+M$ when there is immune control, or an asymptotic equilibrium where $V+M$ settles in the absence of host immunity. In this section, we describe three ways in which the differentiation process may influence within-host parasite fitness F: by guaranteeing transmission, by favouring host survival, and by modulating the pace of antigen turnover during the period of infection.

3.3.1 Guaranteeing transmission

In the absence of host immunity ($a_i = 0$, for all i), which represents an ideal scenario for the parasite, the parasite population within the host, following the dynamics in Eqs.(2.1) - (2.8) tends towards the carrying capacity K (Figure 2.8). This is independent of the antigenic archive. At this steady state, $V^* + M^* = K$, and in particular the total stumpy cell population M^* and total slender population V^* are given by:

$$M^* = K\frac{r}{r+\delta_M}; \qquad V^* = K\frac{\delta_M}{r+\delta_M}, \tag{3.5}$$

where r is the intrinsic growth rate of slender cells and δ_M the intrinsic mortality rate of the stumpy cells. The ratio $M^*/V^* = r/\delta_M > 1$ (see Table 2.1) indicates the parasite-intrinsic tendency to favour transmission over reproduction in any host.

Naturally, a sufficient requirement for ensuring host-to-vector transmission in this best-case scenario is $M^*_{fly} > z$. If $M^*_{fly} < z$, $\beta = 0$ for all time. This can be translated into a threshold criterion for M^* and hence for either $r/(r+\delta_M)$ or K:

$$M^*_{fly} = \frac{M^*k}{\Omega} = \frac{Krk}{\Omega(r+\delta_M)} > z \implies \beta = 1. \tag{3.6}$$

So, in any given host, independently of the archive structure, the within-host carrying capacity, K must be sufficiently large relative to its blood volume Ω, in order to ensure sufficiently high concentration of parasites in a tsetse bloodmeal, and hence a non-zero transmission probability. Equivalently, for a fixed K, the ratio $r/(r+\delta_M)$ must be sufficiently high to fulfil transmission requirements, imposing a balance between the slender growth rate r and the stumpy cell natural mortality rate δ_M. Notice

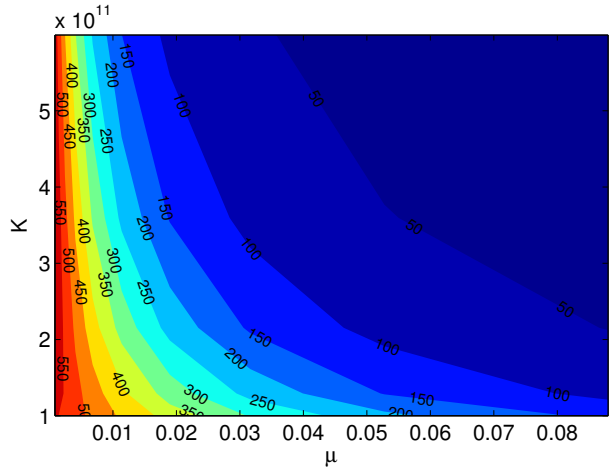
that, the inequality (3.6) exposes a continuum of within-host differentiation and kinetic strategies of the pathogen that can guarantee a nonzero transmission probability. The interplay between within-host growth parameters such as K, r and δ_M, the vector transmission threshold z, and physiological variables such as host blood volume Ω and vector bloodmeal volume k, provides the parasite much flexibility and freedom for parasite-host-vector adaptation, as long as $Krk/\Omega z(r+\delta_M) > 1$ is satisfied.

3.3.2 Favouring host survival

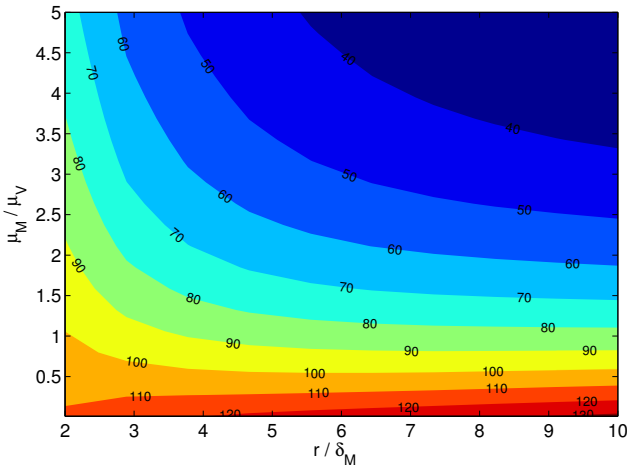
Another significant effect of parasite slender-to-stumpy differentiation may be related to host survival, $\varphi(t)$, which appears in the formula for F (Eq. 3.4). Because the overall host pathogenesis is a function of the total parasite population $V(t)+M(t)$, the maximal parasite burden reached in the host plays an important role in modulating this maximal level of host harm. At a general level, if $K = K_{max}$, the pathogenesis induced in the host if the parasite persists or peaks at K is maximal. In the persistence case, it can be easily seen that host survival declines rapidly to zero via: $\varphi(t) = \exp(-(u+\mu)t)$. Alternatively, if $K < K_{max}$, a faster differentiation process, by limiting sooner total parasite growth in the host, may lead to a smaller reduction in host survival and hence a higher overall infection fitness.

Within each parasitaemia peak, once K is bigger than a minimum threshold required for transmission when z is fixed (see Section 3.3.1, Eq. 3.6), further increases in K can be counterbalanced by corresponding reductions in parasite virulence μ to yield the same fitness of the parasite in a single host (see Figure 3.3(a)). This points towards the significant interplay at the within-host level between the differentiation process from slender to stumpy parasite form, generally responsible for the absolute magnitude of each peak, and the reduction in host survival caused by each pathogen cell living in the host bloodstream.

On the other hand, if the two forms of the parasite have different virulence, (e.g. $\mu_M \neq \mu_V$) the relative ratio V/M, besides simply their sum, becomes important for host survival. If stumpy cells reduced host survival by a smaller per-capita amount than slender cells ($\mu_M < \mu_V$), differentiation would provide the parasite a 'niche' that not only enhances transmission β, but also enhances host survival φ. Since the relative dominance of the two parasite forms for each variant depends primarily on the ratio



(a) Equal slender/stumpy virulence



(b) Differential virulence

Figure 3.3: a) Contour plot of within-host parasite fitness F as a function of parasite virulence μ and carrying capacity K. The same within-host fitness can be achieved by the parasite if it is highly virulent but differentiates rapidly (K small), or if it is less virulent and differentiates at higher parasite loads (K large). Red regions indicate high values of F, whereas blue regions indicate low values of F. Parameters as default. Switch matrix: $N_{blocks} = 4, \eta = 3, \varepsilon = 0.01$. b) Contour plot of within-host parasite fitness F as a function of differential virulence μ_M/μ_V of the two parasite forms, and their relative dominance within each peak $M^*/V^* = r/\delta_m$. The parasite can obtain the same within-host fitness from combinations of high stumpy dominance and low stumpy virulence, or high slender dominance and low slender virulence. Parameters as default. Switch matrix: $N_{blocks} = 1, \eta = 1$.

between the intrinsic slender growth rate and the intrinsic stumpy mortality rate r/δ_M (and subsequently on differential immune clearance rates), we obtain a fitness landscape where different combinations of r/δ_M and μ_M/μ_V can lead to the same parasite success in a single host (Figure 3.3(b)).

These observations suggest that both the rate of differentiation and the relative partitioning of the parasite population into transmissive and replicative life-stages are important factors in determining host survival over infection.

3.3.3 Pacing antigen turnover

In realistic antigenic variation scenarios, specific host immunity is present and mediates parasite clearance for each variant, and consequently a series of parasitaemia waves emerge. Each peak generally contains a few variants, thus has a limited reproduction potential. This reproduction potential (e.g. peak total slender cell number V_{max}) enables the parasite to switch to new variants, which is the primary mechanism of infection prolongation. This means that aside from the minimum level of stumpy cells required for transmission, the slender composition of a peak must also be sufficiently high in order to stochastically give rise to new variants over the course of an infection. Conceivably, the precise magnitude of the slender cells in each peak will affect the actual timing of the next wave of antigenic variants.

In the model, in addition to the differential immune clearance of these two types of cells, the relative ratio M/V within each peak, depends also on r/δ_M, as explained in Section 3.3.1. However, this effect becomes particularly evident when considering a simple modification of the model by adding a parameter α that represents the sensitivity of the inoculating variant ($i = 1$) to differentiation. This changes the dynamic equations of the first variant, altering the functional form (i.e. the strength) of the differentiation process, from $(V+M)/K$ to $[(V+M)/K]^{\alpha}$, and consequently the slender-stumpy composition of the initial peak. When α is high, the initial peak composed of the first variant consists mostly of stumpy cells, and subsequent variants take longer to emerge and grow, instead when α is low, the first peak consists mainly of slender cells and the pace of emergence and growth of new variants is faster, leading to more rapid antigen turnover and hence shorter duration of infection. Keeping everything else fixed, we find that there is an optimal sensitivity to differentiation for the first

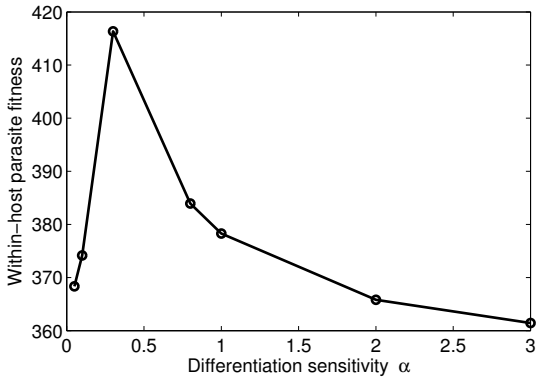


Figure 3.4: The effect of differentiation sensitivity of infecting variant. There is an optimal sensitivity to differentiation of the first variant initiating the infection, where within-host fitness is maximized. This optimum reflects the slender-stumpy composition of the first peak that optimizes the pace of antigenic variation within the host. Within-host parameters as in Table 2.1, with $K = C = 10^8, T = 1000$. Host-specific parameters: $K_{max} = 10^9, \mu = 8.8 \times 10^{-3}, u = 1.1 \times 10^{-4}, \Omega = 2 \times 10^{-3}$. Non-hierarchical switch matrix: $\eta = 2, N_{blocks} = 2, \varepsilon = 0.0001$.

variant (in this case $\alpha_{opt} = 0.5$), where within-host parasite fitness is maximized (Figure 3.4). With α_{opt} below 1, this suggests that at the beginning of infection, it is better if differentiation-mediated parasite growth arrest is downregulated to favour more reproduction. Such an optimum indicates that increases in immediate reproduction and transmission of the parasite must be traded off against decreases in future transmission potential.

Interestingly, we do not find a similar optimum arising when all variants display the same sensitivity α to differentiation. In that case, we observe instead, that as α increases and the entire parasite population differentiates faster, within-host parasite fitness, F, only decreases. This is perhaps due to the fact that the majority of variants are already generated within the first infection peak, thus the slender-stumpy composition of this peak is the most important factor governing the pace of antigen turnover.

To summarize, as it has been noted for vector-borne diseases (Alizon & van Baalen, 2008b), the processes partitioning the parasite population into replicative and transmissive parasite life-stages can have important epidemiological implications also for trypanosome infections.

3.4 How the archive structure impacts within-host fitness F

In Chapter 2, we investigated in detail how antigenic archive structure affects within-host dynamics. In this section, we address the implications of archive structure for within-host parasite fitness, taking into account transmission to the vector and host survival.

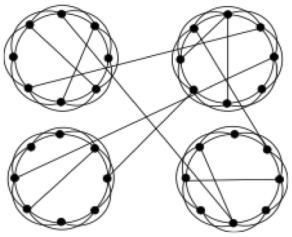


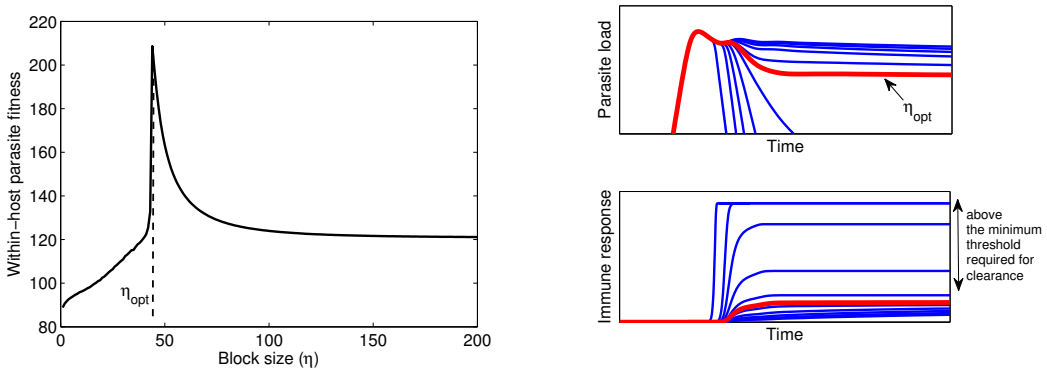
Figure 3.5: Illustration of the genetic architecture of the antigenic archive of the pathogen. Switching within blocks of closely related genes usually happens at a higher rate than switching between blocks.

By our construction of the switch matrix, we can analyze independently three different archive parameters: the size of a single block of variants η, the number of blocks N_{blocks}, and the between-/within-block switch ratio ε (Figure 3.5). We expect that different combinations of these parameters over an infection lead to different transmission success of the parasite. Conceivably, these archive parameters must be under selection to optimize within-host parasite dynamics. The quantitative resolution of the trade-off between transmission and virulence must lie at the core of the optimal configuration of the parasite antigenic archive.

3.4.1 The optimal block size η

In Figure 3.6(a), we explore how the block size η can evolve to maximize parasite within-host fitness for a given set of parameter values. Initially gain in fitness is possible by increasing η. The additional transmission gained by longer parasite persistence through differentiation, outweighs the disadvantage caused by higher host mortality. When the block size increases further, density-dependence regulates the total parasite

population, specific immunity is suppressed, and all variants of the block can persist indefinitely. This reduces host survival but does not produce substantial gains in transmission probability. Consequently, within-host fitness is maximal at an intermediate



(a) Optimal block size

(b) Dynamics at the optimum

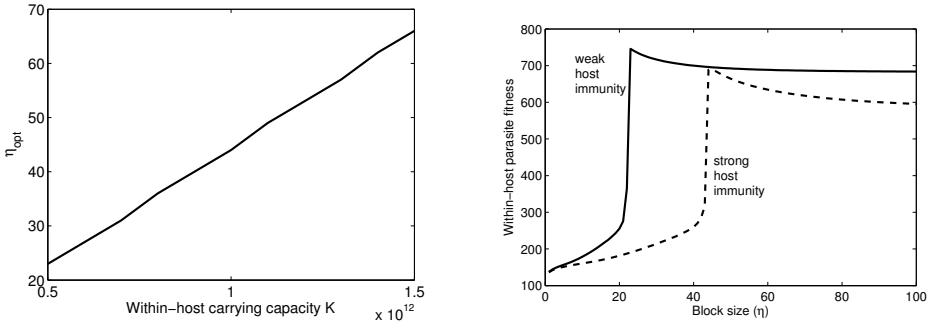
Figure 3.6: (a) Within-host parasite fitness F as a function of single block size η. An intermediate block size η_{opt} maximizes F, resolving the transmission-virulence trade-off. Within-host parameters as in Table 2.1 with: $C = K = 10^{12}, T = 3000, N_{blocks} = 1$. Epidemiological parameters: $\mu = 8.8 * 10^{-3}, z = 100, u = 10^{-5}$. (b) The infection dynamics for increasing block sizes (blue lines) and at the optimal block size ($\eta_{opt} = 44$), given in red. The optimum block size is achieved once the minimum level of stumpy cells needed for transmission is reached. Notice that η_{opt} in terms of parasite within-host fitness is close to the critical block size threshold approximation, η_{crit}, at the within-host level, as given by Eq. 2.22 ($\eta_{crit} = 36$).

value η_{opt}, which is similar in magnitude to η_{crit} (see Section 2.4.3), a parameter which divided block wave scenarios into two regimes: fast clearance and long persistence at quasi-steady state, as illustrated in Fig. 3.6(b).

Near η_{opt} the F curve is very sharp, due to the drastic qualitative change in within-host dynamics (Figure 3.6(a)). Small variations in η may have strong impact on the fitness of the parasite. Thus, the cost of expressing a block size above or below the optimum may be very large. This effect is related to the saturating nature of the transmission function, typical of vector-borne pathogens. As long as the number of parasites exceeds the transmission threshold, their exact number does not matter (Maudlin & Welburn, 1989; Paul *et al.*, 2007; Van den Bossche *et al.*, 2005). Importantly, for

small departures from η_{opt} exceeding the block size is better for the parasite than undercutting it.

Among several factors that may be involved in the evolution of single block size, the strength of variant-specific versus general parasite control in the host is the most crucial determinant at the within-host scale. Reduced parasite density-dependent differentiation, implies weaker general parasite control, thus stronger specific host immunity. The more immune-competent the host, the larger the parasite antigenic blocks it can control. Hence, a parasite adapting locally to a particular host, must increase the size of its antigenic blocks if this host is more competent in building up specific immunity (Figure 3.7(a)).



(a) Shifting optimum (b) Pathogen performance in two hosts

Figure 3.7: (a) Optimum block size increases linearly with within-host carrying capacity K. As K/C increases, specific immunity becomes relatively stronger and a higher number of variants is needed to overwhelm the host. Simulation parameters as in Table 2.1 with: $C = 10^{12}, T = 3000, N_{blocks} = 1$. (b) The same archive performs differently in two host types: weak immunity ($K/C = 0.5$), strong immunity ($K/C = 1$), the optimum η is larger in the second host, but the overall pathogen fitness is smaller.

This suggests that when transmission and virulence parameters are held constant, but there is heterogeneity in host immune-competence, a parasite trait under selection may be the size of its antigenic blocks. An increase in η may be driven by more immune-competent hosts. If such an optimal archive trait can be gained at no physiological cost, the parasite should evolve towards η_{opt}. However, if higher η comes with intrinsic physiological costs, the parasite may not be able to evolve towards the

optimal block size. In that sense, immune-competent hosts may well play a role in constraining single block size evolution and conferring sub-optimal fitness to the parasite. Moreover, a pathogen that has evolved one block size and then transfers to a second host, which is less immune-competent could be super-virulent (Figure 3.7(b)). The virulence in 'non-preferred' hosts could be driven by over-riding selection in the dominant host.

3.4.2 The optimal number of blocks

Compared to the single block size, the number of blocks (N_{blocks}) does not have such a drastic impact on the qualitative nature of the parasite dynamics in the host. When the parasite expresses many blocks of variants sequentially, new parasitaemia peaks develop, thus prolonging infection and the transmission time-window of the parasite. The lower the switch rates between blocks, the more decoupled the dynamics of each block and the greater the separation between subsequent peaks. We find that increasing the number of blocks increases within-host parasite fitness, up to the point when generation of further blocks of variants does not produce any additional gains in transmission, i.e. F tends to a constant.

Consider the case when $\varepsilon \ll 1$ and η is moderately small, thus the dynamics of each block are determined by their intrinsic rates of growth, differentiation and clearance, and there is minimal overlap between infection peaks. We can define the duration of a block wave D approximately as the time it takes the parasite to reach the carrying capacity K growing exponentially, plus the time delay it takes host immunity to begin its action, given by τ, and the time it takes host immunity to reduce parasite numbers below an extinction threshold: $D \approx 1/r \ln(K/V_0) + \tau + 1/d \ln(K/V_{ext})$ (Figure 2.5). This approximation assumes that the growth phase of a block is approximately exponential increase in parasite numbers up to the maximum K, that the non-growth phase lasts approximately τ units, and the block decline phase is a simple exponential decay with per capita rate d. Successful transmission to the vector depends on the concentration of stumpy cells in the average vector bloodmeal and the transmission threshold, thus the transmission window for a block of variants can be approximated by some fraction of block wave duration $\bar{\beta} < D$, equal for all blocks. Next, denote by t_B the time interval between two consecutive block waves. Since the blocks appear sequentially, host

survival after each block gets smaller. By the symmetry between blocks we have:

$$F(N_{blocks}) \approx \bar{\beta} \sum_{n=1}^{N_{blocks}} e^{-n[u(t_B+D)+\mu I]} \tag{3.7}$$

where $I = \int_{\text{Block wave}} \frac{V(t)+M(t)}{K_{max}} dt$ is the same for all blocks. Thus,

$$F(N_{blocks}) = \bar{\beta}\left[\frac{1-e^{-(N_{blocks}+1)[u(t_B+D)+\mu I]}}{1-e^{-[u(t_B+D)+\mu I]}} - 1\right] \to \frac{\bar{\beta}e^{-[u(t_B+D)+\mu I]}}{1-e^{-[u(t_B+D)+\mu I]}}, \tag{3.8}$$

as N_{blocks} increases. Notice that this upper limit decreases with the interval t_B between blocks and with the block wave duration D, thus we predict that within-host parameter changes such as longer immune delay τ, lower parasite clearance rates d, δ (increasing D), or lower between-block switch rates ε (increasing t_B), will make parasite fitness, F, saturate faster with the number of blocks and reach a lower final level. Similarly, changes in parameters that increase the total parasitaemia contained in a block wave, I, will act to reduce the parasite advantage of expressing more blocks over infection. Such changes could be a higher carrying capacity K or a higher antigen detection threshold C. Conversely, reducing I or D, thereby obtaining a relatively more competent host, will increase the benefit of expressing additional blocks. As shown in Figure 3.8, when within-host kinetics are held constant, F saturates faster in shorter-lived hosts (high u) and hosts experiencing higher pathogenesis (high μ).

The above observations seem to suggest paradoxically that the ability of the host to adapt in dealing with each block of variants through a strong and efficient immune response (lower block wave duration D, or lower pathogenesis experienced by each block), is detrimental to the host, providing selective pressure for the parasite to express more blocks in order to increase its fitness. A situation where immunity promotes virulence evolution has been described previously in a malaria model (Mackinnon & Read, 2004). It is likely that wide variations in natural lifespan and pathogenesis rates observed across host species in the range of the pathogen play a decisive role in the evolutionary dynamics of the number of blocks in the parasite antigenic archive. Stronger and longer-lived hosts will select for larger archives with a larger number of variant blocks. Ultimately, divergent selective pressures on this archive parameter (N_{blocks}) would have to be resolved depending on the relative densities of host types and their contribution to the transmission cycle of the pathogen.

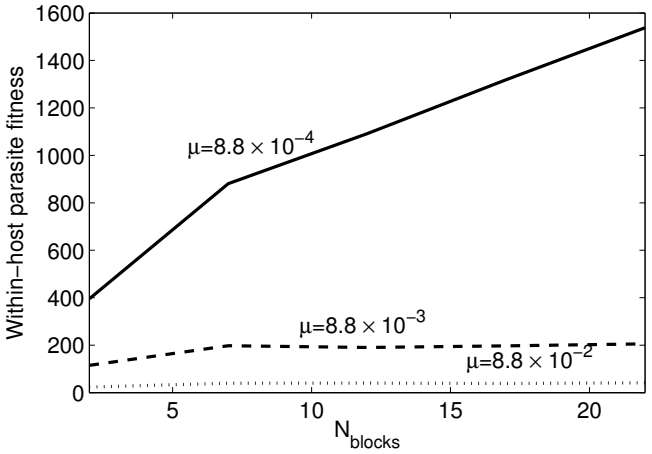


Figure 3.8: Within host fitness as a function of the number of archive blocks for different values of parasite-induced pathogenesis. Simulation parameters as in Table 2.1 with: $K = C = 10^{12}, T = 4000, \eta = 3$.

3.4.3 The optimal between-block switching rate

For ε, the between-/within-block switch ratio, we observe that reducing ε increases within-host parasite fitness, resulting in greater separation between two subsequent infection peaks, and thus, a smaller pathogenesis induced in the host. However, if ε gets too low, no new subsequent peaks can be generated and the parasite is cleared. This phenomenon described also in Chapter 2, and is related to demographic stochasticity in the switching process and the limited reproduction potential of each block wave. Depending on the type of switch matrix, whether hierarchical or non-hierarchical we have two cases: either there is a sharp drop in F when ε gets below the threshold required for stochastic variant generation (S non-hierarchical), or a gradual decrease (S hierarchical) toward the value of F obtained from expression of the first block, when the two blocks overlap. Thus, as in the case of single block size η, an optimal intermediate between-/within-block switch ratio exists, where within-host parasite fitness is maximized.

Clearly, the three parameters of the switch matrix: η, N_{blocks} and ε have a different qualitative effect on infection fitness and a different quantitative interaction with other within-host parameters such as K, C, τ and with more physiological parameters such

as maximal carrying capacity K_{max} and virulence μ. The focus of the next section is exploring more in detail precisely the latter interactions. So far, we have varied the archive parameters and seen their effect on F. Next, we compare discrete combinations of η and N_{blocks} and vary virulence and transmission parameters. We aim to understand the conditions that might select for one archive configuration versus another.

3.5 Infection scenarios and parasite fitness

The antigenic archive structure has a strong impact on within-host infection scenarios. As seen both in the previous chapter and in this chapter, the size of a single block of variants can modulate the balance between specific and general control of the pathogen, leading to persistence of all variants if their number exceeds a certain threshold. Thus, this archive parameter, η, mediates the transition between two infection scenarios: an oscillatory infection, with many parasitaemia waves, where each wave contains a small block of variants, and a stationary infection, where all the variants of a single large block continue to persist. A natural question is: which infection dynamics are better for the parasite? One where the antigenic block size is very large and the infection is stationary, or another, where the antigenic block size is smaller and the infection is oscillatory?

On the other hand, the archive parameter that determines how many parasitaemia waves will occur in an infection is the number of blocks, N_{blocks}. It is easy to imagine a situation where there is only one block of variants expressed, leading to an acute form of the infection, and a situation where the antigenic archive has more blocks, leading to chronic infection with more than one parasitaemia peaks. Previously, we saw that as the number of blocks increases the marginal gain in parasite within-host fitness tends to zero. However, it is not entirely clear how this effect changes when N_{blocks} varies together with the maximal parasite load or with the transmission threshold.

3.5.1 Oscillatory and stationary infection

To asses the impact of these two types of infection scenarios: oscillatory and stationary, on the parasite fitness within a host, we calculate parasite fitness from within-host

dynamics in two cases: one where the antigenic block size is large (stationary infection), and the other, where the antigenic block size is small (oscillatory infection). The relative success of these two scenarios depends on the pathogenesis induced in the host by the total parasite load.

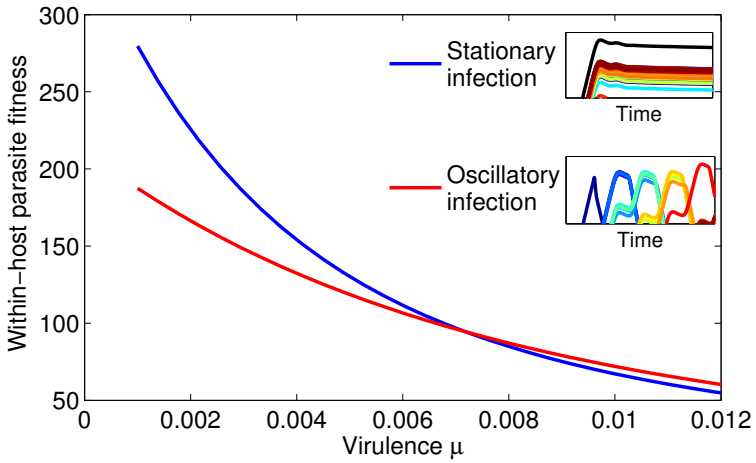


Figure 3.9: Within-host parasite fitness F as a function of parasite virulence for two infection scenarios. When the parasite is moderately virulent, the stationary parasite load, mediated by large antigen blocks ($\eta = 60$), is more advantageous than the oscillatory infection profile generated by small antigen blocks ($\eta = 5$). For higher levels of virulence, it is always better for the parasite to employ an antigen archive structure that gives rise to an oscillatory infection. The virulence level where both block sizes yield equal parasite fitness increases with the transmission threshold z.

When the per capita reduction in host survival by a single parasite (μ) is small, parasite fitness is considerably higher in the stationary infection scenario, despite the higher parasite load. Instead, if each parasite cell contributes to greater host mortality, parasite fitness is higher when the infection is oscillatory (Figure 3.9). Thus, all else being equal, reduction in host survival by one parasite cell influences the best antigenic archive strategy (i.e. single block size), to be used by the remainder of the parasite population. While an oscillatory infection is always better for the host, because the cumulative parasitaemia is smaller, an oscillatory infection is not always better for the

parasite. When smaller virulence and larger block size can be co-expressed as parasite traits, the parasite derives greater fitness from a persisting stationary infection than an oscillatory one.

Two different evolutionary outcomes may arise qualitatively, depending on the correlation between block size and virulence. If η and virulence are independent parasite traits, their evolutionary dynamics are decoupled. In that case, we might expect existence of parasite strains which specialize in one strategy or the other: either expressing enormous antigenic diversity at the start of infection, at the promise of being less harmful to their hosts (stationary parasite load), or expressing moderate antigenic diversity at any time, but being more virulent (oscillatory parasite load). If the two traits are correlated instead, virulence and block size will no longer evolve independently and only a limited number of parasite species may be observed.

Notice that the per-capita virulence parameter μ in the nested model (Eq. 3.4) inevitably depends on properties of both the host and parasite, and is either referred to as parasite virulence when interest focuses on the parasite, or host tolerance, when interest focuses on the host. The above considerations, seen from the perspective of host tolerance, imply that strains found in trypano-tolerant hosts will be more likely to express very large antigenic diversity at any time during infection and to cause persistent parasitaemia. In contrast, trypanosome strains found in trypano-sensitive hosts are more likely to exhibit smaller antigenic diversity at any time and to give rise to an oscillating parasite load that constrains the overall harm done to the host. Comparison of longitudinal infection dynamics across different host types is needed to validate this model prediction against data, and to check whether host-adapted pathogen strains are really optimizing within-host infection fitness via specific antigenic variation strategies.

3.5.2 Acute and chronic infection

The archive parameter N_{blocks}, appearing in the construction of the switch matrix, mediates the length of infection. We consider two cases: 1) infection with a parasite having a small archive of just one block of variants (acute), 2) infection with a parasite having a larger archive of 4 blocks of variants, leading to more parasitaemia peaks and longer infection duration (chronic).

One interesting observation is that the length of infection impacts the sensitivity of within-host pathogen fitness to different model parameters. All else being equal, in the acute infection case, parasite success F is more sensitive to the within-host carrying capacity K than to the virulence parameter μ. This is due to the fixed transmission threshold assumed in the host-vector interaction. If K is too low relative to the total blood volume of the host (Ω), no successful transmission can occur, as the stumpy cell concentration per tsetse bloodmeal is insufficient to infect the vector. In contrast, when infections are chronic, parasite success F is more sensitive to parasite virulence, which is a parameter that has instantaneous impact and determines host harm over time.

Varying the transmission threshold, in Figure 3.10 we compare two hypothetical trypanosome strains: one that causes a severe acute infection (1 antigenic block) and differentiates slowly (K high) and another strain that causes a mild chronic infection (4 antigenic blocks) but differentiates faster (K low). The same maximal carrying capacity K_{max} and host epidemiological parameters are assumed for the two scenarios. Notice that depending on the transmission threshold, one strain does better than the other in terms of within-host fitness. When the transmission threshold is low, the strain causing mild chronicity is more successful, instead, when the transmission threshold is higher, the strain causing a severe acute infection is more successful. For a particular transmission threshold, both strains can do equally well.

This means that distinct combinations of differentiation and antigenic variation traits can be adopted by the parasite leading to the same transmission success. Seen in the context of trypanosome epidemiology, this observation suggests that perhaps the two main T.brucei strains, *T.b. rhodesiense* and *T.b. gambiense* (Section 1.2.1), reflect differences in antigenic variation and differentiation strategies of the parasite, i.e. lower N_{blocks} and slower differentiation in *T.b. rhodesiense* causing an acute form of the infectious disease, and higher N_{blocks} and faster differentiation in *T.b. gambiense* leading to milder chronic infections. A detailed empirical comparison of these strains is needed to test this proposition.

Such outcome suggests that in addition to mammalian host life-history traits (e.g. immune competence, trypano-tolerance, etc.) an important environmental factor influencing parasite life-history traits is the vector and its susceptibility to infection (transmission threshold). Natural selection is expected to optimize the allocation of resources between different processes, so as to maximize parasite fitness, assuming

sufficient genetic variation exists to allow the optimum to be reached. If differentiation at higher parasite loads (high K) is costly to the parasite in terms of the biological machinery needed, and if structured antigenic variation (many blocks) similarly requires investment of resources, then what the model shows is that these two processes can be traded-off against each other and yield the same overall success to the parasite.

Fundamental to the evolution of parasite phenotypes, such as antigenic variation strategies, is thus the interaction between constraints at different points of the transmission cycle. Adaptation may occur not only in relation to a specific mammalian host, but also in relation to a specific vector. Inevitably, which infection profile is optimal in a given host will depend on properties of the host itself and also on properties of the associated vector.

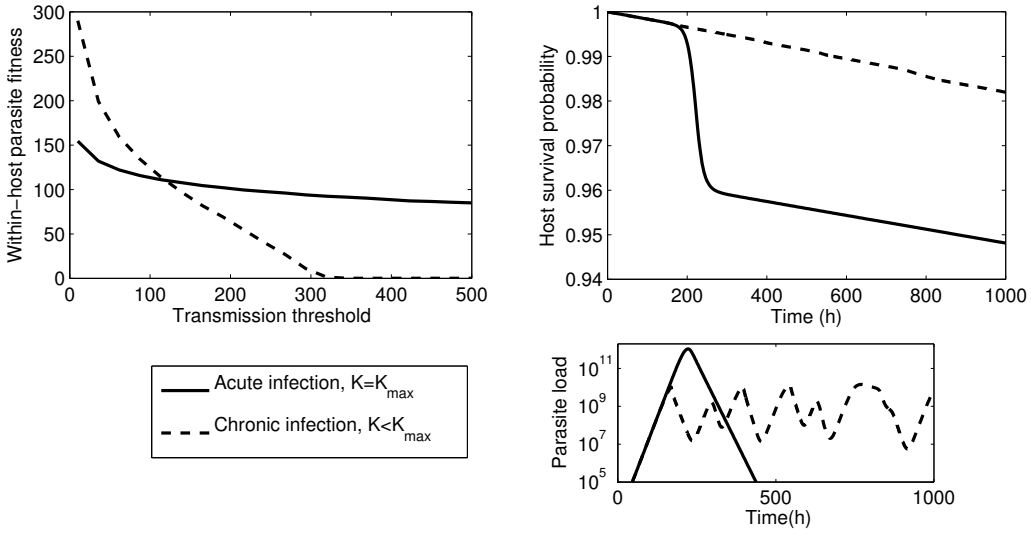


Figure 3.10: Severe acute vs. mild chronic infection. The transmission success depends on the threshold number of stumpy cells needed to infect the tsetse. When the transmission threshold is low, a pathogen strain differentiating at higher parasite loads and expressing a single block of variants obtains higher infection fitness than a strain differentiating faster and expressing many antigen blocks.

3.6 Optima across different hosts

So far we have considered hypothetical hosts in our analysis. In this section, we combine changes in K with realistic changes in host parameters, and investigate how parasite fitness in typical hosts depends on the antigenic archive employed by the parasite. In the calculation of within-host fitness F, we expect parameters such as total host blood volume Ω, maximal carrying capacity K_{max}, within-host carrying capacity K, host natural mortality rate u, and parasite virulence μ to vary across hosts. The transmission threshold, by its dependence on the vector, can be assumed to be generally host-independent.

To understand the full function of antigenic variation strategies of parasites, we must view this parasite trait in the context of hosts that typically get infected in the field. A pervasive question throughout this chapter has been: how does the parasite within-host fitness obtained in different hosts vary depending on the antigenic archive? Here, we address this question in the context of realistic variations in host features. Is there an optimal host for a fixed antigenic archive structure? If yes, what are the characteristics of this optimal host? Can they help explain the current variant repertoire found in trypanosomes? To answer these questions, we begin by giving a simple illustration of parasite fitness in two very different trypanosome hosts: the mouse and the cow, one used in the laboratory, the other prevalent in the field. Then, we propose a new way of looking at within-host fitness across hosts by invoking allometric scaling between different model parameters.

3.6.1 A simple illustration

As described in Chapter 1, for trypanosomes, an important class of host reservoir in the field are domestic cattle and other livestock. For simplicity, we consider the cow as a representative host in the field and the mouse as a representative host in the laboratory. These two hosts have body sizes differing over orders of magnitude. Some of the epidemiological parameters across the two hosts can be found in the literature, others remain largely unknown, such as the host mortality induced per unit of time by each parasite cell occupying one unit of the maximal carrying capacity (μ), typically hard to quantify. The values we use are listed in Table 3.1. Although there are mouse

species that are resistent to infections and display higher rates of survival, most mice experience high rates of infection-induced pathogenesis. Similarly for cattle, although there are some breeds that are more sensitive to infection, trypano-tolerant cattle such as N'Dama play an important role in transmission in affected areas.

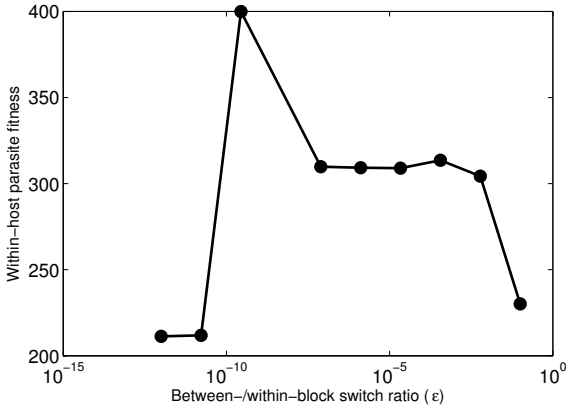
Table 3.1: Estimated epidemiological parameters for two host species.

Parameter	Mouse	Cow
Weight	0.02 kg	500 kg
Blood volume (Ω)	0.002 l	20 l
Life span ($1/u$)	1 year	7 years
Disease-induced mortality (μ)	8.8×10^{-3}	8.8×10^{-4}
Tsetse blood meal (k)	0.005 ml	0.005 ml
Stumpy cell number threshold (z)	100	100

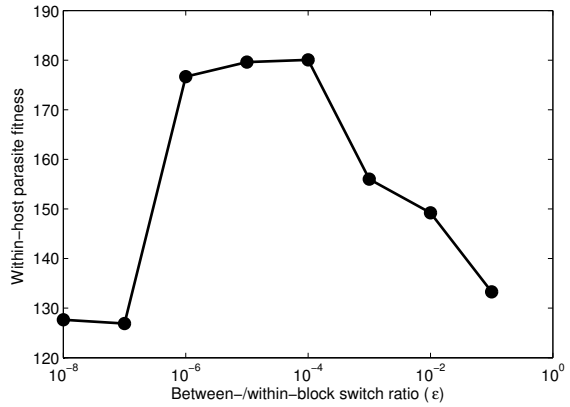
We assume that across the two host species, mouse and cow, the quality of the specific immune response against antigenic variants is the same. This is represented in the ratio C/K and the parameters c, x, d, δ which are host-invariant. Instead, for parameters such as the carrying capacity K, we assume that it increases with host size and $K = K_{max} \sim \Omega$. From empirical infection data it can be seen that mice experience higher rates of infection-induced pathogenesis than cows (Barry, 1986). We roughly approximated μ from data in (Radwanska *et al.*, 2008) and (Barry, 1986). The average natural lifespan of the two hosts also shows marked differences.

In light of the results of Section 3.3, we find that for a fixed antigenic archive, the overall magnitude of within-host pathogen fitness, F, is smaller in the smaller size host. This is unsurprising, given the higher rate of pathogenesis the mouse experiences as a result of infection, and the fact that the maximal pathogenesis is directly related to the within-host carrying capacity. Interestingly, this difference in pathogen fitness is reduced if the strength of within-host specific immunity relative to parasite differentiation (e.g. K/C) increases. In that case, the host body size (thus, K) is less relevant

in determining maximal pathogenesis because the peak parasite load is controlled by host immunity as opposed to parasite differentiation.



(a) Cow optimum



(b) Mouse optimum

Figure 3.11: The optimum between-block switch rate is smaller in a large size host than in a small size one, but the overall parasite fitness is higher. Parameters as in Table 2.1 with: $C = K, T = 3000, N_{blocks} = 2$. (a) $K = 10^{12}$, (b) $K = 10^{8}$.

For these two hosts, numerical observations suggest that the optimum block size, η_{opt}, is independent of host size, if parasite-immune system kinetics parameters (e.g. $K/C, d, \delta$) are assumed invariant. The effect of antigenic block size on parasite within-host fitness depends primarily on the immune competence of the host rather than the absolute value of the within-host carrying capacity. This implies an immune competent

cow and an immune competent mouse will select for the same size of archive blocks, despite differences in blood volume and pathogenesis rates.

However, as shown previously, the archive parameters that control the duration of infection such as between-block separation, ε, and N_{blocks} are likely to show dependence upon within-host carrying capacity and parasite virulence. We observe that the optimal between-block connectivity in a small size host, such as a mouse, experiencing both lower within-host carrying capacity and higher virulence per parasite cell, is higher than the optimal between-block connectivity in a large size host, such as a cow (Figure 3.11). This implies that variants that can be easily generated stochastically in the larger host, may appear much later in a small-size host or may never arise at all over infection. Similarly, the number of blocks that increase pathogen within-host fitness saturates faster in the mouse than in the cow, as the expected lifespan of the infected mouse is lower than the expected lifespan of the infected cow. These results illustrate that the intrinsic coupling between different host physiological parameters can have a significant impact on the within-host fitness of the parasite and the evolution of its antigenic archive traits.

3.6.2 Towards a general scaling theory

Sections 3.4-3.5 suggest that within-host parasite fitness is highly dependent on the interaction between parasite antigenic archive characteristics (η, N_{blocks}, ε) and host characteristics, such as immune-competence, pathogenesis rate, natural lifespan and carrying capacity. So far, we have considered changes in these host characteristics in an isolated manner, largely neglecting possible systematic correlations between them. However, a more realistic scenario is obtained when these characteristics are interdependent or directly linked to a host trait. Size is perhaps the most crucial host trait, as many important physiological characteristics such as metabolic rate, complexity, body temperature, life span, and strength scale allometrically with body size (Peters, 1983). Larger hosts are expected to have longer life expectancy, smaller reproductive rate, slower dynamics and lower population densities when compared to smaller host species.

Allometric scaling has recently been invoked to describe certain quantitative aspects of between-host infection dynamics (Bolzoni *et al.*, 2008; Dobson, 2004). Here,

we discuss some implications of allometric theory on the within-host dynamics of antigenically varying pathogens, such as the African trypanosome. Clearly, systematic differences of orders of magnitude in mass and blood volume across mammalian host species, must lead to similar differences in their respective parasite carrying capacities, and consequently in the opportunities or challenges they provide for parasite growth and antigenic variation. Indeed, an early comparative study (Barry, 1986) has shown that the rate of trypanosome antigenic turnover for the early VSG group may be linked to the within-host carrying capacity across a group of host species. Furthermore, these differences in carrying capacity are likely to be associated with parallel changes in epidemiological parameters affecting transmission between hosts.

Cable *et al.* (2007) have found an allometric relationship for the rates of pathogenesis and the time to disease progression for a set of pathogens in different host species. Similarly it has been argued that quantitative processes of the cellular immune responses are affected by allometric scaling. General principles on how the number of naive T-cells scales with body size have been derived (Perelson & Wiegel, 2004). Further, lymphocyte trafficking, the circulation of T-cells through the blood, tissues and the lymphatic system have also been suggested to obey general scaling laws (Perelson & Wiegel, 2009).

Using the modeling framework developed in this chapter, we can study variation in host body size by altering some within-host and epidemiological parameters, thereby exploring the influence of host body size (W) on parasite antigenic archive evolution. In calculating within-host parasite fitness $F(W)$ (Eq.3.4), we assume: 1) the within-host carrying capacity scales as W; 2) natural host mortality scales as $W^{-1/4}$; 3) the per-capita increase in host mortality from one parasite cell scales also as $W^{-1/4}$; 4) other kinetics parameters, including immune response and parasite intrinsic growth rate are host size-invariant (Cable *et al.*, 2007; Perelson & Wiegel, 2004).

The allometrically scaled model reveals that trypanosome within-host fitness increases initially as a function of host body size, and then decreases (Figure 3.12). This can be explained by the fact that infections established in hosts of larger size are likely to be associated with lower pathogenesis rates and better host survival, thus increasing the chances of parasite transmission. However, if the host size increases further, the corresponding increase in carrying capacity leads to fast antigen turnover, which implies the archive is expressed too quickly, shortening the overall infection duration.

This reduction in the transmission window of the parasite is apparently larger than the benefit obtained from lower pathogenesis in larger hosts, therefore within-host parasite fitness decreases. Thus, there is an optimal intermediate host size, in which a given archive is best exploited. The interplay between within-host and between-host constraints seems to be crucial in determining this optimum.

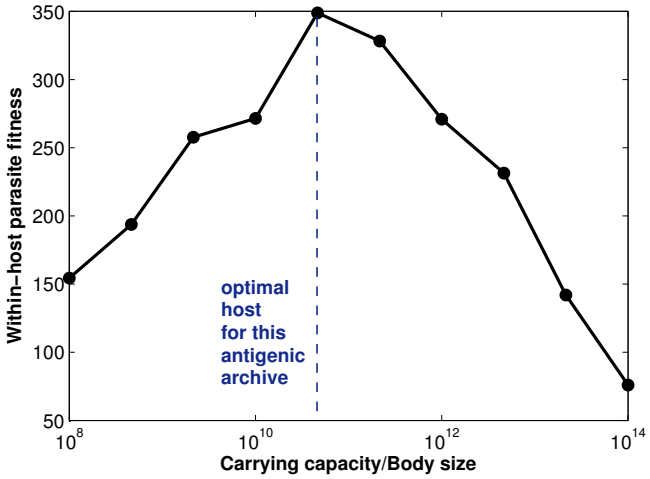


Figure 3.12: Allometric scaling reveals that the within-host parasite performance first increases with K ($\sim$ host body size) and then decreases. There is an optimal host size where a given archive is maximally exploited without damaging substantially host survival. Simulation parameters as in Table 2.1 with: $C = K, T = 10000$. Switch matrix non-hierarchical, $\eta = 5, N_{blocks} = 5, \varepsilon = 0.01$.

Our results further confirm that the sensitivity of within-host fitness to changes in N_{blocks} and ε increases systematically with host body size, suggesting that the evolutionary dynamics of these parasite traits will be strongly coupled to the host trait of body size. If one assumes that specific host immune-competence does not change with host size, as the kinetics of processes in immune systems are scale-invariant, the sensitivity of F to the size of a single block, η, in the parasite antigenic archive will remain generally host size independent.

We suggest that there may be at least two ways for the parasite to evolve the antigenic archive in order to better adapt to large size hosts, namely increasing the number of blocks of variants and increasing the separation between them. A plausible expla-

nation for the size and modular structure of the modern trypanosome antigenic archive may partly be the large selection pressure for these archive traits, coming typically from large size hosts (e.g. cattle and livestock), dominant in the transmission cycle of this parasite. Indeed, the trypanosome archive vastly exceeds that of other parasites, particularly when the added effect of combinatorial creation of mosaic and expressed genes is considered. The exceptionally broad trypanosome host range might provide an explanation.

Conceivably, another positive effect of constant antigenic archive expansion emerges when the parasite reinfects partially-immune hosts. Since bigger hosts will have been exposed to a larger proportion of the archive over any single infection, upon re-infection, expression of antigens dissimilar to these types is highly advantageous to the parasite. Thus bigger hosts, by displaying a larger extent of relevant immune memory, will select for more antigenic variability of the pathogen. Previous research has already linked the divergent selective pressures from population classes of different immune memory to some variability observed across pathogens in age-class bias, reproductive rate, and antigenic variation (Frank & Bush, 2007). In analogy, the divergent selective pressures coming from hosts of different size may partly explain eventual differences between trypanosome strains in host-bias and antigenic variation strategies.

3.7 In the field: global archive optimality vs. plasticity

Despite many simplifications, our modelling has revealed that different traits of the host population can act as evolutionary drivers for different aspects of the antigenic archive, reinforcing once more the importance of host heterogeneity in the evolutionary dynamics of parasites (Regoes *et al.*, 2000). Characteristics such as within-host carrying capacity and tolerance to infection may have a bearing on the evolution of archive modularity in terms of number of antigen blocks and their intra-connectivity. Instead, characteristics such as specific immune competence will influence the evolution of single block sizes.

The magnitudes of these different selection pressures will inevitably depend on host demography such as relative abundances between host types, host susceptibilities, and ecological factors such as contact rates with the vector, which have been shown to affect the evolution of multi-host pathogens (Gandon, 2004; Gandon *et al.*, 2002).

These ecological determinants will influence whether the global optimum, where parasite fitness in the field R_0 is maximized, is closer to the local optimum in one host type or another.

To illustrate the effect of host demography, we consider a simple case (see Figure 3.13), where the host community is composed of immune-competent and less competent hosts (variable K/C) of the same species (i.e. all other epidemiological parameters are equal). This composition is pivotal for the evolution of the optimal block size in the antigenic archive of the parasite. Indeed, when a larger proportion of the host population consists of immune-competent hosts, the globally optimal block size is closer to the local optimum in these hosts, and as a result, the ultimate magnitude of the maximal basic reproductive ratio is smaller, in fact less than 1, which ensures disease extinction in the long run.

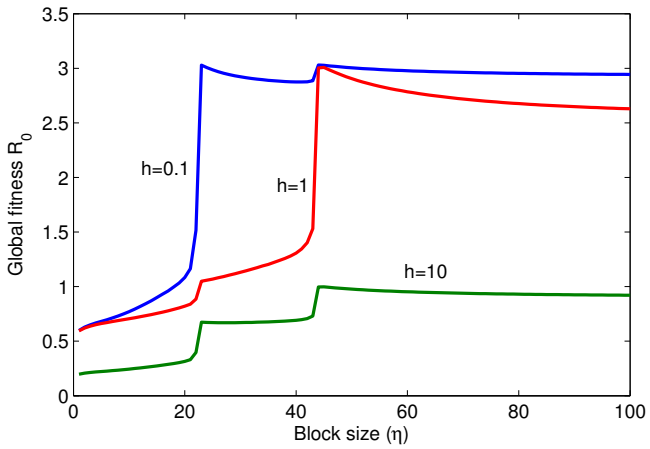


Figure 3.13: Global parasite fitness R_0 as a function of the abundance ratio h between two host types of: weak ($K/C = 0.5$) and strong immune-competence ($K/C = 1$). The two hosts have the same epidemiological and ecological features otherwise (ref. cow in Table 3.1). Simulation parameters as in Table 2.1, with $C = 10^{12}, T = 1000, N_{blocks} = 1$. For the parameters in R_0: $Z = 50, \alpha_H = 0.1, \gamma = 7 \times 10^{-4}, b = 0.005, H_1 + H_2 = 500$.

In addition, although neglected in this study, many other factors such as the coevolution of the mammalian host population and vector population is likely to play a very important role in the nature and speed of the molecular evolution of the pathogen.

One must however not forget that parasites have very short generation times, thus they may evolve much faster than their hosts. One possible response of multi-host parasites is to evolve a compromise life-history, one that produces the best average response to all environmental conditions, i.e. is globally optimal across hosts. Another possibility is that the organism evolves a plastic response: in each environment (host) the pathogen displays a different life-history suited to that environment. Whether a fixed or plastic response is optimal depends on the costs and benefits of sensing and regulation versus the benefits of plasticity. A plastic archive would require easy genetic control of switching off, activation and mutual pathways between different VSGs, as well as a biological machinery to distinguish between different host types (tolerance, immune competence, size, age, immune memory, etc). At the moment, the link between genetic mechanisms operating at the molecular level and the antigen switching process is not sufficiently understood (but see Appendix B.1), at least not to a substantial quantitative level. Therefore, archive plasticity remains a speculation, however an interesting one. More evidence is also needed about trypanosome-host adaptation, in terms of possible host-specific replication characteristics of the parasite or particular gene expression, in order to establish whether the pathogen can distinguish between different host-types in the first place, and what are the costs and benefits of that recognition.

3.8 Antigenic variation and parasite genetics

Up to now we have explored the mechanistic links between antigenic variation, within-host dynamics and parasite transmission. A natural extension comes from the need to build mechanistic frameworks that bridge between the within-host level and the genetic level of the parasite. Nested modelling has a young but robust history of connecting the within-host and the between-host level, to provide insight into the evolutionary dynamics of pathogens. The next step is to construct nested frameworks that take into account genetic processes.

In these last two chapters, we have demonstrated the various ways in which the antigenic archive interacts with other within-host parameters, the effects of such interactions on parasite fitness across biological scales, and the resulting selection pressures toward optimal structures. Clearly, structural properties of antigenic archives arise

as direct consequences of genetic mechanisms and processes acting at the molecular level. The ultimate recipient of population-level feedbacks is the genome of the parasite. Only via genetic processes can optimal archive configurations, optimal switching rates, and groupings between variants be achieved.

Across vector-borne pathogens that display antigenic variation, preferential expression of certain variants over others is mediated by genetic factors that govern *var* gene switching, in addition to other non-genetic factors that may play a role, such as immune-selection in the case of the malaria parasite. Despite differences in the size of the repertoire, variant gene organization in such pathogens displays a common hierarchical structure, grouping together genes that easily switch to each other and appear under the same peak during infection (Morrison *et al.*, 2009; Recker *et al.*, 2011). Despite extensive research efforts (Barbour *et al.*, 2006; Barry, 1997; Barry & McCulloch, 2001; Borst *et al.*, 1997; Donelson, 1995; Kyes *et al.*, 2007; Morrison *et al.*, 2005), the mechanisms driving antigenic diversity and switching in these pathogens remain only partially understood. However, with the increasing resolution of the genetic structure of these archives, it becomes crucial to link this understanding mechanistically with the within-host pathogen population dynamics and between-host transmission.

In Appendix B.1, we propose one such general integrative framework: a gravity model, by which each switch rate between two antigenic variants can be specified on the basis of their genetic properties. The set of switch rates can then be assembled into a switch matrix that becomes one representation of the archive structure. Subsequently, at the level of the archive, the emerging magnitude configuration of the switch rates can take multiple forms (Frank, 1999; Lythgoe *et al.*, 2007), a special case being the block structure considered so far, where within-block variant switching is higher than between antigen blocks. An important factor in determining switch rates between different variants, especially in the mosaic phase of infection is genetic identity between the two sequences (Marcello & Barry, 2007a). The processes governing changes in genetic identity between silent genes in the VSG archive will be treated in detail in Chapters 4 and 5.

3.9 Discussion

In this chapter, we have nested a within-host parasite dynamics model into an epidemiological framework to investigate the influence of trypanosome life-history traits on its transmission success. Despite the difficulties involved in this approach, by using a small number of plausible assumptions on host species-invariant infection kinetics, the ecology of vector-host interaction, and the nature of parasite-induced host pathology, we gained valuable insight into the evolution of this parasite. While the behaviour of the within-host model and the evolutionary dynamics investigated depend on our assumptions, the mechanistic framework used for this analysis does not. Changing aspects of within-host dynamics, scaling across host species and fitness definition would necessarily change the resulting conclusions. Hence, while our findings are in some sense limited by our model assumptions, the approach linking parasite genetics to parasite fitness can be applied to other host-parasite systems including any level of biological detail.

Our analysis shows that an important measure of within-host parasite fitness can be derived from the infectivity of the single host over all the infection period, which is very sensitive to the antigenic archive of the trypanosome. Previous models of within-host antigenic variation have taken the total parasitaemia as an index of parasite success (Frank, 1999) and have shown that modular switching pathways between antigenic variants maximize parasite success. We have used a slightly more refined index of success, taking into account the saturating nature of vector-borne transmission and the role played by stumpy cells. Many infection characteristics such as peak parasite load, duration and amplitude of oscillations depend on the parasite antigenic archive structure. Furthermore, the interplay of this structure with host ecological and epidemiological parameters influences significantly within-host pathogen fitness, and consequently the overall basic reproduction number R_0.

Considering systematically discrete changes in the archive structure, we find that: 1) parasite within-host fitness is maximized at an intermediate archive block size, which changes with the specific-general parasite control balance; 2) despite more blocks initially leading to longer infection and more transmission, parasite success in a given host becomes insensitive to the number of archive blocks in the long run,

due to parasite-induced host mortality; 3) increasing archive modularity is more advantageous in trypano-tolerant hosts and hosts with longer lifespan, thus we expect these are the types of hosts exerting a strong selective pressure on parasite archive expansion; 4) there exists an intermediate rate of switching between blocks, dependent on the within-host carrying capacity, that optimizes within-host parasite fitness.

The above trends are independent of host body size. However, when characteristics such as host immune competence, parasite-induced host mortality and carrying capacity depend on host body size, then optimization of the archive structure can be a direct consequence of this primary host trait. More specifically, archive parameters such as the number of blocks and between-block connectivity can be expected to depend on K, while the size of an arbitrary block can be considered as K-independent.

We used allometric scaling to link host body size to epidemiological and ecological host factors. Our findings reveal that for a given antigenic archive, within-host parasite fitness is maximized at an intermediate host body size. Such optimum arises largely from the increase of K with body size, which accelerates antigen turnover within a host, and the saturating nature of vector-mediated transmission, which makes increases in parasite load futile after a certain threshold. Uncovering the biological mechanisms that govern K and the rate of differentiation, whether parasite- or host-intrinsic, thus becomes mandatory, if we are to envision control strategies for trypanosome infections. The induction of differentiation at lower parasite loads, via drugs that target the parasite or trigger host action, emerges as a possibility that may well constrain the reproduction potential of the pathogen and limit infection duration and host damage.

3.9.1 Outlook

In summary, our models revealed that there are larger-scale consequences of parasite genetic architecture that influence within-host and between-host population dynamics. Indeed, the selection forces that shape the parasite genome are likely to be determined primarily by these larger scale processes. The optimal structure of the antigenic archive is likely to be selected for on the basis of between-host dynamics and the ecological and epidemiological factors driving it. A crucial goal is to understand how various mechanistic components determine complex aspects of host-parasite interac-

tion at these higher levels, and how evolutionary processes in turn have shaped and are shaping the underlying mechanisms.

An important aspect of the formulation of within-host infection fitness was the assumption that host pathogenesis was a function of the total parasite load, indirectly dependent on the antigenic archive of the pathogen. A more realistic situation would be to assume that in addition, host pathogenesis is due also to host immune activation. In fact, much research in immunology seems to point in this direction, and even assign host immunity the greatest part in infection-induced pathogenesis (Graham *et al.*, 2005). For example, the amount of variant-specific immune memory that is built up during infection and later maintained may require a huge investment of host vital resources or may induce immunopathology, two effects that could substantially reduce host survival, and thus play a role in the trade-off between transmission and virulence. Precise quantification of pathogen virulence effects, and host tolerance effects, both in relation to specific life-stages of the parasite or particular variants would help make the current nested framework both more biologically realistic and biologically applicable.

Another possible extension in the exploration of between-host dynamics could be to use two additional allometric scalings in R_0: of host-vector contact rates and host abundances. If host-vector contact rates depend roughly on the surface area of the body of the host, they would scale allometrically with host size as $W^{2/3}$. Species abundance has been suggested to obey an allometric scaling law with power exponent equal to -1 (Peters, 1983). These relationships would have a direct bearing on the relative contributions of different host species on parasite fitness, and would provide further insight into the ecological mechanisms driving pathogen evolutionary dynamics.

Ultimately, the results of all these evolutionary processes lie encrypted within the contemporary parasite genomes we study today. In this post-genomic era, we are able to deduce details of parasite genomic architecture that govern the generation of antigenic novelty within and between hosts, and the factors possibly constraining this process at the genetic level. It might be that certain antigenic archive traits are too hard to change genetically, and that alternative life-history traits, such as reproductive restraint (Reece *et al.*, 2010), or sensitivity to differentiation (Mathews, 2011), could evolve instead in order to yield equivalent fitness benefits to the parasite. The difficulty is that we do not know what life history alternatives are available to pathogens, and unless those options are known, prediction of their plastic or evolutionary responses is hard.

Increasing knowledge of parasite genetic details, however should enable integration of population dynamics across different biological scales, and unveil the selective pressures on the structure of antigenic archives and the way they are used. Uncovering the inter-dependencies across different levels of biological organization remains key to understanding how parasites have evolved to be the way they are, and what control strategies are ecologically and evolutionarily sustainable.

In the next chapters, we consider in more detail some of the evolutionary processes responsible for the generation, maintenance and access of genetic diversity in the antigenic archive of African trypanosomes.

Chapter 4

Quantifying local VSG diversification using hidden Markov models

4.1 Introduction

The last two decades have seen a considerable increase in the understanding of the structure and organisation of the VSG archive of trypanosomes, comprising more than 1600 member genes, and molecular mechanisms involved therein (Barry, 1997; Berriman *et al.*, 2005; Borst *et al.*, 1997; Marcello & Barry, 2007b). As shown by Marcello & Barry (2007a), about 60% of VSGs are unique, the rest occur in subfamilies of two to four close homologs ($>$ 50% peptide identity). Generally during an infection, intact array genes are activated by duplication after two weeks, and mosaic VSGs assembled from pseudogenes start to be expressed by week three and dominate by week four (Morrison *et al.*, 2009; Thon *et al.*, 1990). The small subfamily structure of the archive seems to be fundamental in providing the interacting donors for mosaic formation (Marcello & Barry, 2007a), and thus for the maintenance of chronic infection through antigenic variation. The predominant paradigm for explaining the variation between antigen genes and their huge number has been diversifying selection acting on these genes, an important evolutionary process that poses the major barrier to disease control, as demonstrated by the difficulties associated with vaccine development against antigenically diverse pathogens.

The VSG antigenic archive of African trypanosomes is an example of a multi-gene family, which generally represent genomic regions that include multiple similar copies

in close proximity, generated by duplication from a common ancestor gene. Across species, multi-gene families or gene clusters are of special interest because of their genetic and molecular importance. In humans, they are often involved in diseases having a genetic component, such as cancer and immune system disorders. In pathogens, multi-gene families are often involved in immune evasion processes. To understand how these gene clusters are implicated in biological function, it is helpful to examine their evolutionary histories.

In this context, the architecture of the trypanosome VSG archive offers an important opportunity to study the evolution, maintenance, and function of repositories of antigenic diversity, so fundamental to the life-history strategies of protozoan pathogens. The two most significant factors driving subsequent change after gene duplication, in terms of evolutionary adaptation and diversification, are point mutation and gene conversion (Ohta, 2010). Gene conversion is the process by which one participant in a recombination event copies a short tract of DNA from a donor partner, resulting in a mosaic of DNA from different ancestors. Point mutation instead is the process of random nucleotide substitution on any given sequence. The ratio between these two processes differs largely among organisms; generally recombination among subtypes is rarer than point mutation. The action of these two processes on the VSG archive of trypanosomes is considered to be the main driver for changes in pairwise identity between loci encoding important surface antigens, but their precise ratio is not yet known.

A natural question is: what are the parameter combinations of these constituent processes that shape the distribution of pairwise homologies in the silent archive? Across the VSG archive, gene conversion can lead to homogenization, thus preventing genetic divergence and leading to evolutionary conservation of the genes. In contrast, mutation diversifies the members of the multi-gene family. However, locally, gene conversion may act as a promoter of genetic diversity, by bringing in new genetic material from the global pool into small subfamilies of recently duplicated VSG genes, and presumably, at the protein level by altering structure. Such events have the potential to accelerate gene diversification within subfamilies, and spread any beneficial mutations acquired by older VSG genes across all members of the archive.

4.1.1 Quantifying evolutionary processes

The advent of molecular genetic techniques, in particular high-throughput nucleotide sequence determination, has made most of the genetic variation present in the trypanosome VSG genes accessible for characterization. Thus, there is a significant amount of data on pairwise genetic relatedness between any two genes, resulting from their long-term evolutionary dynamics. This data is open for investigation and interpretation, thus allowing the quantification of key evolutionary processes through modern inference frameworks.

However, the analysis of multi-gene families poses significant computational challenges. One of the major difficulties is the process of gene conversion among the duplicated regions of the family, which can obscure their true relationships (Song *et al.*, 2011). Constructing a phylogenetic tree or a multiple sequence alignment is the the most frequent first step in analyzing multi-gene family evolution. Both of these methods assume that all the positions in the duplicated copy will display similar divergences from the original segment, so we expect a single phylogeny for a given set of DNA sequences and similarly, a single multiple alignment between them. However, current tree construction approaches give rise to different tree topologies depending on which part of the duplicated segment is taken as the input data, whereas multiple-sequence alignment tools sometimes align non-orthologous parts of the sequences. Conversion events, by producing mosaic DNA segments with varying divergences from the common ancestor are difficult to detect.

There are currently many computational methods for detecting gene conversion across a set of genetic sequences (Song *et al.*, 2011), among which most prominently, phylogenetic-based methods (e.g. recHMM by Westesson & Holmes (2009)) that identify gene conversions by finding breakpoints that change the tree topology, and similarity-based methods, which search for segments of unusually high similarity within two homologous sequences in the set (e.g. GeneConv by Sawyer (1989)). The evaluation of such methods is commonly performed using simulation data with varying levels of recombination, genetic diversity and mutation rate, but most importantly using real datasets, where the "true" conversions are already known. These approaches are powerful in detecting recombination events, especially when the potential set of donor sequences is known, when the recombination tracts are long and the amount

of data is large. However, they are usually of a less-parametric nature, thus making it difficult to quantify dynamic rates that can be explicitly linked to the mechanistic processes involved.

In this chapter, we concern ourselves with the generation of diversity within closely related members of the VSG gene family in African Trypanosomes. In particular, we study the interplay between point mutation and gene conversion at the scale of high-identity gene subfamilies in the VSG archive. These subfamilies arise primarily via gene duplication, which is subsequently followed by gene diversification. To assess the impact of the two most important processes involved in gene diversification, we develop a mathematical model that reflects the genetic dynamics of mutation and gene conversion on aligned gene pairs, neglecting other more minor genetic processes.

Our aim is to model the concurrent processes of point mutation and gene conversion, relying entirely on patterns of genetic identity and mismatches between two aligned sequences in a gene subfamily. In contrast to other methods, we don't use any explicit information about donor genes that have contributed the genetic material through the conversion events, although we can infer properties of these donor genes. We are interested only indirectly in detecting the locations of conversion events. Rather than replacing current approaches, we see our method as a potentially valuable tool for generating new insight into the relative roles of these evolutionary processes and for parameterizing the mechanisms that underlie them.

Our hypothesis is that mutations occur randomly and homogeneously on each gene, whereby the mismatches introduced through point mutation on their pairwise alignment follow a Bernoulli process. Each aligned nucleotide can thus be considered to mutate with an associated given probability. The mismatches introduced through gene conversion, instead are more likely to occur in clusters. We assume first that the most relevant conversions occur between a member of the pair and a gene fragment outside the subfamily, bringing in new genetic material. Secondly, the clustering of mismatches within converted tracts occurs because donor genes outside the high-identity gene subfamily have been evolving and diverging for a longer time, thus on average each of them contributes a fragment with a higher density of mismatches to the gene pair in the young subfamily.

4.1.2 Data

Our data consist of 15 high-identity ($\geq$80%) genes, organized as 5 closely related triplets, from the VSG antigenic archive of African Trypanosomes (Figure 4.1). The sequences were obtained from the VSGdb database (http://www.vsgdb.net/) and each gene triplet was multiply aligned via CLUSTALw. The triplets selected were: **1)**Tb927.5.5260, Tb09.160.0100, Tb11.38.0005; **2)** Tb09.244.1860, Tb11.57.0027, Tb09.244.0130; **3)** Tb927.3.400, Tb08.27P2.680, Tb09.244.0900; **4)** Tb09.244.1850, Tb09.244.0 140, Tb11.57.0026; **5)** Tb09.v2.0430, Tb09.v4.0178, Tb927.6.5210.

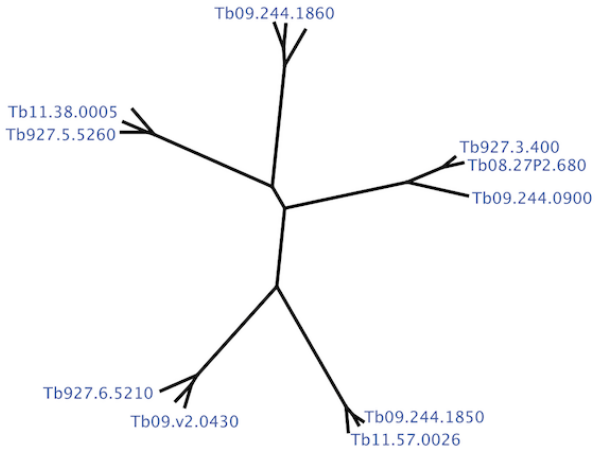


Figure 4.1: Illustration of the phylogenetic structure in the data. Phylogenetic relationships between the 5 VSG triplets studied are constructed on the basis of full-length comparisons between their sequences. In our analysis, only alignments between the N-domains of genes within the same triplet are used.

Subsequently each aligned pair was restricted to only the N-domain of the given gene sequences, given the hypervariability of this region, a key factor in VSG antigenicity. The alignments were then transformed to numerical vectors taking the values of 0 and 1 at each position (0-mismatch, 1-identical nucleotides). Most mismatches obtained in this manner (83-100%) constitute genuine nucleotide substitutions between the two genes compared, rather than insertions or deletions (see Appendix C.1 for details). Each aligned pair is treated as an independent observation.

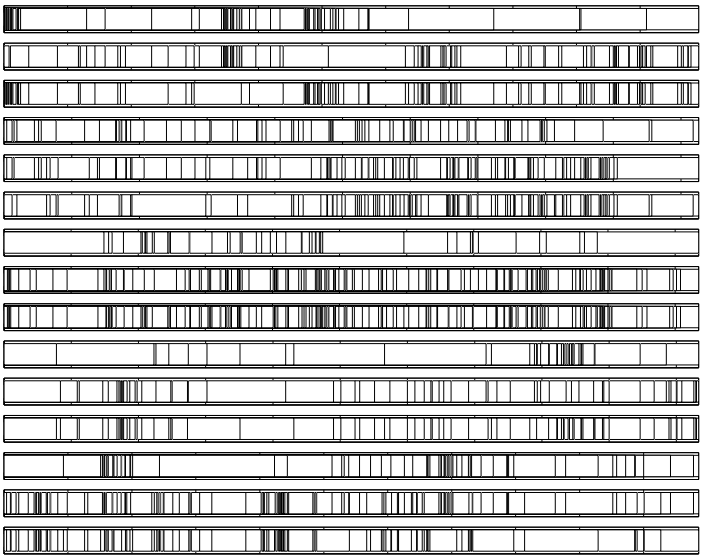


Figure 4.2: The data consist of 15 VSG alignments from 5 triplets of closely related genes. Each of the pairs within a triplet is aligned and presented in the order: (1,2), (1,3), (2,3). The dark bars refer to mismatches between nucleotides in the N-domains of the two sequences. These domains are located respectively in the following nucleotide regions: 1-1092 for triplet 1, 1-1026 for triplet 2, 1-1035 for triplet 3, 43-1075 for triplet 4, and 1-1086 for triplet 5. The next-mismatch distances in each alignment starting from the first mismatch serve as our observations, which we model.

As one scrolls along any given alignment, reading the mismatches from left to right, there are segments of low mismatch density and segments of high mismatch density (Figure 4.2). The first segments, most likely correspond to regions affected only by mutation, the second correspond to areas where a gene conversion with an outside partner has occurred. Because the data do not show the borders between clusters of mismatches, it is very difficult to determine where the conversions have occurred. This calls for a probabilistic model that can enable us to distinguish between these two spatial scales: the short inter-mismatch distances within clusters, and the larger inter-mismatch distances between clusters. This characterization is necessary to assess the relative contributions of these two evolutionary forces, namely point mutation and gene

conversion, on the local diversification within gene families and ultimately, on the evolutionary dynamics of the trypanosome antigenic archive as a whole.

4.2 Model formulation

The simplest characterization of the pairwise identity/diversity between two genes is obtained through their pairwise alignment. Each alignment is a vector $X^{(n)}$, $n = 1,..N$, of length L, where each element $X_i^{(n)}$ takes the values of 0 or 1, to indicate respectively a mismatch or identity at nucleotide position i between the two genes. We assume the mismatch occurrences form a stochastic process on the alignment from left to right, similar to a Markov process, where the next mismatch location depends only on the current mismatch and not on the entire history of previous mismatches.

As a conversion is initiated, a mismatch is positioned at the start of a conversion with probability λ_{begin} per nucleotide. Then there are only two events that can happen: either an internal mismatch is introduced after a certain distance with probability μ per nucleotide, or the conversion terminates with an end-mismatch with probability λ_{end} per nucleotide. This implies that a conversion segment must have at least two mismatches. After a conversion terminates, there are two possible events: either a mutation is found after some distance at probability m per nucleotide, or a new conversion is initiated with probability λ_{begin}.

The simulation of this process can be carried out through a discrete version of the Gillespie Algorithm (Gillespie, 1977) (see Section 4.2.1 for details). Because of the memoryless property assumed, the distances to the next-event, i.e. to the next mismatch, are geometrically distributed with parameter corresponding to the total probability of events that can happen at that current point, analogous to the exponentially-distributed inter-event times in general Markov processes (Karlin & Taylor, 1975). A consequence of the model is that the distribution of conversion tract lengths (defined conservatively from the first to the last mismatch within a cluster) is geometrically distributed with parameter λ_{end}. The geometric tract length distribution appears to describe well the mechanistic basis of gene conversion, and has been applied in previous literature (Betran *et al.*, 1997; Hilliker *et al.*, 1994). Some properties of the model are summarized in Section 4.2.2.

4.2.1 Process simulation

Here, we describe the algorithm for the stochastic occurrence of mismatches on an alignment of length L. We have the following input parameters: $\lambda_{begin}, \lambda_{end}, \mu, m$. Begin with empty vectors $\mathbf{S}$ and $\mathbf{y}$. For simplicity, use: $\lambda_2 = \lambda_{end} + \mu$ and $\lambda_1 = \lambda_{begin} + m$.

1. Denote by type '2' to the first next-mismatch segment, i.e. between-conversion: $s_1 = 2$. Then $\mathbf{S}$ gets updated $\mathbf{S} = [\mathbf{S}, s_1]$.

2. Generate the first next-mismatch distance through the geometric distribution with mean $d_1 \sim Geo(\lambda_1)$, and update $\mathbf{y}$: $\mathbf{y} = [y, d_1]$.

3. While $\sum_i y_i < L$, repeat: $\{$
   $u \sim U(0,1)$,
   If $s(end) = 1$:
   if $u < \lambda_{begin}/\lambda_1$, then $s_{new} = 2, d_{new} \sim Geo(\lambda_2)$,
   else
   $s_{new} = 1, d_{new} \sim Geo(\lambda_1)$;
   else if $s(end) = 2$:
   if $u < \lambda_{end}/\lambda_2$, then $s_{new} = 1, d_{new} \sim Geo(\lambda_1)$,
   else
   $s_{new} = 2, d_{new} \sim Geo(\lambda_2)$.
   Finally, update $\mathbf{S} = [\mathbf{S}, s_{new}], \mathbf{y} = [\mathbf{y}, d_{new}]$. $\}$

4.2.2 Model properties

The conversion length distribution

The memoryless property, characterizing the Markov process, imposes a geometric distribution on conversion lengths, with parameter λ_{end} (Figure 4.3), consistent with previous literature (Betran *et al.*, 1997; Hilliker *et al.*, 1994). Once a conversion is initiated, the probability per nucleotide that it terminates is λ_{end}. This implies that the mean conversion length is $1/\lambda_{end}$.

The number of conversion events

By similar arguments, a geometric distribution with parameter λ_{begin} is also obtained

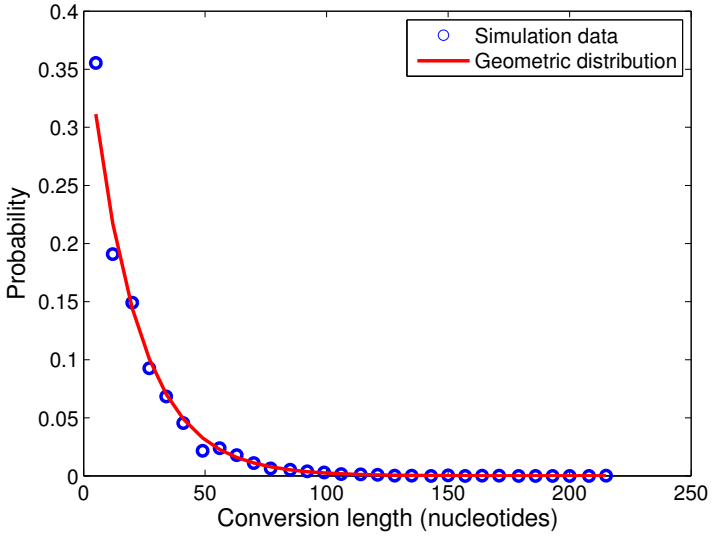


Figure 4.3: Conversion lengths are geometrically distributed with mean $1/\lambda_{end}$. Mean distribution after 500 runs with $(\lambda_{begin},\lambda_{end},\mu,m)$=(0.01,0.05,0.25,0.03).

for the distances between consecutive conversions. This implies that on average this random process on an alignment of fixed length L gives rise to the following number of conversions: $\bar{N_c} = L/(1/\lambda_{begin}+1/\lambda_{end})$, which simplifies into $\bar{N_c} = L\lambda_{begin}\lambda_{end}/(\lambda_{begin}+\lambda_{end})$.

The number of point mutation events

Point mutations occur with probability m per nucleotide, but only in the regions between conversions, namely in $L-\bar{N_c}/\lambda_{end}$ which is the space remaining on average after N_c conversions. Substituting $\bar{N_c}$, we obtain that the number of point mutations is given by: $\bar{N_m} = Lm\lambda_{end}/(\lambda_{begin}+\lambda_{end})$. Thus the ratio between the mean number of conversion and point mutation events is $\bar{N_c}/\bar{N_m} = \lambda_{begin}/m$. Furthermore, simulations confirm that the Poisson probability distribution is a very good approximation for the distribution of the number of conversions and point mutations (Figure 4.4).

The total intensity of the process

In order to obtain an estimate for the total intensity of mismatches, λ, denoting the mean number of mismatches per unit length, we must add up all the mismatches oc-

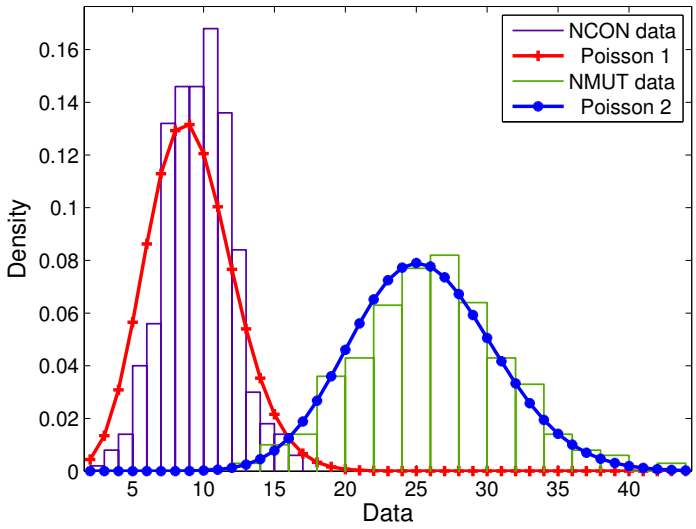


Figure 4.4: The number of conversion and point mutation events in the model are Poisson distributed with rates $\lambda_{begin}L\lambda_{end}/(\lambda_{begin}+\lambda_{end})$ and $mL\lambda_{end}/(\lambda_{begin}+\lambda_{end})$. Empirical distributions (bars) are plotted after 500 process simulations with parameters $(\lambda_{begin},\lambda_{end},\mu,m) = (0.01,0.05,0.25,0.03)$. Superimposed are the corresponding Poisson densities.

curring in conversions to all mismatches occurring in spaces between conversions. Because each conversion extremity also introduces a mismatch, we must add this into the total expression. Then the total intensity is given by:

$$\lambda = \frac{\mu\lambda_{begin} + m\lambda_{end} + 2\lambda_{begin}\lambda_{end}}{\lambda_{begin}+\lambda_{end}}.$$

The distribution of next-mismatch-distances (nnd)

The probabilities that a particular segment between two consecutive mismatches is of *within*- or *between*-conversion type are: $p_w = \frac{\bar{N}_c(\mu/\lambda_{end}+1)}{\lambda}$, $p_b = 1-p_w$, where $\mu/\lambda_{end}+1$ refers to the number of next-mismatch-distances within a conversion. From this, we can write the total cumulative distribution function of next-mismatch distances (nnd) as: $P(nnd < x) = p_w[1-(1-(\lambda_{end}+\mu))^x] + p_b[1-(1-(\lambda_{begin}+m))^x,]$ which simplifies into $P(nnd < x) = 1 - p_w[1-(\lambda_{end}+\mu)]^x - p_b[1-(\lambda_{begin}+m)]^x$. This approximation accurately describes the simulation data, as seen in Figure 4.5).

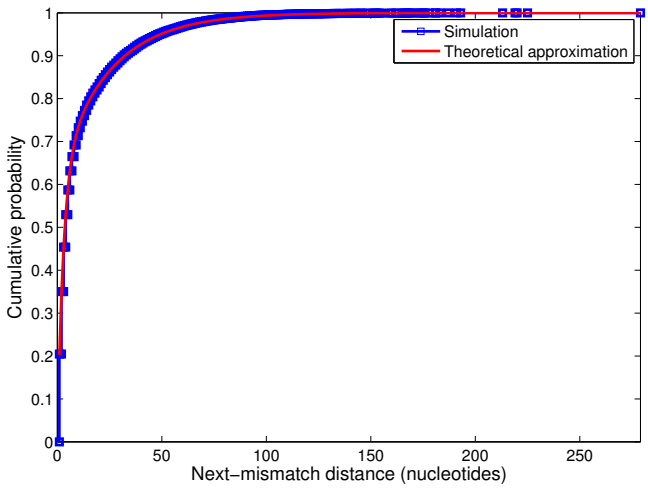


Figure 4.5: The theoretical cumulative distribution of next-mismatch distances matches closely the empirical cumulative distribution obtained from 500 independent runs with parameters: $(\lambda_{begin}, \lambda_{end}, \mu, m) = (0.01, 0.05, 0.25, 0.03)$.

4.3 Modelling VSG alignment data: 4 alternatives

Instead of focusing on mismatches themselves and their locations, it is more convenient to transform the data into inter-mismatch distances. Suppose alignment $X^{(n)}$ has M mismatches. We can thus consider that each observation $y_i (i = 1, ...M)$ of next-mismatch-distances along an alignment is associated with an unobserved hidden state $s_i = k, k \in \{1,2\}$: 1, corresponding to inter-mismatch distances *between clusters*, and 2, corresponding to inter-mismatch distances *within-clusters*.

Conditioned on the type of the hidden state, each y_i observation is assumed to be independently drawn from a geometric distribution: $\Phi_k(d) = P(y_i = d | s_i = k) = (1-\lambda_k)^{d-1}\lambda_k$, with parameters $\lambda_1 = \lambda_{begin} + m$, for between-cluster distances, and $\lambda_2 = \mu + \lambda_{end}$, for within-cluster distances, where $d = 1, 2, 3, ...L-1$. This formulation implies that the modes of both small-scale and large-scale next mismatch distances are the same and very small (equal to 1).

The model described above is an instance of a large class of models known as Hidden Markov Models (Durbin *et al.*, 1998), widely used in biological sequence analysis. The numerical observations are generated by hidden states, forming an ordered

sequence known as the path. In the path, the probability of the next hidden state depends only on the current hidden state, that is why the path follows a simple Markov Chain. The transition matrix between states in our model is given by:

$$T = \begin{pmatrix} \frac{m}{\lambda_{begin}+m} & \frac{\lambda_{begin}}{\lambda_{begin}+m} \\ \frac{\lambda_{end}}{\lambda_{end}+\mu} & \frac{\mu}{\lambda_{end}+\mu} \end{pmatrix}, \tag{4.1}$$

where entry $T_{1,2} = P(s_i = 2|s_{i-1} = 1)$ and so on, expressing the probabilities for the mismatches to persist within clusters or to jump between clusters.

If the 4 basic genetic parameters, $\lambda_{begin}, \lambda_{end}, \mu, m$ are known, the transition probabilities $T_{i,j}$ and the two geometric distributions for the next-mismatch segment lengths ("emission" probabilities) are uniquely determined. Conditioned on the sequence of hidden states $S = \{s_i, i = 1, ..., M\}$, the likelihood of the data $\mathbf{y} = \{y_i, i = 1, ...M\}$ on each alignment is:

$$P(\mathbf{y}|S) = \prod_{i=1}^{M} (1 - \lambda_{s_i})^{y_i - 1} \lambda_{s_i}. \tag{4.2}$$

The joint probability of the observations and a particular hidden path is given by:

$$P(\mathbf{y}, S) = T_{0k} \prod_{i=1}^{M} (1 - \lambda_{s_i})^{y_i - 1} \lambda_{s_i} T_{s_i s_{i+1}}, \tag{4.3}$$

where T_{0k} is the transition probability from an artificially introduced begin state to state k, and can be thought of as the probability of starting in state k. Because many different hidden paths can give rise to the same sequence of observations y, to obtain the full likelihood of $\mathbf{y}$, we must sum over all possible sequences of hidden states $P(\mathbf{y}) = \sum_S P(\mathbf{y}, S)$.

Given the observations of next-mismatch distances in N alignments, one is interested in estimating the genetic parameters (λ_{begin},λ_{end}, μ, m) that are most likely to have generated the entire dataset. Naturally, by independence between gene pairs, the total likelihood of the data will be a product over the individual alignment likelihoods.

Since the relationship between the genetic parameters governing the mismatch process, and transition and "emission" probabilities, governing the hidden Markov model is one-to-one, one can estimate the latter and subsequently obtain the original parame-

ters. For example, with the simpler notation:

$$T = \begin{pmatrix} 1-p_2 & p_2 \\ p_1 & 1-p_1 \end{pmatrix}, \tag{4.4}$$

$$\Phi_k(d) = (1-\lambda_k)^{d-1}\lambda_k, \quad k \in \{1,2\}, \tag{4.5}$$

we can recover explicitly the 4 genetic parameters as follows:

$$\begin{aligned} \lambda_{begin} &= p_2\lambda_1, \qquad m = (1-p_2)\lambda_1, \\ \lambda_{end} &= p_1\lambda_2, \qquad \mu = (1-p_1)\lambda_2, \end{aligned} \tag{4.6}$$

after the auxiliary parameters $p_1, p_2, \lambda_1, \lambda_2$ have been estimated.

Different models can be constructed to describe the same dataset, based on different assumptions on independence between aligned pairs (see Figure 4.6). In the following, we present results for 4 models that we consider most relevant and biologically plausible:

1. **Global fit model** The simplest assumption is the hypothesis that the same parameter values apply to all ($N = 15$) aligned pairs simultaneously. All gene pairs come from the same process, thus sharing the same probability of conversion initiation, conversion termination, mutation and the same mismatch density per conversion. This model results in 4 parameters that should explain the mismatch data of every pairwise alignment.

2. **Triplet fits** Alternatively, the data may be seen as a collection of 5 independent unrelated triplets, where each triplet of genes is governed by its own set of primary parameters. This formulation entails $5 \times 4 = 20$ parameters in total.

3. **Triplet ages** The data may consist of 5 partially-related triplets, which are governed by the same density of mismatches within conversions μ, and same probability of conversion termination λ_{end}, but differ in the mutation probability m, and probability per nucleotide to start a conversion λ_{begin}. If relative to the first triplet, the latter probabilities scale by the same factor in the other triplets, we could introduce a new parameter in the model, namely: the relative 'age' of

each triplet. Triplet 1 gets assigned age $A_1 = 1$. Then, for the other triplets $(2, ...N/3-1)$, the 'ages' relative to the first triplet $A_{triplet}$ can be inferred. It is sufficient to estimate λ_{begin} and m only for the first triplet. Such model has $4+4=8$ parameters. When 'ages' are included, the triplet-specific probability of conversion start per nucleotide λ_{begin}, and the triplet-specific probability of mutation per nucleotide m, after scaling become the products $\lambda_{begin}A_{triplet}$ and $mA_{triplet}$, which by definition must fall in the range $[0,1]$. Intuitively, in aligned pairs from an 'older' triplet we should expect more conversion events and more mutation events on average than in a 'younger' triplet.

4. **Individual ages** Here we consider the case where each gene pair shares the same μ and λ_{end} with the other gene pairs, but differs in λ_{begin} and m. Assuming that the latter parameters scale equally among gene pairs, we can introduce again a scaling parameter, similar to a pair-specific 'age' relative to pair 1. This yields a set of 4 primary parameters governing all alignments, but with the exclusion of the first alignment, for the other alignments, parameters λ_{begin} and m have to be effectively scaled (as in the Triplet ages model), according to their relative age $A_i, i = 2, ...N$. The number of parameters here becomes $4+14=18$. This model also results in the same conversion tract length distribution and the same density of mismatches per conversion across all gene pairs.

4.4 Parameter estimation

We adopt a Bayesian approach which allows us to estimate explicitly the switching rates between large-scale and small-scale next-mismatch-distances and the probability distributions associated with each of these states. In addition, this approach enables us to include any prior knowledge about the process, and to quantify the uncertainty present in the alignment mismatch data. We implement the Metropolis-Hastings Algorithm, one of the simplest Markov Chain Monte Carlo algorithms (Gilks *et al.*, 1996), where instead of point-estimates that maximize the likelihood of the data, we look for probability distributions over model parameters. MCMC techniques are useful to get a representative sample from the posterior when the posterior distributions themselves

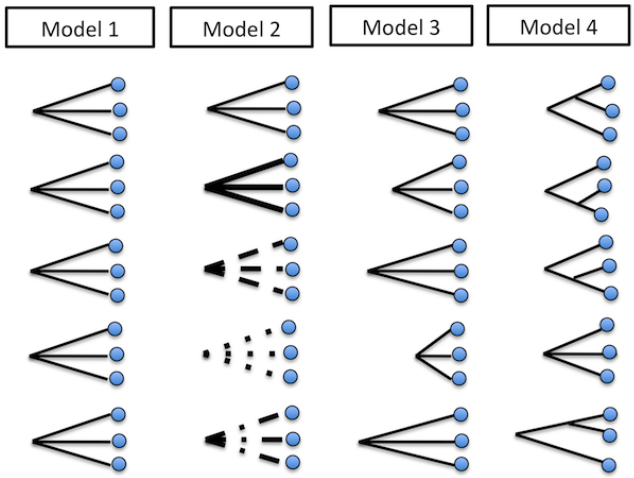


Figure 4.6: Model diagrams. Our 4 models differ in the assumptions they make about the nature of the evolutionary processes (depicted by line type) and the divergence time between the compared sequences (depicted by line length). Model 1 assumes the mutation and conversion process are governed by the same parameters on all gene pairs, and that each pair within a triplet shares the same divergence time with the other pairs. Model 2 assumes the genetic processes occur in each gene triplet at distinct triplet-specific rates, and in addition, it allows for triplet-specific conversion length distribution and conversion mismatch density. Model 3 assumes the processes occur universally at equal rates across triplets, including conversion length distribution and mismatch density, however the divergence time of each triplet may be different, resulting in time-scaled triplet-specific mutation and conversion event probabilities. Model 4, like Model 3, assumes mutation and conversion processes occur universally at equal rates across gene pairs, but it allows for within-triplet variation in divergence time.

are impossible to get analytically. We use uniform priors throughout our model fitting. For calculating the overall likelihood of each sequence of observations, we employ HMM posterior decoding (Durbin *et al.*, 1998), which takes into account all possible hidden paths that might have generated the observations.

The algorithm is implemented in MATLAB, and starts with an initial guess of the parameter values. Then a new guess is generated from a proposal distribution, e.g. a multivariate normal distribution centred at the current value of the parameters. The variance of the proposal distribution is carefully chosen so as to optimize the speed of convergence to the stationary distribution. Then the new likelihood of the data is

calculated for the new values of the parameters. If it is greater than the old likelihood, the new set of parameters is accepted with some probability, otherwise it is rejected. As the above steps are repeated over and over again, the MCMC is expected to converge to its stationary distribution.

For any model, it is customary to run more than one MCMC chain from different starting points, until no autocorrelation remains and convergence to the stationary distributions in parameter sample paths is reached. Several statistical measures of convergence can indicate whether convergence is satisfied. One such measure is the Gelman-Rubin convergence statistic, as modified by Brooks & Gelman (1998), a test that compares variance within and between various Markov chains. After convergence, the Markov chains are let to run for several thousand more iterations, after which the posterior distributions of the parameters and other statistics of interest can be computed. For a full description of estimation procedures we refer to Appendix C.2.

4.5 Model comparison and goodness of fit

In maximum likelihood model selection, models can be compared using the Akaike's Information Criterion (AIC) (Akaike, 1974). In Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation, a similar criterion is used. This criterion is called the Deviance Information Criterion (DIC) and has been proposed by Spiegelhalter *et al.* (2002). Like AIC and BIC (Bayesian Information Criterion) it is an asymptotic approximation as the sample size becomes large. It is only valid when the posterior distribution is approximately multivariate normal. As in other model comparison tools, DIC consists of two terms: one representing goodness of fit and the other penalizing for increasing model complexity. Model fit is represented by the expectation of the posterior distribution of the "Bayesian Deviance", which is calculated from the posterior distributions of the set of parameters θ as follows:

$$\mathrm{Dev}(\theta) = -2\log P(y|\theta). \tag{4.7}$$

For model complexity on the other hand, an index measuring the "effective number of parameters" is calculated:

$$p_D = \bar{D} - \mathrm{Dev}(\bar{\theta}), \tag{4.8}$$

where $\bar{D} = E^{\theta}\mathrm{Dev}(\theta)$ is the expectation of the deviance over the MCMC parameter sample, and $\bar{\theta}$ is the mean of the set of parameters θ. DIC is then defined by

$$DIC = p_D + \bar{D}. \tag{4.9}$$

Generally models with lower DIC are preferred over models with higher DIC, although this is not always used as a strict criterion for model choice. As independent tests, we compare also simulated data from estimated parameters to the original dataset. We check whether mismatch patterns obtained with model parameters can produce pair-correlation functions Illian *et al.* (2008) and next-mismatch distance cumulative distributions similar to those observed in the data (see Appendix C.2.3).

4.6 Results

Convergence of the Markov chains was generally reached during the first few thousand iterations. As it is normal procedure in calculating the posterior, we discarded the first 10000 iterations as 'burnin' and let the algorithm run for another 50000 iterations for posterior estimation. Thus for every parameter, the posterior was obtained from a sample of 3×50000 independent MCMC observations. DIC and log-likelihood values for each model are reported in Table 4.1.

Table 4.1: Summary of the DIC indices and the mean log-likelihood values for the 4 models considered

Model	Nr.parameters	Log-Likelihood	DIC
1. **Same age**	4	-4430.8	8232.8
2. **Triplet fits**	20	-4399.4	8140.4
3. **Triplet ages**	8	-4414.2	8188.4
4. **Individual ages**	18	-4390.5	8136.1

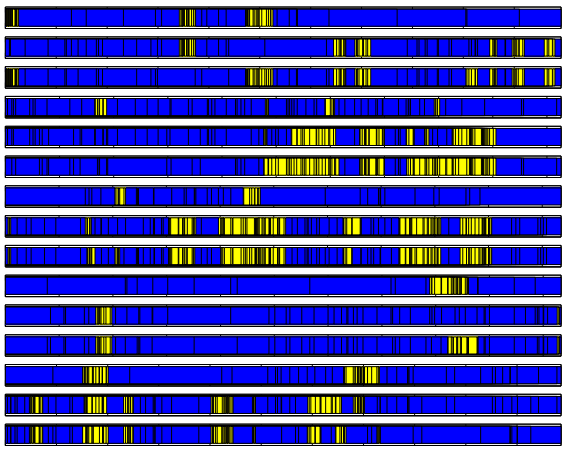
Although the models differ in the assumptions they make, the estimated values of the primary parameters and the DIC values across models do not vary hugely. In particular, all models agree substantially on the estimates of the density of mismatches per conversion μ, and the per-nucleotide probability of conversion termination λ_{end}. Unsurprisingly so, as the two parameters are expected to remain invariant across models.

In Model 1, the estimated mean probability of conversion start per nucleotide was estimated to be 0.0099, whereas the mean probability of mutation per nucleotide was estimated to be 0.0410, i.e. about 4 times higher. This suggests that mutation events are estimated to be more frequent than conversion events in these gene pairs. The average conversion length predicted by this model is $1/0.0387 \approx 25$ nucleotides, which fits nicely with the most likely estimated conversion tracts (Figure 4.7). On the other hand, the density of mismatches within conversion tracts was estimated to be as high as 0.2552 (for more details see Table 4.2).

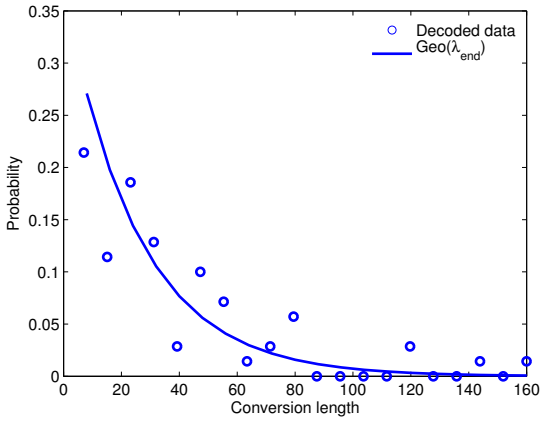
Table 4.2: Parameter estimates obtained for Model 1 (Same age)

Parameter	5% Conf. Bound	Mean	95% Conf. Bound
λ_{begin}	0.0077	0.0099	0.0120
λ_{end}	0.0281	0.0387	0.0503
μ	0.2350	0.2552	0.2764
m	0.0366	0.0410	0.0458

In Model 2, we considered the case of each triplet being independent, and thus its mismatch pattern being governed by distinct values of parameters (Table 4.3). We found that the estimated λ_{begin} was in the range 0.0038-0.0175 across triplets, a result not very far-off the initial estimate obtained with Model 1. For λ_{end}, the mean range was 0.0126-0.0836, implying conversion lengths between 12 and 80 nucleotides. The mean density of mismatches per conversion varied across triplets, ranging from 0.2043 to 0.3469, however its global average, about 0.26 was consistent with the value of μ predicted by the first model. The mutation rate also showed some variation 0.0325-0.0623, yet the values predicted for each triplet stayed within the same order of magnitude. The most likely conversion tracts predicted by this model are shown in Figure 4.8.

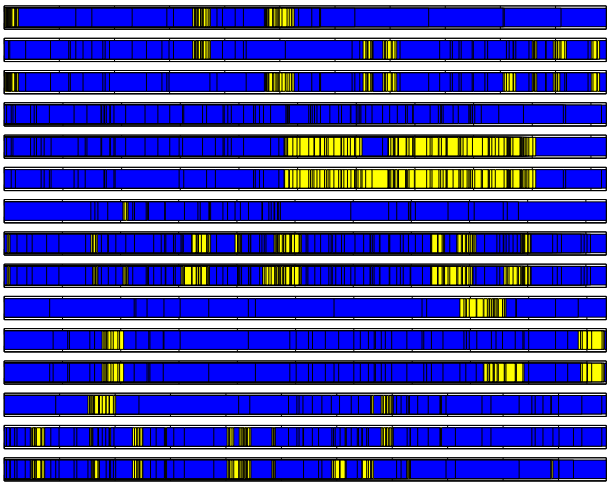


(a) Decoded conversions

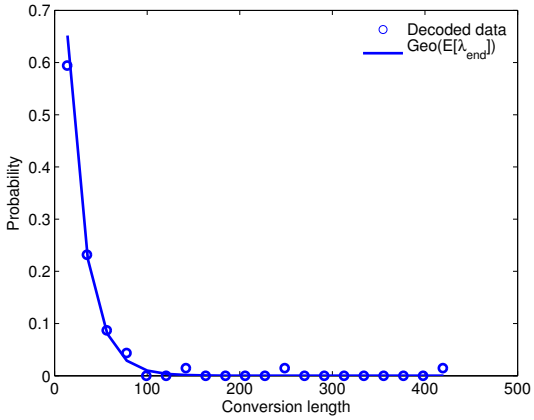


(b) Conversion length distribution

Figure 4.7: Model 1 (Global fit) assumed that all 15 pairs were governed by the same parameters.



(a) Decoded conversions



(b) Conversion length distribution

Figure 4.8: Model 2 (Triplet fits): each triplet governed by its own set of parameter values.

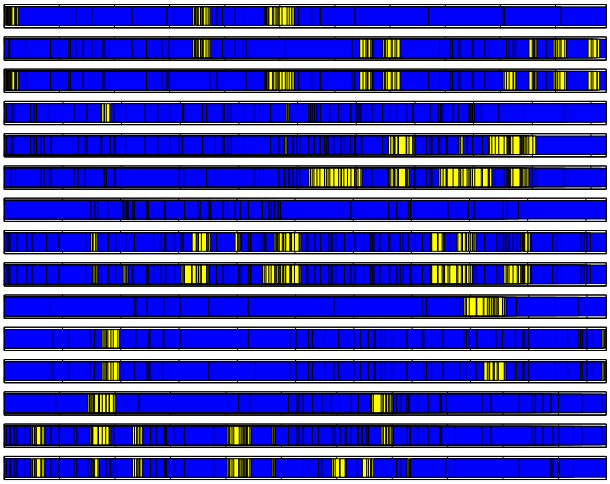
Table 4.3: Parameter estimates obtained for Model 2 (Triplet fits). Only the means are shown.

Triplet i	$\lambda_{begin}(i)$	$\lambda_{end}(i)$	$\mu(i)$	m(i)
1	0.0092	0.0534	0.3469	0.0325
2	0.0056	0.0126	0.2073	0.0578
3	0.0175	0.0632	0.2476	0.0623
4	0.0038	0.0248	0.2043	0.0379
5	0.0138	0.0836	0.3110	0.0382

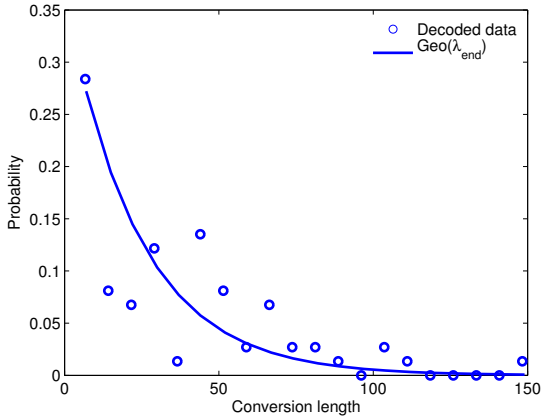
In Model 3, which assumed each triplet had a different 'age' relative to the first triplet of genes, we obtained mean estimates for the relative 'ages' ranging from 1.09 to 2.05 (see Table 4.4), a result that shows a moderate variation with respect to the reference triplet. As a consequence, the estimates obtained for the primary genetic parameters do not vary hugely from the estimates obtained in Model 1 for example. The ratio between m and λ_{begin} increases slightly, ≈ 4.7, reinforcing the dominance of mutation events over conversions. The result of the decoding procedure is given in Figure 4.9.

Table 4.4: Parameter estimates obtained for Model 3 (Triplet ages). The triplet-specific λ_{begin} and m are obtained via multiplication of the baseline values given here with the corresponding relative age.

Parameter	5% Conf. Bound	Mean	95% Conf. Bound
λ_{begin}	0.0052	0.0069	0.0088
λ_{end}	0.0354	0.0473	0.0594
μ	0.2536	0.2717	0.2910
m	0.0273	0.0331	0.0392
$A_{\text{triplet 2}}$	1.3534	1.7220	2.1556
$A_{\text{triplet 3}}$	1.6253	2.0594	2.5672
$A_{\text{triplet 4}}$	0.8523	1.0918	1.3731
$A_{\text{triplet 5}}$	0.9914	1.2733	1.6046



(a) Decoded conversions



(b) Conversion length distribution

Figure 4.9: Model 3 (Triplet ages) assumed that the process of gene conversion and point mutation across triplets happen at the same rates and characteristics, but there is difference in the "age' of each triplet.

In Model 4, which assumed each gene pair is characterized by a different evolutionary 'age' relative to the first gene pair, more variation in estimated ages was found than in Model 3, with an approximate range from 0.96 (gene pair 10) to 5.28 (gene pair 9), which strongly correlated with the differences in pairwise identity between alignments (see Figure 4.10). However, this variation is still within a factor of 5, which is unsurprising given the fact that all the genes in the dataset display similar pairwise nucleotide identity ($\approx 90.7\%$), i.e. presumably they should have been diverging from a common ancestor for a relatively short time. Because the effective probabilities of conversion initiation and mutation on each aligned gene pair are obtained by the baseline values multiplied by the corresponding relative ages, the λ_{begin} and m values inferred in this model for the reference pair are lower than the values obtained in Model 1. The ratio m/λ_{begin}, however is invariant across gene pairs and independent of their relative age. We observe that mutations occur at least 5 times more frequently than conversion events. We notice that in this model, the density of mismatches per conversion is estimated to be 0.2877, higher than in Models 1 and 3, whereas the mean conversion length is lower than in other models, approximately 18 nucleotides. These results are summarized in Table 4.5.

Table 4.5: Parameter estimates obtained for Model 4 (Individual ages). The alignment-specific λ_{begin} and m are obtained via multiplication of the baseline values given here with the corresponding relative age.

Parameter	5% Conf. Bound	Mean	95% Conf. Bound
λ_{begin} (pair 1)	0.0021	0.0035	0.0052
λ_{end}	0.0400	0.0551	0.0718
μ	0.2611	0.2877	0.3127
m (pair 1)	0.0124	0.0191	0.0270
A_2	1.4725	2.3980	3.6719
A_3	1.5531	2.5507	3.9701
A_4	1.8741	2.9923	4.5708
A_5	2.6478	4.2168	6.3249
A_6	1.7537	2.9912	4.6698
A_7	1.3536	2.3153	3.5950
A_8	3.0133	4.7650	7.1719
A_9	3.3190	5.2838	8.0093
A_{10}	0.5063	0.9645	1.5966
A_{11}	1.6073	2.6257	4.0232
A_{12}	1.6795	2.7357	4.1687
A_{13}	1.0776	1.8165	2.8530
A_{14}	1.9615	3.2001	4.8341
A_{15}	1.5089	2.5283	3.9155

Common to all models is the small variance in the posterior distribution of one parameter, namely the density of mismatches per conversion, μ. Instead, for the other parameters, there is a larger variance in the posterior distributions. The obtained posteriors are generally unimodal and symmetric around the mean, resembling the normal distribution (see Figures 4.11, 4.12).

DIC values for each model indicated that rank order performance of these four formulations supports Model 4 as the best model, and then Model 2, Model 3 and

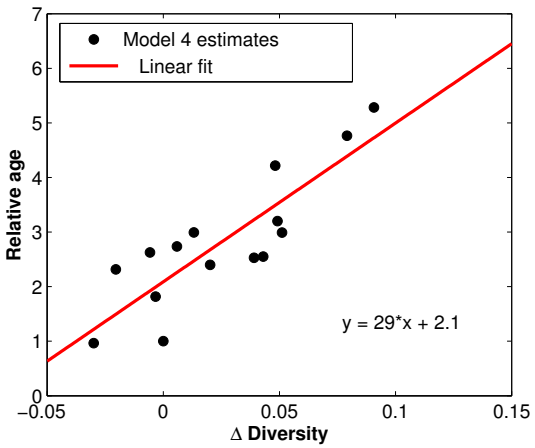


Figure 4.10: There is a high correlation (estimated to be 0.8531 by a standard correlation test) between the relative ages inferred by Model 4 and the differences in diversity between pairs relative to pair 1.

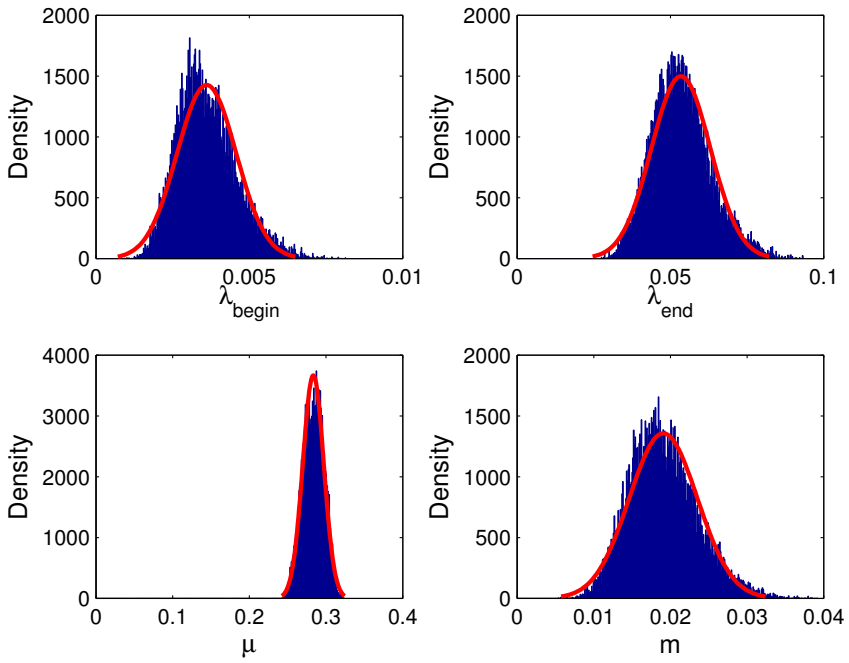


Figure 4.11: Posterior distributions obtained for the 4 baseline parameters of Model 4 (Individual Ages), which was ranked as the best model from our model selection procedure.

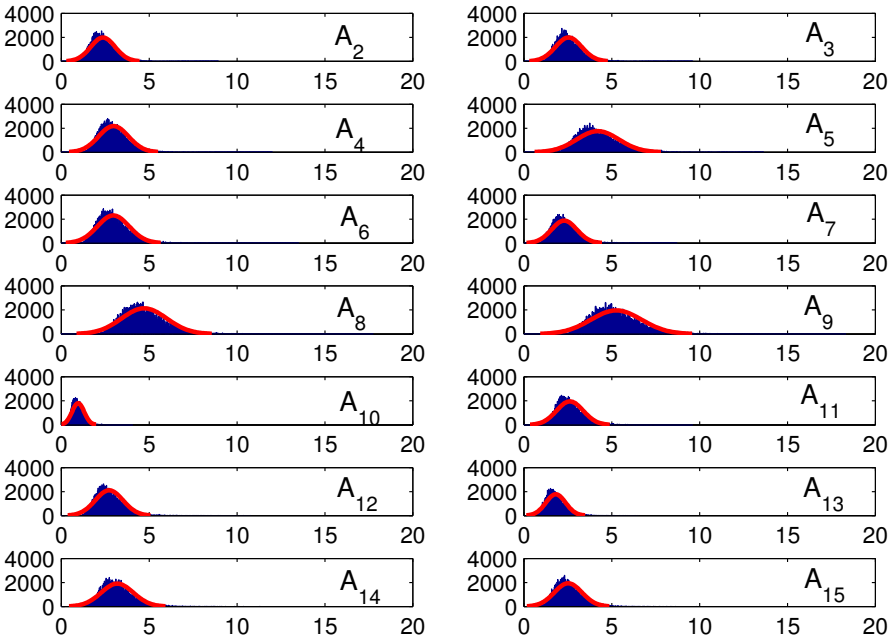


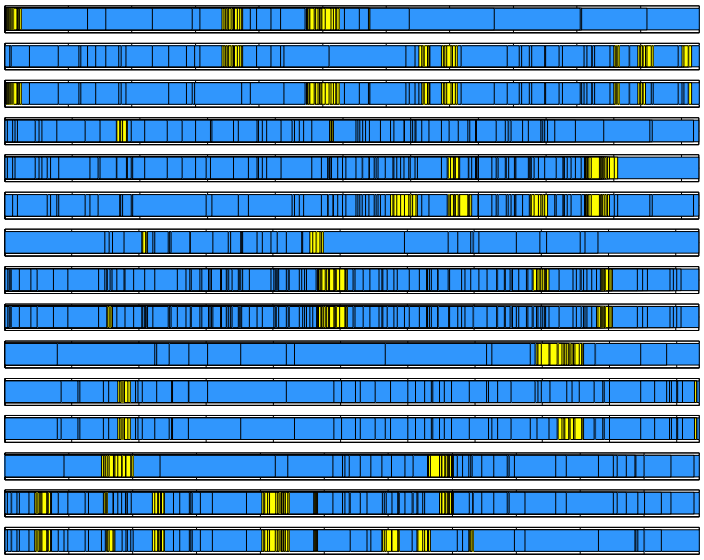
Figure 4.12: Posterior distributions obtained for the ages A_i of each gene pair ($i = 2,...15$) relative to the first pair, with Model 4 (Individual Ages).

Model 1, although the differences were not very big. It is however reassuring to see the parameters are not overly sensitive to the model structure. We applied the Viterbi algorithm (Forney, 1973) to the observed mismatch patterns on all aligned pairs, within the framework of Model 4, in order to "decode" the most likely hidden path, i.e. the most likely sequence of between- and within-conversion next-mismatch segments that could have generated the distances observed. The most likely positions of conversion tracts and point mutations along each alignment are illustrated in Figure 4.13. In Figure 4.14, in contrast to the most likely sequences of hidden paths for Model 4, we show the probabilities of finding a conversion segment along each alignment.

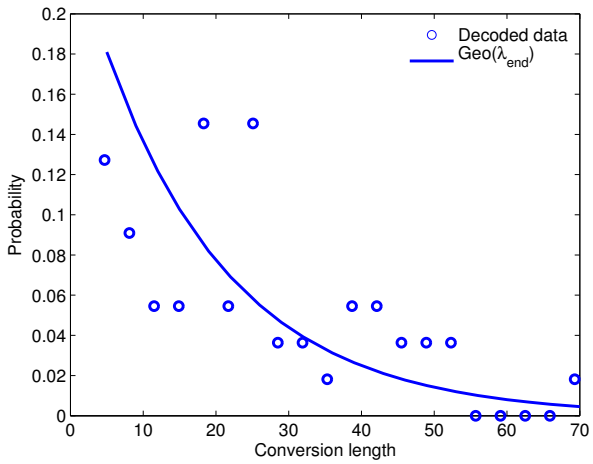
Further, as independent goodness-of-fit tests, we compared pair correlation functions in the original dataset with pair correlation functions (Illian *et al.*, 2008) of simulated data for the best model. Similarly, we also compared the cumulative distribution of next-mismatch-distances in the real alignments and in simulated alignments with model parameters, to verify the quality of fit of Model 4. As shown in Figure 4.15, the simulated statistics very closely matched the statistics from the original dataset. This suggests that the individual ages model provides a very good description of the diversity pattern found in the aligned VSG pairs.

4.7 Discussion

We have presented a general modelling framework that can describe pairwise identity patterns within gene families and an inference framework that can disentangle two genetic processes: gene conversion with partners outside the family and point mutation. Although applied to the *VSG* genes of African trypanosomes, our approach has several advantages that may apply also to other, similar contexts: 1) it uses abstract, global-level information about mismatch occurrence between two aligned gene sequences, without requiring information about the underlying DNA; 2) it accounts for the spatial ordering of the identity pattern; 3) it allows direct estimation of switching rates between two different scales: short and long inter-mismatch distances; 4) it provides a means of quantifying the mutational processes that mechanistically give rise to the observed identity pattern; 5) its results can be applied to the case when another process acts instead of gene conversion but with the same effect of introducing clustered mismatches (e.g. localized point hypermutation, such as in immunoglobulin gene somatic



(a) Most likely conversion tracts



(b) Conversion tract length distribution

Figure 4.13: Decoding results for Model 4 with parameter means as in Table 4.5. a) The 15 alignments from 5 triplets of closely related VSG genes are presented as horizontal bars in the order: (1,2), (1,3), (2,3) for each triplet. The vertical bars refer to mismatches on the aligned N-domains. The most likely conversion tracts estimated by the algorithm are highlighted in yellow, whereas between-conversion segments are given in blue. b) The empirical conversion lengths obtained after "decoding" closely match the theoretical geometric distribution predicted by Model 4 with parameter $E[\lambda_{end}] = 0.0551$.

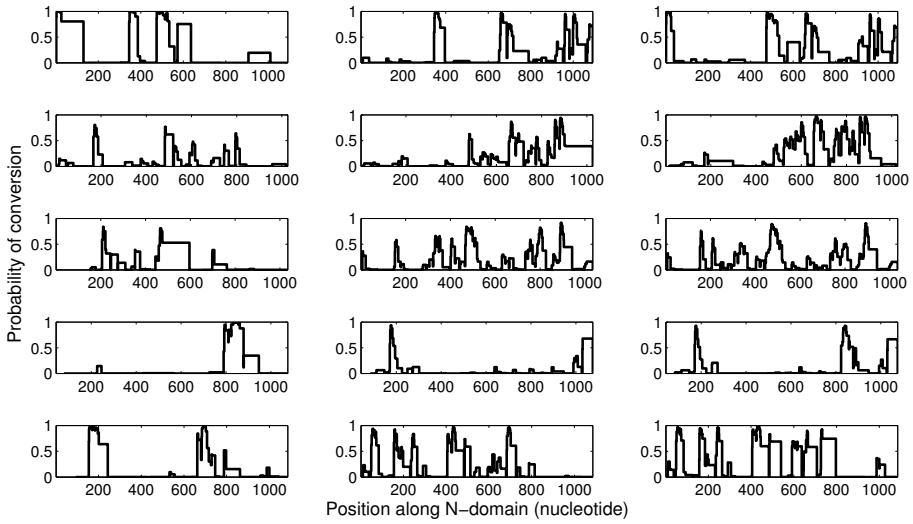


Figure 4.14: Posterior probabilities of finding a conversion segment ('between' type) along each alignment for Model 4. In contrast to the Viterbi algorithm, which gives only the most likely sequence, the posterior probabilities contain more information.

hypermutation); 6) it lends itself easily to further extension through the incorporation of additional factors that may influence the nature of evolutionary processes.

Through our analysis, we find that the patterns of pairwise diversity between the *VSG* genes we have modelled can be explained by two processes: one that results in local clustering of mismatches and one that generates sparse diversity across the entire alignment length. All models considered reveal that the probability of finding a mismatch cluster between two genes is lower than the probability of single mismatches, which suggests a higher propensity for these genes to point-mutate (i.e. diversify independently), than to accept segmental conversion from donor genes in the archive.

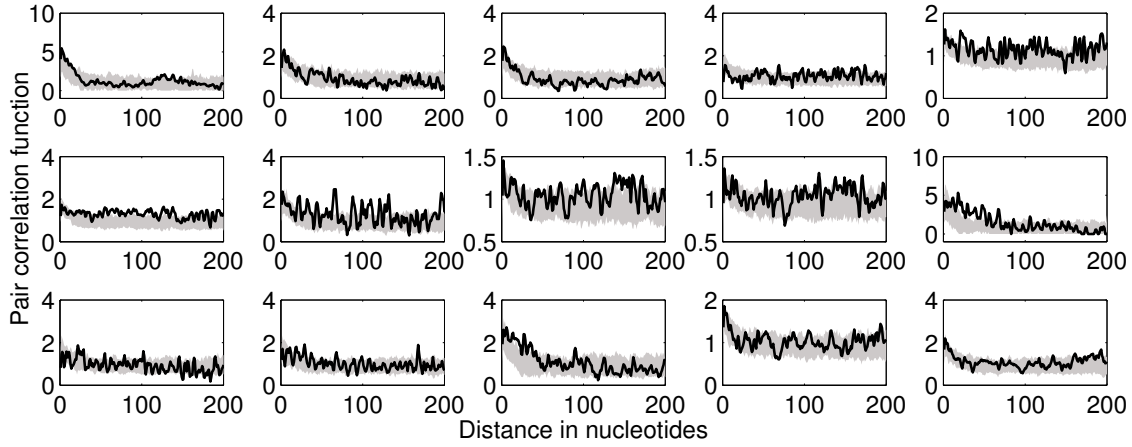
The average density of mismatches per conversion predicted by our models is around 0.25. Combining the estimate of $\mu \approx 0.29$ obtained from Model 4, with the average conversion length obtained by $1/\lambda_{end} \approx 18$, we obtain the average number of mismatches contributed by one gene conversion into the alignment is roughly 5.2. This implies that the relative contribution of gene conversion on pairwise diversity between two genes, on a per-nucleotide basis, is about 0.96 times that of point-mutation. Thus,

although on an event basis mutation is apparently much stronger than gene conversion, on a per-nucleotide basis, the two processes have almost equal impact on genetic diversification.

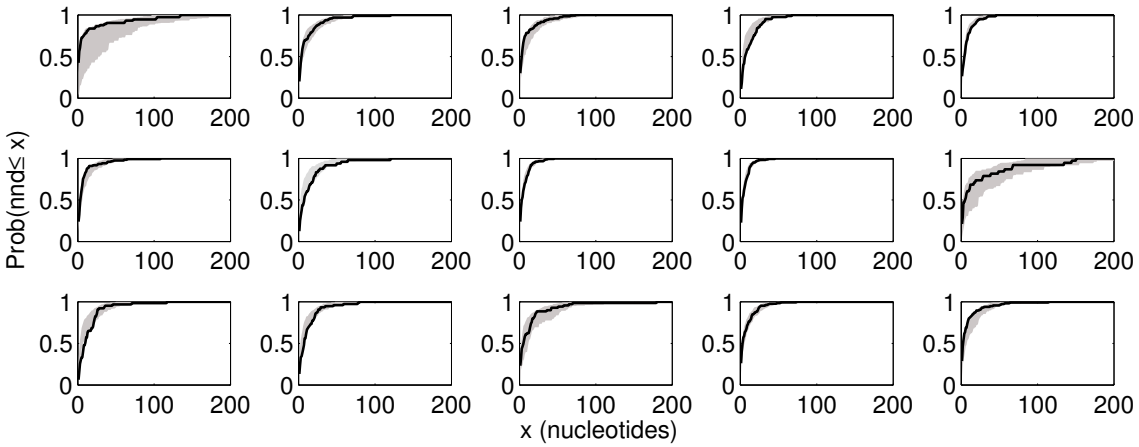
Model 4 ranked as the best model by our model selection procedure. Its inclusion of estimates for the ages of gene pairs relative to the first pair indicates that these parameters hold key information which is crucial for the data fit. Reassuringly, the estimated ages have a small mean of 2.83 and a small standard deviation, confirming a relatively minor variation, expected among gene pairs sharing similar pairwise identity. Notably, Model 4 supports short conversion tracts, between 14 and 25 nucleotides in length, geometrically distributed with mean equal to 18 nucleotides. This characteristic small scale suggests the existence of functional or structural constraints in the conversion process, that may limit the inflow of diversity from the rest of the archive into high-identity VSG subfamilies. Such short conversions resemble human MHC gene conversion events, which display a mode of 14 nucleotides (range 2 to 35) Parham *et al.* (1995), inevitably arising from a mechanistic basis in the recombination pathway involved. Conceivably, the observed length range maximizes effective alteration in *VSG* epitopes.

By using sequence 'decoding' as a first step for finding the most likely locations of conversions and point mutations along each alignment, we can then examine the sequence underlying the mismatches. Comparing their locations obtained from the abstract point pattern, with the original encoded amino acids and nucleotide sequence can reveal what type of sequence is being exchanged through gene conversion. Furthermore, because alignments occur in triplets, the regions where the maximum-likelihood 'decoding' locates conversion segments, can be cross-compared and thus indicate which gene in the triplet received that particular conversion from outside. If two gene pairs within a triplet display a conversion segment in the same location, then the gene they have in common must have been the original recipient of that conversion from an external donor.

Clearly, an important aspect of VSG archive evolution that this framework may elucidate is the rate of divergence between gene families. The assumption that gene conversion with outside partners brings in diversity through a cluster of mismatches in the alignments considered can be translated into a hypothesis for the relative rates of divergence between- and within-families. Within-family (within-pair) identity, on a



(a) Pair correlation functions



(b) Cumulative next-mismatch distance distribution

Figure 4.15: Goodness-of-fit tests for Model 4. The gray shaded area represents 95% credibility intervals for the modeled mismatch patterns (100 replicates, with mean estimates for each parameter as in Table 4.5). The lines represent the observed mismatch data from the N-domains of the 15 aligned VSG pairs.

per-nucleotide basis, is higher than between-family identity or mean pairwise identity across the archive. That is why the density of mismatches is higher in the regions where DNA has been imported from outside through gene conversion, rather than changed from within, through point mutation. If one assumes that point mutation happens at the same rate across all *VSG* genes, as appears to be the case Hutchison *et al.* (2007); Jackson *et al.* (2010); Marcello & Barry (2007a), the difference in mismatch density within and between conversions can only be attributed to differences in evolutionary time, an avenue calling for further investigation.

Furthermore, our inferred mean density of mismatches per conversion, approximately equal to 0.25 could suggest that the mean diversity in the pool of genes where conversions are coming from is around 25%. Notice that if there is no preference in conversion partners, this observation leads to a prediction on the global mean pairwise identity in the VSG N-domains, of around 75%. Alternatively, if conversion with VSG genes outside the given subfamily is preferential, i.e. it occurs only with specific gene subsets of the archive, the 75% identity would only be indicative of N-domains in the subset of the VSG archive with which the given subfamily interacts.

The mean nucleotide identity in the archive is very low (0.26 i.e. $\neq 0.75$), thus conversion appears to be biased towards more similar donor genes. As argued earlier, it is difficult to account for this effect through direct selection of the sequence of individual *VSG*s. A reason for this preferential use of more homologous segments might lie in the need to introduce a segment from the corresponding region of a donor *VSG*, so that the characteristic cysteine pattern of the protein is conserved, rather than being disrupted by random conversion from any region of a donor gene.

4.7.1 Future work

To be able to transform our estimated point mutation and conversion probabilities into actual probabilities per unit of time (or per generation), one would need information on the precise time since duplication of the reference gene pair at least. Once the real evolutionary age of the reference gene pair (with $A_i = 1$) is established, its λ_{begin} and m parameters can be immediately scaled, and subsequently the other pair-specific observed probabilities can be updated. So far, the time information has been missing, but will become available through longitudinal sequencing of field isolates, at which

time it can be readily incorporated into the hidden Markov model, whereby our current estimates would become dimensional parameters of evolutionary processes (rates per unit time, per nucleotide), that can then be compared to similar estimates coming from other methods.

In our models, we assumed that the density of mismatches in each conversion is the same and fixed. Such an assumption might not always hold, as gene conversion donors in the rest of the archive may come from particular subfamilies, each having had its own rate of divergence, thereby contributing a distinct mismatch clustering density. A more general framework in that case, to accommodate this phenomenon, could be to model the mean density of mismatches per conversion, μ, through a probability distribution.

Another simplifying assumption in our study is the spatial homogeneity in the occurrence of point mutations and conversions. It is possible that formulations and inference frameworks accounting for spatial bias might bring additional insight into the nature of gene diversification. From visual inspection of the mismatch patterns, one notices segments of low mismatch density, besides segments of high mismatch density on the aligned pairs. Although neglected in this analysis, such high-identity segments are potentially indicative of within-pair conversion and could be addressed in future studies.

Finally, by considering conversion segment length distributions different from the memory-less geometric distribution assumed here, one might represent other types of gene conversion. A negative binomial or gamma distribution could for example account for a mode of conversion lengths different from 1 nucleotide, thus allowing longer conversions to be more frequent. More flexible distributions would require more sophisticated modelling and possibly a larger dataset, but such alternatives could be explored to better disentangle the various spatial scales that characterize genetic diversity. Nonetheless, the model presented here provides a framework that can easily be built upon as more data become available, offering a valuable tool for a more parametric understanding of genetic diversification processes.

In the next chapter, we zoom out on genetic diversification at the global level of the VSG archive, where the local interactions between genes and subfamilies, such as the ones considered in this chapter, average out. Such averaging gives rise to

macro-properties of the multigene family, among which the mean pairwise identity between genes and the identity probability distribution are very important. These macro-properties depend on the same processes that we have studied in this chapter, point mutation and gene conversion, but we will see that at the global scale, the result of their interaction is different.

Chapter 5

Quantifying global VSG archive diversification using diffusion processes

5.1 Introduction

An understanding of the roles of mutation and recombination in the generation of parasite antigenic diversity is important for addressing key questions about the evolution, adaptation and persistence of parasites. The experimental analysis of antigenic diversity generation poses a massive challenge, even in the case of parasites that have traditionally received relatively more scientific attention, such as *Trypanosoma brucei*, the sleeping sickness parasite, or *Plasmodium falciparum*, the malaria parasite. However, with the increasing availability of gene sequences derived from the *T.brucei* Genome Project (Berriman *et al.*, 2005) comes a motivation to use new data- and modelling-driven approaches to characterise functional properties of proteins and genes at different levels of organisation, to help us in the study of antigenic diversity.

The genes coding for the Variant Surface Glycoprotein (VSG), which determine the variable antigen type for trypanosomes, are crucial in sustaining chronic infection of these parasites in their vertebrate hosts. They form a multigene family, known as the VSG archive, and represent an important level of biological organisation for trypanosomes. In recent years, the structure and organisation of the VSG archive and molecular mechanisms involved therein, have been studied extensively Barry (1997);

Borst *et al.* (1997); Morrison *et al.* (2005). Among the striking results that have emerged are the large size of the VSG multigene family (Berriman *et al.*, 2005), containing more than 1600 silent VSGs, the high probability of antigenic switch per division, the high combinatorial potential for mosaic gene formation from pseudogenes and their huge diversifying potential.

Multigene families are groups of similar genes arising primarily via gene duplications. Across organisms gene duplication may occur at a rate as high as $10^{-4} - 10^{-6}$ per locus per generation. After duplication, gene conversion is a major force shaping multigene families. Gene conversion is a special type of recombination in which one segment of DNA contributes genetic information to another, making the recipient location identical to the donor, but not altering the donor sequence. As a result, sequence identity among gene family members is very high. This is certainly the case for many VSG genes in the trypanosome antigenic archive. Marcello & Barry (2007b) have shown that about 60% of VSGs are unique, the rest occur in subfamilies of two to four close homologs ($> 50\% - 52\%$ peptide identity).

Point mutation is another important process affecting each individual VSG gene randomly and independently introducing diversity to each sequence. The combined effect of gene conversion and point mutation determines the evolution of multigene family identity. Globally, gene conversion makes members of a multigene family more similar, and can help in the fast spread of beneficial mutations through all gene family members. In contrast, point mutation enhances dissimilarity between genes and promotes global diversification. However, as shown in the previous chapter, gene conversion can also act as an accelerator for the diversification of genes locally (e.g. within subfamilies).

The study of mechanisms for the generation of genetic diversity is important because such mechanisms are needed for the survival and adaptation of the parasite in its hosts. Conceivably, *T.brucei* uses such a capacity to generate a massive genetic heterogeneity in order to increase its chances to adapt to the many different hosts it encounters. The analysis of the generation of genomic variation may have concrete applications for the development of novel therapies against these parasites. For example, highly variable genes or gene parts may not be suitable drug/vaccine targets, as opposed to relatively more robust genes. Relatively more robust genetic sequences (i.e. less prone to mutation) by virtue of their smaller ability for generating diversity, might

represent a more feasible target if the gene in question is deemed essential for the survival of the pathogen. Thus, depending on the level of variation of a potential vaccine target, different strategies (e.g. focus on function domains, focus on the vector-specific surface antigens) may be considered in the absence of multivalent solutions.

To determine the diversity of VSG genes and the mechanisms that maintain it, in this chapter, we will study the combined effect of gene conversion and point mutation at the *global* level of the VSG gene archive. At this global scale, gene conversion adds identity between members of the gene family and as a consequence the VSG archive is homogenized. The opposing evolutionary force, point mutation, introduces diversity. The interplay between the two processes alters the pairwise identity between genes in a dynamic manner, and may lead to a complex archive structure, such as groups of genes of a given pairwise similarity between their members. Over the course of evolution, certain distributions of genetic identity among VSG genes in the silent archive of African trypanosomes might be preferred over others, because they might confer a higher fitness to the parasites employing that antigenic archive. On one hand, a large number of highly similar genes could facilitate mosaic gene formation, thus help prolong infection within a host. On the other hand, greater variation between members of the VSG archive could help the parasite avoid cross-reactive immune responses when these variants are sequentially expressed.

Given this context, the rates and characteristics of the processes that diversify and homogenize the silent archive, point mutation and gene conversion are very important. But can the dynamic rates of these processes be extracted from current snapshots of the VSG antigenic archive? Can we estimate these rates from the pairwise identity distribution observed today in the multigene family? To answer these questions, in the following, we apply the theory of stochastic processes and population genetics to the study of the interaction between gene conversion and point mutation, and the emerging diversification of a multigene family. First, in Sections 5.2.1 - 5.3 we investigate how macro-properties of a multigene family arise from singular interactions and random events affecting individual genes, and subsequently, in Sections 5.4-5.7 we apply population genetics and the diffusion approximation to obtain the stationary identity distribution analytically. Our final aim is to estimate the rates of point mutation and gene conversion globally, as evolutionary forces in the VSG archive of African trypanosomes. In Section 5.8, we fit the analytical stationary identity distribution to the

empirical distribution of genetic identity among the VSG N-domains. Finally, in Section 5.9 we discuss the relative contributions of the two most important evolutionary forces mediating VSG archive diversification.

5.2 Model

5.2.1 Probabilistic description

We model the stochastic occurrence of single mutation and gene conversion events in a multigene family as a Markov process, which we simulate using the Gillespie Algorithm (Gillespie, 1977). The global event rates of the two processes per unit of time are respectively defined as γ and μ. Stochastic events (mutations and conversions) are indexed $1,2,...,T \in \mathbb{Z}$, which occur at the times $t_1,t_2,...,t_T \in \mathbb{R}$. The inter-event times are exponentially distributed with mean $1/(\gamma+\mu)$ (see Figure 5.1 for an example of a realisation of the process). Only one event can happen at a time.

Point mutation is chosen with relative probability $\mu/(\gamma+\mu)$, whereas conversion events are chosen with relative probability $\gamma/(\gamma+\mu)$. The size of the gene family is denoted by N. The length of each gene by L. Point mutation occurs only on one gene out of N, at a single discrete position out of L possible positions. For simplicity, we assume that the length of each conversion is fixed and we denote it by l_c. Each conversion affects only 1 pair out of $N(N-1)/2$ possible gene pairs, where the donor and recipient gene are randomly chosen. Following the classical model for point mutation, the positions of mutations are chosen to be random uniform. Similarly, we adopt the simplest model for the occurrence of conversions: their starting points along each gene are also random uniform.

To distinguish between the genes, each gene $(1,2,..N)$ is assigned a particular color at the beginning of the simulation, at time 0. Then as time passes, if two genes convert and one of them (the recipient) copies genetic information from the other (donor), the identity and the color of the converted stretch will change accordingly. For denoting mutations, we assume the infinite allele model, that each mutation is new, and thus, adopt the convention that they can be represented by a grayscale, with the more recent mutation being assigned a darker shade.

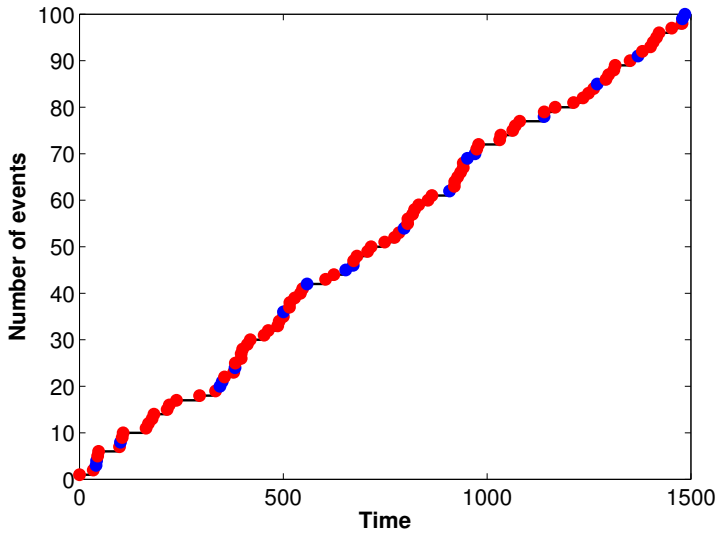


Figure 5.1: Illustration of the stochastic occurrence of events. A realization of 100 alternating mutations (blue dots) and gene conversions (red dots) affecting the multigene family. Parameters used: $\gamma = 0.5, \mu = 0.1$.

After each stochastic event, we align all gene pairs, and compute the pairwise identity between any two genes in the family. The pairwise identity is calculated by comparing the color information and the mutation information of each of the two genes. Since we assume that genes start off as identical, the color notation serves only to keep track of the genes contributing the converted segments, and does not affect their pairwise identity. As a consequence, identity is computed on the basis of the mutations only[1].

In this formulation, both mutation and conversion events are assumed to be independent of time and of the pairwise identity between gene sequences. The event rates are thus constant in gene space and time. All genes have equal probabilities to be selected for each event.

Denoting by $h_{ij}(t)$ the pairwise identity between two genes i and j ($i \neq j$) in the

[1] Alternatively, if the genes do not start off as identical, the color assignment, besides determining ancestry, can then be used to compute pairwise identity by taking into account also the initial configuration.

system at time t, we define

$$h_{ij}(t) = \frac{\text{No. identical positions at time t}}{L}, \tag{5.1}$$

which can take values between 0 and 1 in discrete steps of $1/L$. We note that $h_{ij}(0) = 1$, for all pairs (i, j). The mean pairwise identity in the archive (within an individual genome) at any time is given by the sample mean over all gene pairs:

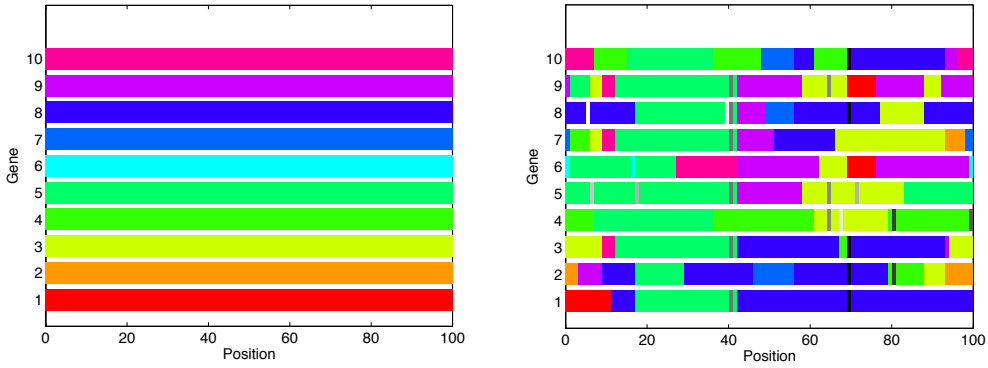
$$\bar{h}(t) = \frac{\sum_{i=1}^{N} \sum_{j>i}^{N} h_{ij}(t)}{N(N-1)/2}. \tag{5.2}$$

The sequence of $\bar{h}(t)$ for $t \in \mathbb{R}, t > 0$ gives a summary statistic from one realization of the stochastic process. Each different realization would correspond to a different genome (family of genes) starting from the same initial conditions. Denote one such realization by the index k. Then, the average of the mean pairwise identity over K stochastic realizations, representing a summary statistic of the entire population, is given by:

$$\bar{H}(t) = \frac{\sum_{k=1}^{K} \bar{h}^k(t)}{K}. \tag{5.3}$$

5.2.2 Simulation results

The simulation model is a flexible tool that allows us to experiment with concrete biological assumptions about the processes of mutation and gene conversion, and explore the consequences of those assumptions. An example of the simulated gene family is shown in Figure 5.2. By visualizing conversions through different colors, we can explore the ancestry of different gene segments in the archive, and observe how each gene becomes a mosaic gene from contributions of other members of the archive. The larger the length of each conversion l_c, relative to the gene length L, the less diverse the mosaics that we see. By this we mean, that each gene, after some time, is composed of segments that originate from only a few original donors. Instead, if the length of each conversion is small relative to L, then the mosaics that arise are more diverse and contain information from more original donors. As time passes, many early mutations may get overwritten by conversions, as well as many late mutations may get transferred across many family members.



(a) Archive initially

(b) Archive after 100 events

Figure 5.2: An illustration of archive change as a function of 100 stochastic events, comprised of partial gene conversions and point mutations. The genes take the form of mosaics as a result of multiple gene conversion events. The effect of mutation events can be seen by the gray lines denoting the individual point mutations that have not been overwritten. Parameters as in Figure 5.1, with $l_c = 10, N = 10, L = 100$.

Our main focus lies in the evolution of pairwise genetic identity within a multigene family and the role played by point mutation and gene conversion. Individually, each process drives the gene pairs toward opposing extremes: mutation towards total diversity, gene conversion towards total identity. Together however, they interact to give rise to intermediate scenarios on the continuum between 0 and 100% identity. Indeed, the mean pairwise identity changes in the system from 1 (all genes identical), at the beginning of the simulation, towards lower values due to mutation, but it does not decline to 0, because conversion events tend to increase similarity between genes (Figure 5.3). Ultimately, the mean pairwise identity will settle at an intermediate equilibrium value. This equilibrium depends on the effects of mutation and conversion events on a per-aligned nucleotide basis.

Besides the mean pairwise identity, the numerical simulations can be used to provide information about the probability distribution of identity across gene pairs. We have several identity classes in the range 0 to 1, equally spaced, with minimum identity difference $\Delta h = 1/L$. Thus, we can obtain a distribution over at most $1/\Delta h + 1$ identity classes. For each identity class s, we can compute the frequency of gene pairs sharing identity $h_s = (s-1)\Delta h$ at any particular time t. Initially all gene pairs share 100% iden-

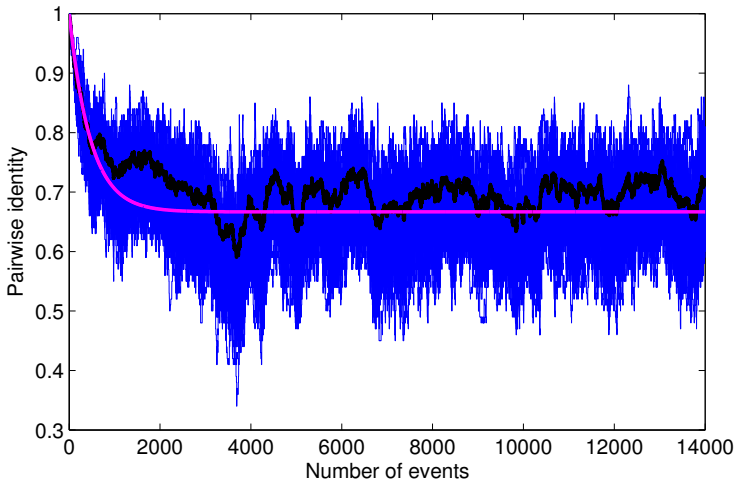


Figure 5.3: Dynamics of pairwise identity in a multigene family. Each gene pair (blue lines) can be thought to represent an independent realization of the stochastic processes of conversion and mutation. The simulation mean over gene pairs (given in black), is very well approximated by our analytical expression for $\bar{h}(t)$ (pink line) . $N = 10, L = 100, l_c = 10, \mu = 50, \gamma = 90$.

tity, implying a Dirac delta function initial condition located at the last identity class. Then, as stochastic events accumulate, the distribution spreads towards lower identity classes (Figure 5.4), until a stationary configuration is established. The exact properties of this distribution depend on the model parameters. In the example illustrated in Figure 5.4 for instance, the stationary identity distribution is centred at about 0.5, suggesting an almost perfect counterbalance of mutation and conversion processes. In the next section, we want to understand precisely how the model parameters μ, γ, l_c, L, N determine the probability distribution of pairwise identity across genes *within one genome*, and properties of its distribution such as the mean and variance.

5.3 Dynamics of identity between genes in the archive

5.3.1 Gene pairwise identity after stochastic events

The changes in identity between any two genes can be approximated by representing the process of gain/loss of identical nucleotides conceptually as a birth and death

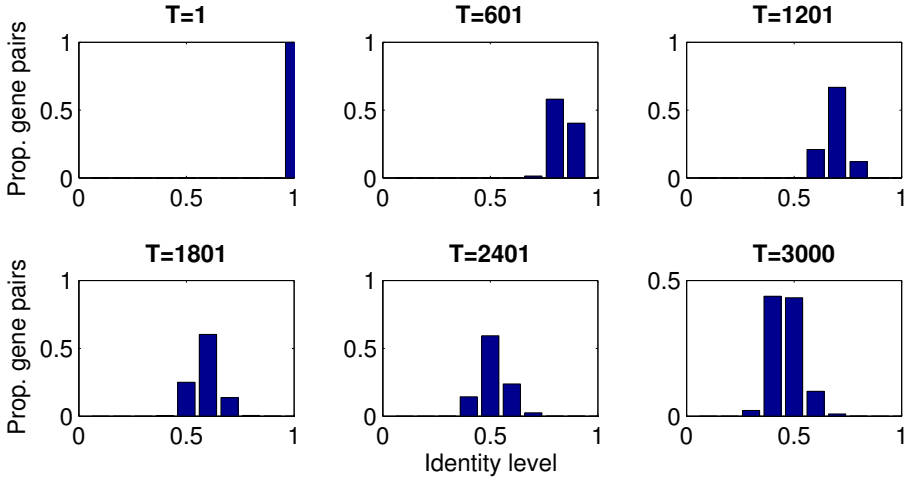


Figure 5.4: Pairwise identity distribution changes in the system. After starting off as identical, genes in the multi-gene family diversify due to mutation. T refers to the number of events that have happened in the simulation. Parameters: $N = 30, L = 100, l_c = 20, \mu = 150, \gamma = 65.25$.

stochastic process on each alignment, and by assuming that all gene pairs are independent. Denote the number of identical nucleotides shared by a gene pair after T events by $n(T)$. This gene pair can gain an identical nucleotide due to gene conversion with probability $1-h(T) = 1-n(T)/L$ per nucleotide, or lose an identical nucleotide due to point mutation with probability $h(T) = n(T)/L$ per nucleotide. The first event would imply $n(T+1) = n(T)+1$, whereas the second event would mean $n(T+1) = n(T)-1$. These two events happen at respective probabilities $\gamma/(\gamma+\mu) \times 2/N(N-1)$ and $\mu/(\gamma+\mu) \times 2/N$. Since a gene conversion affects l_c nucleotides out of a total L, whereas a point mutation only one nucleotide, the expected number of identical nucleotides in a gene pair after $T+1$ events is given by:

$$\bar{n}(T+1) = n(T) + \frac{2\gamma l_c}{(\gamma+\mu)N(N-1)}\left(1-\frac{n(T)}{L}\right) - \frac{2\mu}{(\gamma+\mu)N}\frac{n(T)}{L}. \tag{5.4}$$

With initial condition $n(0) = n_0$, and the simpler notation $\Omega = N(N-1)L/2$, the above recurrence relation can be solved, giving:

$$\bar{n}(T) = n_0\left(1-\frac{\mu(N-1)+\gamma l_c}{\Omega(\gamma+\mu)}\right)^T + \frac{\gamma l_c L}{\Omega(\gamma+\mu)} \times \sum_{i=0}^{T-1}\left(1-\frac{\mu(N-1)+\gamma l_c}{\Omega(\gamma+\mu)}\right)^i, \tag{5.5}$$

for the expected number of identical nucleotides between any two genes after T stochastic events have occurred. The mean pairwise identity follows easily from the above when it's divided by L:

$$\bar{h}(T) = h_0\left(1-\frac{\mu(N-1)+\gamma l_c}{\Omega(\gamma+\mu)}\right)^T + \frac{\gamma l_c}{\Omega(\gamma+\mu)} \times \sum_{i=0}^{T-1}\left(1-\frac{\mu(N-1)+\gamma l_c}{\Omega(\gamma+\mu)}\right)^i. \quad (5.6)$$

In Figure 5.3 we gave an example of the evolution of mean pairwise identity from model simulations. Notice that for N suitably large (e.g. $N \geq 10$) the approximation found through consideration of the birth-death process is very accurate. The steady state of the expected number of identical nucleotides between two genes, and of their pairwise identity, in the limit $T \to \infty$, are given by

$$\bar{n}^* = \lim_{T\to\infty} \bar{n}(T) = L\left[1+\frac{\mu(N-1)}{\gamma l_c}\right]^{-1}, \quad (5.7)$$

$$\bar{h}^* = \lim_{T\to\infty} \bar{h}(T) = \left[1+\frac{\mu(N-1)}{\gamma l_c}\right]^{-1}. \quad (5.8)$$

Although the approximation in Eq. 5.6 is likely to somewhat underestimate real archive identity because we neglect the effect of a conversion between i and j (e.g. i receiver gene, j donor), on the changes in identity in $N-2$ overlapping pairs (i,k), numerical observations suggest that this indirect contribution to pairwise identity is negligible for N large, because the majority of pairs in the archive will not involve genes i and j.

5.3.2 Mean pairwise identity in continuous time

Section 5.3.1 treats the pairwise identity between any two genes in the family as a quantity that changes after each event, thus the notation T stands for the discrete number of mutation and conversion events. However, this quantity may be unknown in practice, and it is more convenient to describe mean pairwise identity by expressing it as a function of *continuous time*, denoted by $t \in \mathbb{R}$, instead. To achieve this description, it is convenient to revisit Equation 5.4 and consider its continuous time analogue. We must express the dependence of $\bar{n}(t+\Delta t)$ on $n(t)$, where Δt is an infinitesimal time step. Recall that γ and μ are event rates per unit of time. Thus in the time interval $[t, t+\Delta t]$, we expect $\gamma\Delta t$ conversion events and $\mu\Delta t$ mutation events on average. After

conversion and mutation have occurred, the number of identical nucleotides between the two genes will be:

$$n(t+\Delta t) = n(t) + 2\gamma\Delta t \frac{l_c(1-n(t)/L)}{N(N-1)} - 2\mu\Delta t \frac{n(t)/L}{N}. \tag{5.9}$$

Taking the expectation, rearranging and dividing both sides by Δt gives:

$$\frac{\bar{n}(t+\Delta t) - \bar{n}(t)}{\Delta t} = \frac{\gamma l_c}{N(N-1)}\left(1 - \frac{\bar{n}(t)}{L}\right) - \frac{2\mu}{N} \times \frac{\bar{n}(t)}{L}, \tag{5.10}$$

which after taking the limit $\Delta t \to 0$ becomes:

$$\frac{d\bar{n}(t)}{dt} = \frac{2\gamma l_c}{N(N-1)}\left(1 - \frac{\bar{n}(t)}{L}\right) - \frac{2\mu}{N}\frac{\bar{n}(t)}{L}. \tag{5.11}$$

The solution of the above differential equation gives the expectation of the number of identical nucleotides between two genes as a function of continuous time:

$$\bar{n}(t) = L\frac{\gamma l_c + \mu(N-1)e^{-\frac{\gamma l_c + \mu(N-1)}{\Omega}t}}{\gamma l_c + \mu(N-1)}, \tag{5.12}$$

where we recall $\Omega = N(N-1)L/2$. The above expression can be simplified further by introducing the parametrization $c_0 = \gamma l_c/\Omega$ and $m_0 = \mu(N-1)/\Omega$, denoting respectively the per-aligned nucleotide probabilities of undergoing a gene-conversion induced change ($0 \to 1$) and mutation induced change ($1 \to 0$) per unit of time. We can derive a similar expression for the pairwise identity $\bar{h}(t) = \bar{n}(t)/L$. Finally, we have:

$$\bar{n}(t) = L\frac{c_0 + m_0 e^{-(c_0+m_0)t}}{c_0 + m_0}, \quad \bar{h}(t) = \frac{c_0 + m_0 e^{-(c_0+m_0)t}}{c_0 + m_0}, \tag{5.13}$$

which depend only on the per-nucleotide probabilities of being affected by mutation or gene conversion in one unit of time. This means that if c_0 and m_0 are fixed, the number of genes in the gene family, N, their length, L, and the conversion length, l_c, do not determine the mean evolution of pairwise identity between genes as a function of time. Our numerical simulations confirm this result. Such analytical expression for the mean pairwise identity in a multigene family can be used in many ways. For example, it allows us to predict how much time is required on average for the pairwise identity between two genes to reach a certain threshold value V, starting from given initial

conditions. In the case of the two genes starting off entirely identical, the expected time for their identity to reach V is given by:

$$E[t_V] = -\frac{1}{c_0 + m_0} \ln \left[\frac{V(c_0 + m_0) - c_0}{m_0} \right]. \tag{5.14}$$

It is also interesting to calculate the variance around the mean identity as a function of time, $var[h(t)]$. Numerical observations suggest that unlike the mean, the variance depends explicitly on parameters like the number of genes N. The mutation-only case is easier to calculate and the corresponding variance is shown in Figure 5.5. For the combined mutation-conversion case, the variance calculation is more challenging (see Appendix D.1 for details.

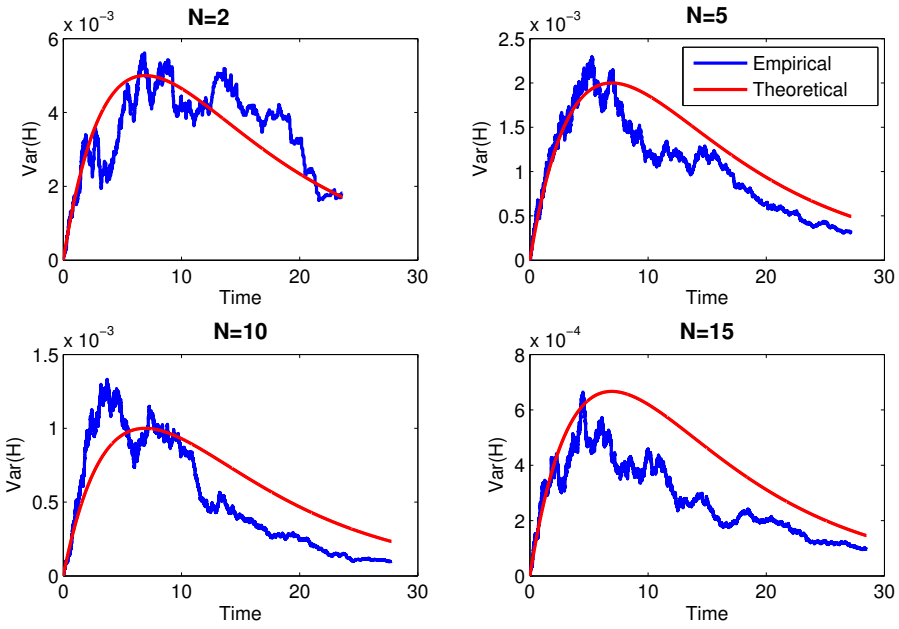


Figure 5.5: The variance around the mean identity as a function of time for mutation-only dynamics matches well with the theoretical approximation $1/L\bar{h}(t)(1-\bar{h}(t))$ given in Appendix D.1. Parameter values: $m_0 = 0.1, L = 50$.

Although, the mean pairwise identity $\bar{h}(t)$ is an important quantity in the dynamics of the family of genes, it represents only a first order summary statistic, with limited applications. For example, it is impossible to estimate, besides the ratio c_0/m_0, the

specific mutation and conversion parameters from the equilibrium value of the mean pairwise identity $\bar{h}^*$. A more complete description of the evolution of the system can be derived from the identity probability distribution, namely the probability $P(x,t) = \text{Prob}\{h_{ij}(t) = x\}$, denoting the probability that at time t, the genetic identity between any pair of genes (i, j) in the family is equal to x, where $0 \leq x \leq 1$.

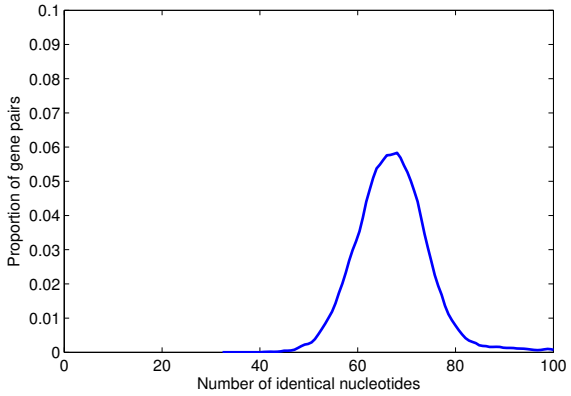


Figure 5.6: The stationary distribution of pairwise identity in the gene family. After starting off all as identical, gene pairs in the multi-gene family reach varying levels of identity as stochastic events accumulate. In the long time limit, the proportions of gene pairs in identity classes do not change: $P(x,t) \to P^*(x)$, shown here. Parameters as in Figure 5.3

Such a probability distribution must inevitably depend on the same process parameters that we have seen so far: γ, μ, L, l_c, N, and will surely contain much more information about the pairwise identity characteristics of the group of genes in consideration. In order to obtain this distribution analytically, we resort to the population genetics framework and diffusion approximation in the coming sections. We plan to use the analytical formula for $P(x,t)$ and VSG data to estimate the parameters that characterize the mutation and gene conversion process in a typical gene archive.

5.4 Pairwise identity and the Wright-Fisher model

In order to work towards the identity distribution $P(x,t)$, we consider the pairwise identity in a family of N genes from another perspective, drawing parallels with the classical Wright-Fisher model (Fisher, 1930; Wright, 1931). The total number of aligned

nucleotides in a gene pair is given by L. Each position in any aligned pair of genes consists of a 0 or 1, depending on whether the two genes are different or identical at that nucleotide[1]. Denote the total number of identical positions as Y, and the number of different positions is thus $L-Y$. The pairwise identity in this gene pair is then Y/L, the frequency of the identical positions.

Assuming that at each generation the family of genes reproduces itself, - by this we mean that the number of genes remains constant, as does the length of each gene L,- we expect that the relative frequencies of identical and different positions on each gene pair change continuously (see Figure 5.7).

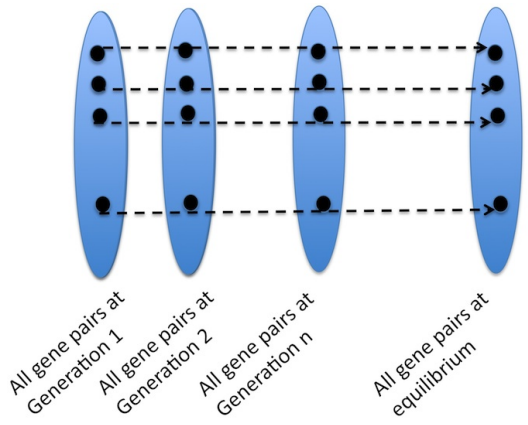


Figure 5.7: Schematic illustrating the evolution of gene pairs in the multigene family model. Each dashed arrow represents an idealized 'independent' trajectory of an arbitrary pair (depicted by black circles). As the number of generations tends to infinity, a stationary distribution in the distribution of pairwise identity is reached, whereby the probability of observing a given identity in a random pair is constant.

When a gene experiences random point mutation, some identical nucleotides in pairwise alignments in the previous generation change in the next generation ($1 \rightarrow 0$). When a gene pair experiences gene conversion, some different nucleotides in their pairwise alignment in the previous generation become identical in the next generation ($0 \rightarrow 1$). Denote by m the probability of mutation per aligned nucleotide per generation, and by c the probability of gene conversion per aligned nucleotide per generation. We

[1] In this parallel with the Wright-Fisher model, the L aligned positions stand for 'haploid individuals' in the classical Wright-Fisher framework, and 0/1 stand for 'alleles'.

want to track the changes in identical/non-identical nucleotide frequencies over time, i.e. changes in pairwise identity between two genes.

So this formulation is similar to the basic Wright-Fisher model (Fisher, 1930; Wright, 1931) for a locus with two alleles (0/1) in a population of size L. Since m and c are not necessarily equal, the mutation process between the two alleles is asymmetric, yet it allows us to apply the same analytical framework. There are three dynamic components affecting the frequency of identical nucleotides between two genes at each generation in our model: the directional drive of mutation toward 0 identity, the directional drive of gene conversion toward total identity, and the process fluctuations arising from random sampling of positions ('individuals') in a population of size L. The latter component would correspond to genetic drift in the classic Wright-Fisher model.

5.4.1 Link to simulation model

Recall from the previous sections that the conversion and mutation event rates in the gene family are given by γ and μ. In other words, γ and μ are the expected number of events of a given type per unit of time. A natural question is: how are γ and μ related to c and m in this new formulation? In fact, c and m denote the probability of $0 \rightarrow 1$ or $1 \rightarrow 0$ transition per generation per aligned nucleotide. By generation we mean one division cycle of the parasite, during which all DNA is replicated. So, during one generation many conversion and mutation events may have occurred.

Denoting generation time by τ, the probability of a mutation per nucleotide ($1 \rightarrow 0$) per unit of time becomes m/τ (= m_0 in Section 5.3.2), similarly c/τ (= c_0 in Section 5.3.2) for the probability of conversion ($0 \rightarrow 1$). A mutation in an aligned nucleotide ($1 \rightarrow 0$) could have occurred in either gene of a pair, thus the expected number of mutational changes in one gene per unit of time is given by $mL/2\tau$. Since there are N genes in the system, the total number of mutation events, in the family, per unit of time is given by $\mu = NmL/2\tau$. Similarly, the expected number of conversion changes per unit of time in one pair is given by $cL/l_c\tau$. Since there are $N(N-1)/2$ gene pairs in the system, the total conversion event rate per unit of time becomes $\gamma = cLN(N-1)/2l_c\tau$.

We notice that through this new parametrization, the mean pairwise identity at equilibrium between mutation and gene conversion (Eq. 5.8), becomes:

$$\bar{h}^* = \frac{1}{1+\frac{\mu(N-1)}{\gamma l_c}} = \frac{1}{1+\frac{m}{c}} = \frac{c}{c+m}, \tag{5.15}$$

a non-dimensional version of the previous formula, independent of conversion lengths, number of genes, length of each gene, etc. The final pairwise identity depends only on the mutation and conversion probabilities per nucleotide per generation; this allows for the maintenance of both diversity and identity in the population of nucleotides.

5.5 The diffusion approximation

To obtain the probability distribution of pairwise identity in the gene family, $P(x,t)$, in the following, we first present the general framework of diffusion processes that can be applied to the frequency of identical nucleotides as a function of time, and subsequently, calculate the precise parameters (infinitesimal mean and variance) associated to this process, dependent on the conversion-mutation interplay in our model.

Genetic drift, is an evolutionary force that tends to decrease the variation in a population and can impact the effectiveness of mutation and selection. One of the key mathematical approaches to deal with genetic drift is the diffusion approximation. Introduced into population genetics by Fisher (1922) and Wright (1945), this approach was substantially extended and developed by Kimura (1955b). Under this approximation, the proportion of individuals of a particular genetic type is treated as a continuous random variable whose distribution follows a diffusion equation. This approach has been applied to derive results that lie at the core of population genetics (Crow & Kimura, 1970).

In simple terms, the diffusion process is a continuous-time stochastic process describing a quantity that changes continuously over time and whose future depends only on the current state. In our case, the quantity is the pairwise identity between two genes in a multi-gene family, or the frequency of identical positions between them in a total of L positions. The motivation for using a diffusion process as an approximation of genetic models for large but finite populations comes from the observation that many finite-size models, when viewed on the right timescale, have the property that as the

population size goes to infinity the change in relative gene frequencies is continuous, and results in a well-defined process, which is easier to study.

Let us denote the diffusion process by $\{X(t) : t \geq 0\}$. Here t denotes time and $X(t)$ the state of the system/gene pair at time t. In this case, $X(t)$ is the relative frequency of identical positions in the total pool of L positions in one pairwise alignment. A diffusion process is characterized by two basic quantities: the mean and the variance of the infinitesimal displacement, called drift and diffusion respectively. The displacement during the time interval $(t, t+\delta t)$ is denoted by $\Delta X(t) = X(t+\delta t) - X(t)$. Then the drift parameter is defined as:

$$a(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} E[\Delta_{\delta t} X(t) | X(t) = x]. \tag{5.16}$$

The diffusion parameter is defined as

$$b(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} E[(\Delta_{\delta t} X(t))^2 | X(t) = x]. \tag{5.17}$$

For small δt, the mean of the displacement $\Delta_{\delta t} X(t)$ during the time interval $(t, t+\delta t)$ is $a(x,t)\delta t$ since:

$$E[\Delta_{\delta t} X(t) | X(t) = x] = a(x,t)\delta t + o(\delta t). \tag{5.18}$$

Similarly, $b(x,t)\delta t$ is approximately the variance of the displacement $\Delta_{\delta t} X(t)$ during the time interval $(t, t+\delta t)$ because:

$$\begin{aligned} \mathrm{var}[\Delta_{\delta t} X(t) | X(t) = x] &= E[(\Delta_{\delta t} X(t))^2 | X(t) = x] - (E[\Delta_{\delta t} X(t) | X(t) = x])^2 \\ &= b(x,t)\delta t - (a(x,t)\delta t)^2 + o(\delta t) \\ &= b(x,t)\delta t + o(\delta t) \end{aligned} \tag{5.19}$$

Once the drift and diffusion are known, one can explicitly write $P(p_0, x; t)$, the conditional probability density that the state of the system is x at time t given that it was p_0 at time 0. The following then holds in the diffusion limit:

$$\frac{\partial P(p_0,x;t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}[b(x,t)P(p_0,x;t)] - \frac{\partial}{\partial x}[a(x,t)P(p_0,x;t)]. \tag{5.20}$$

The above is known as the Kolmogorov forward equation or Fokker-Planck equation and may be derived in different ways from the actual stochastic dynamics underlying $X(t)$. The question is: what is $a(x,t)$ and $b(x,t)$ for an arbitrary gene pair in our model?

If we can quantify precisely $a(x,t)$ and $b(x,t)$, we will have formulated a model that tracks the dynamic changes in pairwise identity in the multigene family, through Eq. 5.20. In the following, we present a simple way of deriving the infinitesimal mean and variance of $X(t)$ from basic principles and some scaling arguments.

5.6 Drift and diffusion for the mutation-conversion model

Here we will calculate the diffusion limit of a haploid Wright-Fisher model with asymmetric "mutation" (mutation and gene conversion). Mutations in a gene pair ($1 \to 0$) occur with probability m per nucleotide per generation. Conversions in a gene pair ($0 \to 1$) occur with probability c per nucleotide per generation. Denote by $Y(n)$ the number of identical positions in the pairwise alignment of length L at generation n. From random sampling, the transition probability is given by:

$$P(Y(n+1) = j | Y(n) = i) = \binom{L}{j} \phi_1^j (1-\phi_1)^{L-j}, \tag{5.21}$$

where

$$\phi_1 \;=\; \frac{i}{L}(1-m) + \frac{L-i}{L}c \tag{5.22}$$

is the proportion of nucleotides that are of type 1 (identical) after mutation and conversion have occurred. To compute the drift and the diffusion, we study the scaled process where $\delta t = 1/L$:

$$X_L(t) = \frac{Y(\lfloor Lt \rfloor)}{L}, t > 0$$

where $\lfloor Lt \rfloor$ denotes the largest integer less than or equal to Lt. The infinitesimal drift parameter can be found by first computing:

$$\begin{aligned} LE\left[X_L\left(t+\frac{1}{L}\right) - X_L(t) | X_L(t) = \frac{i}{L}\right] &= E[Y(\lfloor Lt \rfloor + 1) - i | Y(\lfloor Lt \rfloor) = i] \\ &= L\phi_1 - i \\ &= i(1-m) + (L-i)c - i \\ &= -mi + (L-i)c. \end{aligned} \tag{5.23}$$

In this formulation, the limit $L \to \infty$ is equivalent to the limit $\delta t \to 0$. We scale also the mutation and conversion parameters, namely we assume that $\lim_{L\to\infty} Lm = \theta$ and

$\lim_{L\to\infty} Lc = \sigma$. This means we expect the approximation to be good when the gene length (system size), L, is of the order of the reciprocal of the conversion/mutation probability per nucleotide[1]. We can then use $x = i/L$ to obtain:

$$a(x) = \lim_{\delta t\to 0} \frac{1}{\delta t} E[X_L(t+\delta t) - X_L(t) | X_L(t) = x] = -\theta x + (1-x)\sigma \tag{5.24}$$

To find the infinitesimal diffusion parameter, we compute:

$$\begin{aligned} \Omega\, E\left[\left(X_L\left(t+\frac{1}{L}\right) - X_L(t)\right)^2 | X_L(t) = \frac{i}{L}\right] &= L\frac{1}{L^2} E\left[\left(Y(\lfloor Lt\rfloor + 1) - i\right)^2 | Y(\lfloor Lt\rfloor) = i\right] \\ &= \frac{1}{L} L\phi_1(1-\phi_1), \end{aligned} \tag{5.25}$$

given that $L\phi_1(1-\phi_1)$ is the variance of the binomial distribution with parameters L and ϕ_1. Taking the limit $L\to\infty$, where $\theta = Lm$ and $\sigma = Lc$, with $x = i/L$, we see that $\lim_{L\to\infty}\phi_1 = x$. So, for the infinitesimal diffusion parameter we obtain:

$$b(x) = x(1-x) \tag{5.26}$$

The diffusion approximation in this case then becomes:

$$\frac{\partial P}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\Big[x(1-x)P\Big] - \frac{\partial}{\partial x}\Big[\Big(-\theta x + \sigma(1-x)\Big)P\Big], \tag{5.27}$$

and it describes how the distribution of pairwise identity changes over time in a multigene family. Because no probability mass enters or leaves the system at either boundary, it is appropriate to assume no flux boundary conditions for this partial differential equation. We also assume an initial condition of the form $P(x,0) = \delta(x-p)$, with all probability mass, i.e. all gene pairs, residing in the same identity class p (e.g. $p = 1$). The solution of this equation has been obtained by Crow & Kimura (1956) through the study of the moments of the distribution, and has been found to agree with the fundamental solution with zero-flux boundary conditions derived by Goldberg (1950) in his unpublished thesis. Assuming a Dirac delta function initial condition of the form $P(x,0) = \delta(x-p)$, the time dependent solution is given by:

$$P(x,t) = \sum_{i=0}^{\infty} G_i(x) \exp\left[-i\Big(\sigma + \theta + (i-1)/2\Big)t\right] \tag{5.28}$$

[1] Indeed, it is common practice to replace the limit by an equal sign in such cases: $\theta = Lm, \sigma = Lc$.

where

$$\begin{aligned} G_i(x) &= x^{2\sigma-1}(1-x)^{2\theta-1} \times \\ &\times F(2\theta+2\sigma+i-1,-i,2\theta,1-x)F(2\theta+2\sigma+i-1,-i,2\theta,1-p) \\ &\times \frac{\Gamma(2\theta+i)\Gamma(2\theta+2\sigma+2i)\Gamma(2\theta+2\sigma+i-1)}{i!\Gamma^2(2\theta)\Gamma(2\sigma+i)\Gamma(2\sigma+2\theta+2i-1)}, \end{aligned} \tag{5.29}$$

where F is the regularized confluent hypergeometric function and Γ is the gamma function.

5.7 Stationary identity distribution

As the probability distribution of pairwise identity within the gene family changes dynamically following Equation 5.27, an important case arises when the distribution of identical nucleotides reaches an equilibrium and the proportions of gene pairs in different identity classes do not change anymore. This equilibrium does not depend on the initial conditions, is the limit of $P(x,t)$ when $t \to \infty$, denoted by $P^*(x)$, and can be found by solving:

$$0 = \frac{1}{2}\frac{d^2}{dx^2}\Big[x(1-x)P^*(x)\Big] - \frac{d}{dx}\Big[\Big(-\theta x+\sigma(1-x)\Big)P^*(x)\Big]. \tag{5.30}$$

Integration of the above equation yields:

$$P^*(x) = Kx^{2\sigma-1}(1-x)^{2\theta-1}, \tag{5.31}$$

where K is a normalizing constant, such that $\int_0^1 P^*(x) = 1$. Since $\theta, \sigma > 0$, we find

$$K = \Gamma(2\sigma+2\theta)/\Gamma(2\sigma)\Gamma(2\theta),$$

where Γ denotes the gamma function. An illustration of different types of stationary distributions, given by Eq.5.31, is shown in Figure 5.8. Notice that Equation 5.31 is a special case of a much more general class of equations of allele distributions derived by Wright (1930) and later revisited by Kimura (1955a,b) in connection to diffusion processes. To obtain the mean pairwise identity in the gene family, one can compute:

$$\bar{h}^* = \int_0^1 Kx^{2\sigma-1}(1-x)^{2\theta-1}x\,dx = \frac{\Gamma(2\sigma+2\theta)\Gamma(1+2\sigma)}{\Gamma(2\sigma)\Gamma(1+2\sigma+2\theta)} \tag{5.32}$$

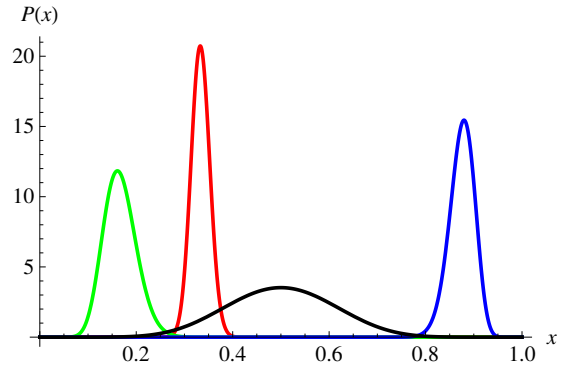


Figure 5.8: The stationary identity distribution $P^*(x)$ can take many forms depending on the values of θ and σ: $\theta = 200, \sigma = 100$ (red line), $\theta = 50, \sigma = 10$ (green line), $\theta = 10, \sigma = 70$ (blue line), $\theta = 5, \sigma = 5$ (black line). The relative magnitudes of σ and θ control the mean of the distribution: the higher θ/σ, the closer to 0 the mean, the lower θ/σ, the closer to 1 the mean. The absolute magnitudes of these parameters, instead control the variance of the distribution: the lower the magnitudes of θ and σ, the more spread the shape of the distribution, and viceversa.

Expressing the gamma functions in terms of factorials $\Gamma(z) = (z-1)!$, we get:

$$\bar{h}^* = \frac{(2\sigma+2\theta-1)!(2\sigma)!}{(2\sigma-1)!(2\sigma+2\theta-1)!} = \frac{2\sigma}{2\sigma+2\theta} = \frac{\sigma}{\sigma+\theta}, \tag{5.33}$$

thus agreeing with the previous theoretical approximations for equilibrium identity in Eq. 5.15, given that $\theta/\sigma = m/c$. Notice that the variance expected around the equilibrium h^* can be calculated easily from 5.31 and is given by:

$$\mathrm{Var}(h^*) = \frac{\sigma\theta}{(\sigma+\theta)^2(1+2\sigma+2\theta)} = \frac{cm}{(c+m)^2[1+2L(c+m)]}. \tag{5.34}$$

Notice that the variance of the pairwise identity distribution at equilibrium depends on the gene length L. In particular, when L increases, the variance of the distribution $P^*(x)$ goes down.

5.7.1 Adding selection

So far, we have considered only two evolutionary processes: mutation and gene conversion, and their interplay with random genetic drift in shaping the genetic identity distribution in a multi-gene family. However, there is another important process, namely

selection, which may act on the evolutionary dynamics of the pool of aligned nucleotides, by providing a greater fitness advantage to nucleotides of one type versus another. In our model, the aligned nucleotides can take only two types: 0 and 1, i.e. they can be different or they can be the same. Mosaic gene formation in African trypanosomes requires high identity between the participating gene sequences (Marcello & Barry, 2007a). So, a natural question is: is there a selective advantage for aligned nucleotides to being identical over being different?

Assume that aligned nucleotides of identical type in a gene pair have a selective advantage with selection parameter s compared to nucleotides of different type. Without going into the details of the mathematical derivation, which is very similar to the steps in Section 5.6, we have that in the presence of selection, the expression for ϕ_1 (Eq. 5.22) changes to:

$$\phi_1 = \frac{i(1-m)(1+s)+(L-i)c}{(1+s)i+L-i}. \tag{5.35}$$

Using appropriate scalings for the mutation, conversion and selection parameters ($\theta = Lm$, $\sigma = Lc$ and $\alpha = Ls$), we obtain the following expressions for the infinitesimal drift and diffusion terms, in the limit $L \to \infty$: $a(x) = \alpha x(1-x) - \theta x + \sigma(1-x)$ and $b(x) = x(1-x)$. This leads to a different equation for the probability density function, which has a new stationary distribution, dependent also on the selection term:

$$P^*(x) = \frac{e^{2\alpha x} x^{2\sigma-1}(1-x)^{2\theta-1}}{\Gamma(2\sigma)\Gamma(2\theta)\mathcal{H}(2\sigma, 2(\sigma+\theta), \alpha)}, \tag{5.36}$$

where $\mathcal{H}$ is the regularized hypergeometric function. Examples of typical forms of this stationary distribution are given in Figure 5.9, where the effect of positive and negative selection can be seen. Positive selection on identical nucleotides pushes the stationary distribution to the right, as expected, whereas negative selection on identical nucleotides pushes the distribution to the left, whichever the combination of mutation and conversion parameters.

We note however that selection is probably not that relevant in the evolutionary dynamics of pairwise identity in a multigene family, as the shape of the stationary distribution remains largely unaffected by additional selection on 0/1 aligned nucleotides, suggesting that the same stationary distribution can be obtained by changes in mutation and conversion parameters only, without the need to invoke explicitly selection

on identity *per se*[1]. Furthermore, if one wants to estimate all three parameters σ, θ, α from the empirical stationary identity distribution in a gene family, the problem of non-identifiability arises, where there are many combinations of these three evolutionary parameters that can give rise to the same distribution.

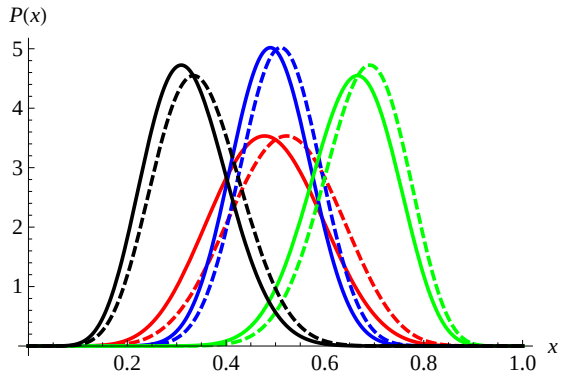


Figure 5.9: Stationary distribution of genetic identity in the presence of selection. The solid lines refer to negative selection ($\alpha = -0.8$), the dashed lines refer to positive selection ($\alpha = 0.8$). The red and the blue lines depict scenarios with equal mutation and conversion parameters, $\theta = \sigma = 5$, and $\theta = \sigma = 10$ respectively. The green and the black lines depict scenarios with conversion or mutation dominating, $\sigma = 10, \theta = 5$, and $\sigma = 5, \theta = 10$ respectively.

In the next section, we neglect selection, and assume that the trypanosome VSG archive evolves primarily and most significantly under the effects of mutation, gene conversion and neutral drift. These processes, by acting over many generations of this parasite, among other things, have shaped the distribution of pairwise identity between VSG genes. Given the estimated age of *T. brucei* of the order of hundreds of millions of years, one could assume that the current configuration of the VSG archive represents a stationary state, reached through the balance between mutation and gene conversion as time t tends to infinity. The availability of VSG genetic data today enables us to estimate the rates of these processes and thus quantify their relative interaction.

[1]It is likely, however, that natural selection operates on the rates of mutation and gene conversion, to drive the gene family towards optimal genetic identity configurations.

5.8 Fitting the diffusion model to VSG archive data

As a first step, we show in Figure 5.10 that the stationary probability density for the pairwise identity within a multigene family (Eq. 5.31) fits well to the empirical stationary distribution obtained through the simulation model presented earlier. We use the same parameters as in Figures 5.3 and 5.6. The quality of the fit, especially when the conversion length is small and the number of genes is large, suggests that the independence between gene pairs and other assumptions in the diffusion approximation are reasonably met by the simulation model under these conditions. An interesting difference lies in the variances of the two distributions, with the distribution coming from simulations displaying a larger variance than the diffusion approximation. This is possibly related to the conversion length $l_c > 1$ in the simulations, causing greater jumps in pairwise identity than what is assumed by the diffusion model, and to the indirect effects of conversion events on third party gene pairs, contributing to the right-end tail of the actual distribution (see Figure 5.11).

We explained in Section 1.2.3 that the VSG N-terminal domains come in two important subfamilies nA and nB. The two subfamilies have respectively 412 and 362 genes (Marcello & Barry, 2007a). We have obtained the pairwise identity distribution by aligning and comparing only the N-domains of the member genes. Analysis of these N-domains shows that each N-domain varies in length between 900 and 1050 nucleotides, motivating our approximation of gene length by the mean $L = 975$.

The availability of the empirical identity distribution in these two VSG subfamilies offers an opportunity to use the diffusion approximation model analyzed so far, in order to extract the mutation probability per nucleotide (m), and conversion probability per nucleotide (c) per generation. An important aspect of our approach is the first assumption that these parameters are constant over time, thus the VSG archive configuration we see today results from two time-homogeneous processes. Secondly, the diffusion approximation implicitly requires the mutation and conversion probabilities to be of the order of L^{-1}. Furthermore, we would expect m and c to be constant also across the VSG archive, thus resulting in the same parameter estimates for nA and nB genes. However, the presence of these two defined N-domain types gives us a good opportunity to test this hypothesis.

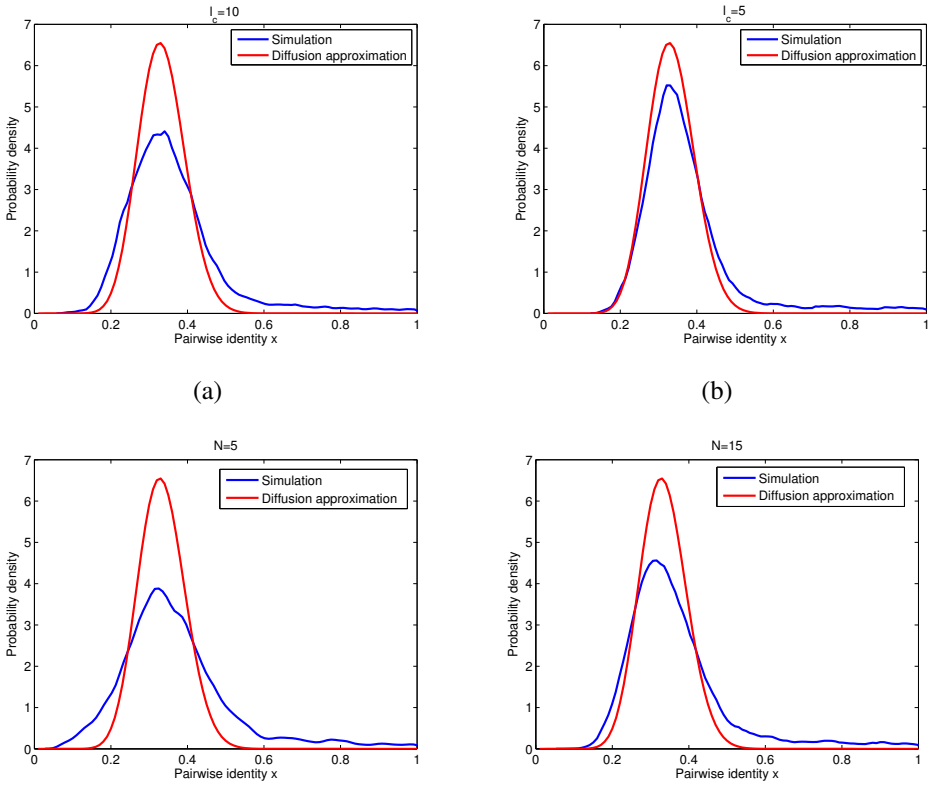


Figure 5.10: The stationary probability distribution of genetic identity from the simulation of conversion and mutation events within a gene family and the diffusion approximation. The quality of the fit (Eq. 5.31) improves with decreasing conversion length, l_c (hence l_c/L) in our simulations, and increasing number of genes in the family, N. The probability density obtained from the simulation represents the proportion of pairs sharing a given identity level at equilibrium. Parameter values: $c = 0.2, m = 0.1, \tau = 1, L = 100$, and a) $N = 10$, b) $l_c = 10$.

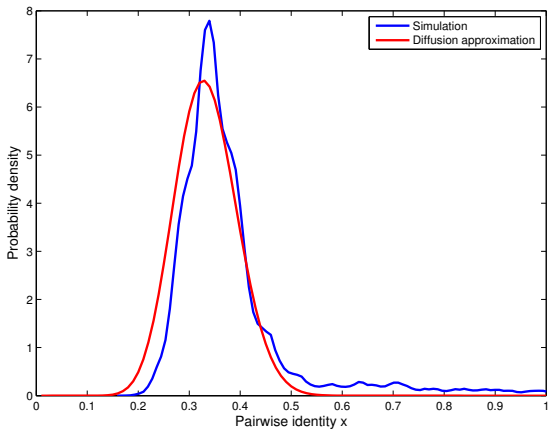


Figure 5.11: Diffusion approximation vs. model simulation for $l_c = 1$. Parameter values: $c = 0.2, m = 0.1, \tau = 1, L = 100, N = 20$. This represents a best-case scenario for the two stationary distributions to match, because each conversion tract is only 1 nucleotide long. However, because of the intrinsic lack of pure independence between gene pairs in multigene family evolution, the numerical simulations exhibit a longer right-tail in the stationary distribution than what's predicted by the diffusion approximation.

We estimate the evolutionary parameters m and c by fitting the empirical identity distribution of VSG gene pairs from the two subfamilies nA and nB, to the theoretical formula for $P^*(x)$ given by Eq. 5.31. The fit is performed using a nonlinear least square routine in Matlab. The maximum-likelihood estimates of m and c are the ones that minimize the sum of squared deviations of the empirical distribution from the theoretical formula. The result of the fit is presented in Figure 5.12.

Our estimated ratio m/c equals approximately 5.4, providing direct evidence for the relative dominance of mutation versus gene conversion in the N-domains of VSG genes, independently of their type. Mutation happens in the gene family on average at a probability of 22.24×10^{-3} per base pair per generation, while conversion happens on average at a probability of 4.14×10^{-3} per base pair per generation, implying that genetic diversification exceeds homogenization, at least by one order of magnitude.

To convert our estimates into mutation and conversion probabilities *per gene*, we must divide the current values by 2. Thus, the mutation probability per nucleotide per gene per generation is 11.12×10^{-3}, whereas the conversion probability per nucleotide

per gene per generation is 2.07×10^{-3}. Note that across the two subfamilies, nA and nB, the specific estimates do not differ much, confirming our previous expectation that the mutation and gene conversion processes occur at the same rates across all VSG N-domains.

The average pairwise identity computed from the empirical distribution is 24.5% for N-domains of nA genes, and 27.4% for N-domains of nB genes. These values are higher than the corresponding fit estimates from Eq. 5.33, $\bar{h}^* = c/(c+m) = 15.07\%$ for nA, and 16.49% for nB. This discrepancy is expected and may be due primarily to the fact that the mean pairwise identity is just one number, reducing the information contained in the whole distribution to a point estimate. Other factors may play a role, such as assumptions required for the diffusion approximation that are not wholly met by the data. For example, the length of a typical conversion tract, being higher than 1 and potentially variable, should increase the variance of identity change and its mean at each conversion event, in a biologically realistic scenario, due to indirect effects on other gene pairs. Such effects are masked under the assumption of independence between gene pairs, required in the diffusion approximation. Other factors can be the bin size in the empirical distribution ($\Delta_x = 1\%$ identity), which may still be too large for the continuous diffusion approximation to hold exactly, or the value of the population size/gene length $L = 975$, which may be relatively small.

5.9 Discussion

In this chapter, we have presented a modeling framework for studying the global genetic diversification of a gene family (VSG genes in African trypanosomes) shaped by forces such as mutation, gene conversion and genetic drift that results from random sampling of nucleotides. Although the mathematical techniques employed were not new, the application of this genetic theory to the VSG antigenic archive of African trypanosomes had not been attempted before. By using a diffusion equation as a large population size approximation, we were able to infer the rates of the mutation and gene conversion processes for the two main subfamilies of N-terminal VSG domains.

Without needing explicit assumptions about gene conversion tract lengths, mechanisms underlying gene conversion or the precise type of point mutations, our model estimated the per-nucleotide conversion probability per generation and the per-nucleotide

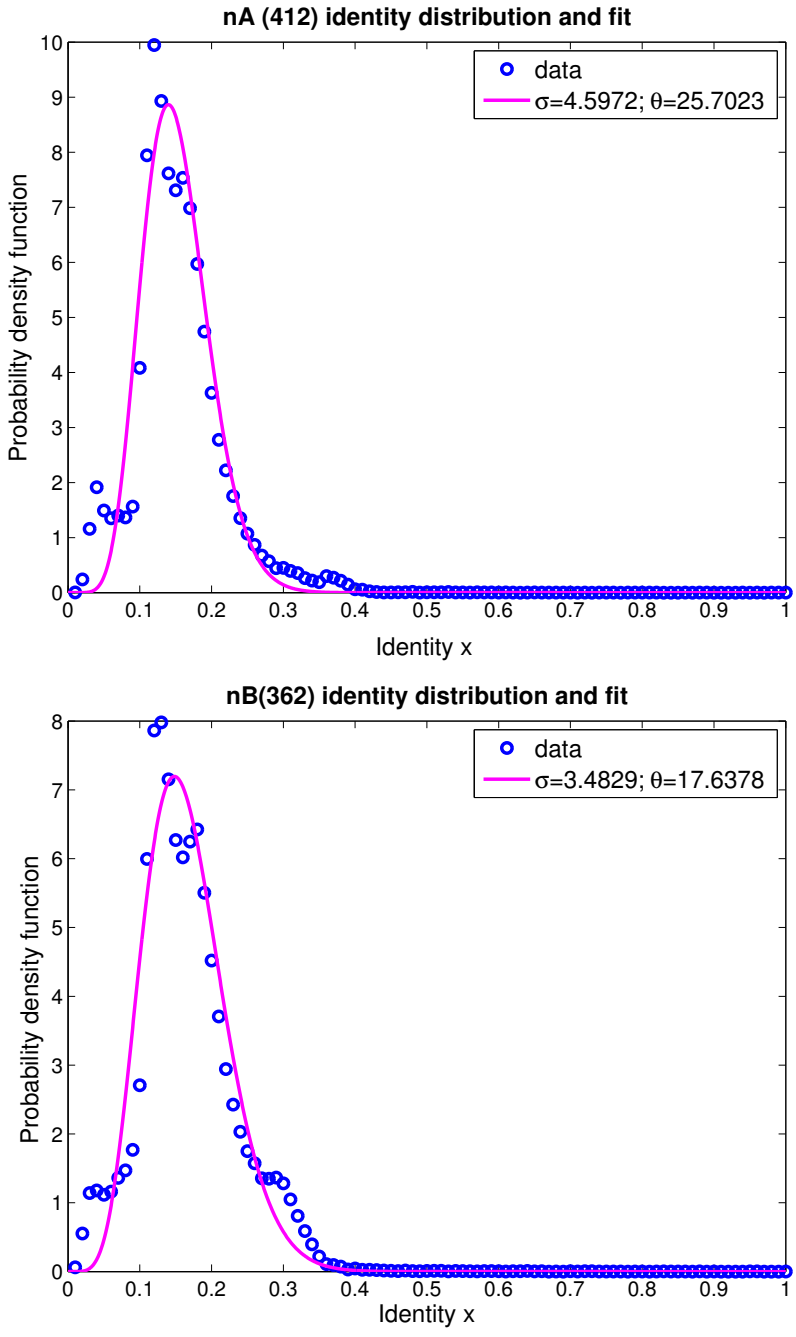


Figure 5.12: The empirical VSG identity distribution within one trypanosome genome for two subfamilies (nA, nB), and the diffusion approximation fit. The best model fit (purple line) was found using nonlinear least squares optimization routines in MATLAB. The left-skewness of the distribution is reflected in the dominance of the mutation process in both cases where $\theta \gg \sigma$. When we scale σ and θ by the inverse of L, we obtain $c = 4.71 \times 10^{-3}, m = 26.4 \times 10^{-3}$ for nA, and $c = 3.57 \times 10^{-3}, m = 18.09 \times 10^{-3}$ for nB, resulting in concrete estimates of the probabilities of mutation and conversion per aligned nucleotide per generation.

mutation probability per generation. The only requirement for the application of the theory of diffusion processes was the assumption that the total population size of aligned nucleotides in the family is approximately of the order of the reciprocal of the mutation and conversion probabilities. Lacking the precise evolutionary time information, we assumed that the VSG genes have been evolving for a sufficient number of generations such that the pairwise identity distribution we observe in the VSG archive today represents a stable equilibrium distribution. This allowed us to fit the theoretical stationary distribution to the empirical distribution and extract the evolutionary parameters of interest. Interestingly, if we scale our average mutation probability per gene per generation by the average number of gene pairs in each subfamily, we obtain 1.5×10^{-7}, a value which is in very close agreement with independent estimates of the mutation rate per base per generation ($4.5 \times 10^{-8} - 1.5 \times 10^{-7}$), obtained by Lindsey Plenderleith (pers. comm.) using detailed genetic comparisons of two *T.brucei* genomes from 1960 and 1977. This suggests a clear link between the two approaches.

Reassuringly, we notice that the parameters extracted for the two VSG N-terminal domain groups, nA and nB, do not differ much from each other, implying that diversification and homogenization processes proceed at the same rate across VSG genes of African trypanosomes. The relative values of mutation and conversion probabilities reveal that an elevated mutation rate is the primary factor making the VSG N-terminal domain a very variable genomic region. As mentioned earlier, this is likely to have important implications for antigenic variation within hosts in the chronic stages of infection, by aiding new mosaic gene formation.

5.9.1 The link with the hidden Markov model

In contrast to this chapter, in the previous chapter we considered local genetic diversification of recently duplicated genes, under the influence of point mutation and gene conversion with older donor genes in the VSG archive. The hidden Markov model approach that we applied to high-identity VSG triplets (N-domains) enabled us to quantify the density of mismatches in converted gene segments, the rate of conversion with partners outside the subfamily, the rate of mutation and the typical conversion length distribution (see Table 4.5 for example).

We expect that the average density of mismatches in converted segments in the HMM, μ, either corresponds or is related to $1-\bar{h}^*$, in the diffusion approximation framework, where $\bar{h}^*$ is the mean pairwise genetic identity in the VSG archive at equilibrium. However, from Chapter 4, we notice that $\mu \approx 0.25$ and from this Chapter we have $1-\bar{h}^* \approx 0.74$. The fact that $\mu < 1-\bar{h}^*$ suggests two possibilities: 1) either conversion with genes outside subfamilies (in the HMM) is biased towards higher-identity donors, as opposed to the entire archive; or 2) μ reflects the result of global archive diversity minus conversion within the subfamily itself. The latter process was entirely overlooked in the HMM formulation, but its effect would be to reduce the diversity throughout the aligned gene sequences and in particular on converted segments. Although we cannot determine which one among the two hypotheses is correct, both of them represent interesting scenarios to be explored further.

Another quantity at the interface between the two approaches (HMM-local diversification and Diffusion equation -global diversification) is divergence time, or the number of generations that have passed since any two genes in the high-identity subfamily/triplet shared a common ancestor. So far, the time information has been missing from the HMM approach. Instead, our diffusion approximation and corresponding fit, yielded estimates for the mutation and conversion probabilities per base pair per generation. Suppose, the number of generations separating two genes in a triplet is denoted by W. Then, if we account for two types/scales of conversion in the VSG archive: *within-subfamily* and *outside-subfamily*, the global probability of conversion per base pair per generation, c_{diff} ($= c$ in the diffusion model), could be written as the sum of the two processes weighted by the relative sizes of the subfamily (r) and the rest of the archive ($1-r$):

$$\begin{aligned} c_{diff} &= \frac{1}{W}\left(r(1-\bar{h}^*-\mu)+(1-r)(\frac{\bar{N_C}l_c}{L}\right) \\ &= \frac{1}{W}\left(r(1-\bar{h}^*-\mu)+(1-r)\frac{\lambda_{begin}\lambda_{end}l_c}{\lambda_{begin}+\lambda_{end}}\right) \end{aligned} \tag{5.37}$$

where $\lambda_{begin}, \lambda_{end}, \mu$ are as defined in Chapter 4 and $r = 15/387$. The assumption of a molecular clock is implicit, and we can use the results of Section 4.2.2 for the number of conversion events in the HMM. Similarly we could compare the mutation probabilities across the two formulations. Using the results from the HMM (Model 4, reference

gene pair 1) in Chapter 4, and the above formula, a first heuristic estimate for the number of generations separating genes Tb927.5.5260 and Tb09.160.0100 from their most recent common ancestor is: $W = \lambda_{begin}\lambda_{end}/c_{diff}(\lambda_{begin}+\lambda_{end}) = 227$ generations. This low number is unsurprising given the evident high similarity between these genes ($> 80\%$). Although it is tempting to link the two approaches, a more rigorous and thorough analysis is necessary to understand and quantify diversification occurring at different scales of the gene family.

5.9.2 Outlook

Even without invoking selection on identity per se, we were able to capture the distribution of pairwise genetic identity in the VSG antigen archive by a simple difference in mutation and gene conversion rates, further evidence of the power of neutral molecular evolution. Throughout our analysis, one important assumption was that the probabilities per nucleotide per generation of conversion and mutation are constant. An interesting question that arises, however is whether these evolutionary rates are subject to selection and to what extent they are adaptive. Mutation rates must depend on an evolutionary compromise between the need to create diversity - a basis for adaptive evolution as in the case of surface protein genes,- and the requirement to preserve core or essential cellular functions, as in the case of genes encoding nuclear proteins. Are there any environmental cues within hosts that trigger higher mutation or alternatively higher conversion rates in the VSG genes of trypanosomes? Are there any structural genomic constraints that limit the generation of diversity, and if yes, what are their features? Understanding these aspects of VSG archive diversification could prove crucial in the design of control strategies, for example drugs that interfere with the capacity of the pathogen to mutate or diversify.

The current model ignores the functionality of regions of DNA. An interesting avenue for future studies is the consideration of diversity at the protein sequence level, as opposed to the nucleotide level. Simulation studies and experimental approaches that can characterise and model the generation of diversity via mutation and gene conversion at the amino-acid level would have the advantage of being more realistic, because they could track those changes that are relevant for protein expression, and could offer concrete biological insight into the mechanisms and particular genes involved. As

usual, the appropriate modelling approach and level of biological detail will depend on the particular question being asked about the parasite.

Finally, the incorporation of other types of mutational processes into the modelling, such as identity-driven gene conversion, could prove crucial in explaining the maintenance of the characteristic identity distribution of VSG genes, skewed towards the low identity spectrum, even in the face of potentially high conversion rates. The appropriate mathematical quantification of these processes and their features could in part also resolve the apparent discrepancy between the HMM conversion diversity parameter μ and the global genetic diversity observed empirically in the VSG archive.

In general, despite advances driven by molecular biology and genomics, there is a need to gain a deeper understanding of key mechanisms that may facilitate generation of diversity across biological systems and scales. The characterisation of surface protein families or other multigene families on the basis of their capacity to generate variation is important, and necessitates models for explaining the role of the genetic processes involved. As illustrated in this chapter, computer models and mathematical models may be implemented to explore, visualise and estimate the dynamic diversification capacity of gene families.

Chapter 6

Discussion

The models presented in this thesis have elucidated the importance of antigenic variation in the life-cycle of the African trypanosome, starting from the within-host level, continuing with the between-host level, and finally exploring its roots in the parasite genome. With increasing availability of genomic data, the frontier for systems biology is the integration of parasite genomic architecture into models that bridge between these different levels of organization. As illustrated in this thesis, there are many mathematical approaches and tools such as population dynamics applying ODE and PDE models, stochastic processes (Markov, Poisson processes), population genetics and Bayesian statistics that can be used to address the interesting questions at the interface between biology, parasitology and genetics.

6.1 Structure of the antigenic archive

The structure of the parasite antigenic archive has been a pervasive topic throughout the thesis. While in Chapters 2 and 3 this structure was reflected in the switch matrix between parasite antigenic variants, and its role in determining infection profiles was investigated, in the last two chapters, the structure of the antigenic archive emerged in the genetic identity between VSG genes. By presenting a gravity model (Sen & Smith, 1995) as an interface between the two approaches, I have shown that it is possible to mechanistically formulate and parametrize explicit switching models. The switch rates, giving rise to the switch matrix, can be directly expressed in terms of characteristics of the genes involved in switching and their genetic distance. This allows then

genetic processes at the level of the parasite genome, such as gene conversion and point mutation, discussed in Chapters 4 and 5, to be linked to population level processes within and between hosts. Although the between-host dynamics of trypanosomes was not addressed explicitly in this thesis, appearing only in the nested within-host parasite fitness and R_0 formulation in Chapter 3, a more detailed description of epidemiological processes based on the classical *SIR* framework is possible. Such integrated model, like the one proposed by McKenzie & Bossert (2005) for the malaria parasite, would help us to understand how the dynamics of trypanosomes within an individual host relate to parasite dynamics in a population of hosts, what is the role played by the vector population, and how these hierarchical relationships may influence the outcome of different types of interventions, targeting not solely antigenic variation.

Clearly, the within-host and between-host dynamics of trypanosomes are linked in multiple ways, antigenic variation being an important one. The magnitude structure of switch rates, organized hierarchically in blocks, where switching within blocks is faster than between blocks, was shown to be a crucial factor influencing the outcome of an infection. In Chapter 2, we observed how the size of a block, indicating the relatedness between variants in terms of switching pathways, affects the peak parasite load and the duration of a single block wave, not only qualitatively, in line with previous literature on antigenic diversity thresholds (Antia *et al.*, 1996; Nowak *et al.*, 1990; Sasaki, 1994), but also quantitatively. Our nested modelling in Chapter 3 showed that there exists an intermediate optimal block size driving the infection dynamics at a balance between transmission and virulence.

Similarly, the number of blocks, originating from the size of the antigenic archive, was shown to affect the number of peaks and overall infection duration. Theoretical studies in adaptive dynamics have shown that the generation of diversity in both hosts and parasites is dependent on the shape of the trade-off relationships, but is more likely in long-lived hosts and chronic disease with long-infectious periods (Best *et al.*, 2010). We showed that the effectiveness of a particular number of blocks depends on host characteristics such as parasite tolerance, natural lifespan and immune-competence, suggesting a continuous arms' race in the coevolution of the parasite and its hosts, where longer-lived hosts promote archive expansion and diversification, thus increasing parasite infectivity.

Concerning the between-block switch rate, we showed that this archive characteristic is important in ensuring infection maintenance, especially in very resistant hosts that limit the reproduction potential of the pathogen, and in hosts with a small carrying capacity, where parasite stochastic extinction is more probable. Interestingly, the characteristic delay it takes the host to mount a specific immune response against any parasite variant is found to affect directly the optimal switch rate between antigenic blocks. This result points in the same direction as studies that link the evolutionarily stable mutation rate between two phenotypes, to the rate of periodic environmental changes (Ishi *et al.*, 1989).

The results of Chapters 2 and 3, taken together, indicate that such effects at the within-host level have implications for the between-host level, where parasite transmission depends on the peak parasite load, infection duration, and host survival. In many cases we find a continuum of antigenic variation strategies in combination with other within-host processes (e.g. density-dependent differentiation) that confer the same fitness to the parasite. This result is similar to fitness continuums shown to arise for RNA viruses as a result of trade-offs between immunogenicity and antigenic variability (Haydon & Woolhouse, 1998). We thus confirm the idea that a parasite like the African trypanosome has an enormous potential to diversify and adapt to particular hosts, not only by changing the structure of its antigenic archive, but also by changing other life-history traits, for example, the conversion from replicative to non-replicative life-stages (Alizon & van Baalen, 2008b; Sasaki & Iwasa, 1991), typical of most vector-borne pathogens.

Although we have illustrated how selective pressures on the structure of the antigenic switch matrix might act top-down, from population level processes and parasite transmission requirements in the field, we have not presented a generative/adaptive framework whereby this structure might evolve: for example, a model of evolutionary dynamics of the switch matrix, where the switch rates would tend to organize into blocks of different magnitudes, as formulated by Frank (1999). This might be an interesting avenue for the future, when new genetic findings regarding switch pathways can be incorporated, and the associated genetic processes driving this evolution, explicitly quantified.

6.2 Within-host parasite control

In this thesis, we have investigated several aspects of parasite control within the host over the course of a trypanosome infection, that we discuss in the following. For antigenically varying pathogens, like the African trypanosome, the major form of within-host parasite control comes from antigen-specific host immune responses, mediated by B-cells and antibodies. Our modelling showed that the interplay of this variant-specific control with parasite density-dependent self-regulation processes, such as differentiation into stumpy cells, can have important consequences for the outcome of an infection (Savill & Seed, 2004; Seed & Wenck, 2003). As a result of a large antigenic diversity, expressed at the same time within the host, (large block size in the switch matrix), immune responses may receive sub-optimal stimulation due to faster parasite density-dependent limitation of each variant. Thus, the balance between specific and general control of the pathogen relies heavily on the amount of antigenic diversity present, hence on the structure of the antigenic archive. This balance, besides simply total parasite numbers, impacts the ratio of the slender and stumpy cells over infection, which has a direct bearing on the ability of the parasite to be transmitted to the vector and possibly also on host survival. Stumpy cell production may actually benefit the individual host, but by virtue of its role in parasite infectivity, may harm the host population.

The picture gets complicated further when general control can come from the host as well, such as when host immune responses are cross-reactive, or when there is a component of the immune response that equally attacks all variants (innate immunity). We have seen that in the presence of cross-reactivity, specific-immune response compete for stimulation both amongst each other and with the density-dependent differentiation process. This can result in the parasite being cleared more easily, but paradoxically, cross-reactivity can also facilitate variant persistence and prolong infection within a host, albeit at a reduced parasite load. In analogy to the interplay between parasite density-dependent differentiation and specific immunity, innate immune responses may prevent the host to build up sufficient levels of acquired immune responses against some variants, thus permitting later re-infection. These sorts of factors may have important implications for host immune history and transmission processes. Unfortunately, due to absence of cross-reactivity data for trypanosome variants, the

model only allows for exploration of hypothetical outcomes. Thus more experimental data are needed to properly include and quantify the effects of cross-reactive and innate immune responses in within-host dynamics and beyond.

Our results show the dominance of transmission stages during chronic infection is mediated primarily by the parasite-intrinsic ratio of slender growth rate and stumpy cell mortality rate, and further reinforced by the differential killing rates by the host immune responses. This control has implications for the allocation of parasite resources between antigenic variation (only slender cells can switch) and transmission (only stumpy cells can infect the vector). By restricting the number of proliferative slender forms, trypanosomes can reduce their VSG switch frequency over infection, thus optimizing parasite transmission probability without substantially harming host survival, especially if the virulence of stumpy cells is lower than that of slender cells. Such pathogen strategies reflect the trade-off between longevity (local reproduction) and fecundity (global reproduction), thought to be an important factor driving pathogen diversity (Frank, 1996).

One of the fundamental questions in many infectious diseases is what determines a pathogen's ability to cause an acute or chronic infection (Alizon & van Baalen, 2008a). In particular, there are two human-infective trypanosome strains *T.b. rhodesiense* and *T.b. gambiense* that seem to be associated to these two disease forms. In Chapter 3 we have investigated how the structure of the antigenic archive, namely the number of variant blocks available to the parasite for expression, can account for such distinction in infection profiles, crucially combined with differences in transmission requirements, for example a higher stumpy cell number needed to infect the vector. There exists an intermediate transmission threshold for which the severe acute and chronic infection profile provide the same infection fitness to the parasite. This raises interesting questions on the role of parasite-vector interaction (Ewald, 1983) that can account for infection differences in the field, as opposed to or in addition to varying efficacies of host immune responses. Clearly, unravelling the epidemiological significance of the observed trypanosome strain divergence (Hutchison *et al.*, 2007) remains key for understanding the evolution of this parasite.

6.3 VSG archive diversification

As seen in Chapters 4 and 5, genomic data provides unique opportunities to capture mechanisms of generation, maintenance and access of antigenic diversity in pathogens. By modelling VSG genetic diversity, both locally, at the level of high-identity subfamilies, and globally, at the level of the entire archive, we have quantified two important dynamic forces in the evolutionary dynamics of these genes: point mutation and gene conversion.

The relative dominance of mutation versus gene conversion in N-terminal VSG domains reveals that trypanosome surface antigens in principle tend to diverge, most probably as a defense mechanism against strong and variable host immune responses. It would be interesting to compare estimates of mutation rates across the C- and N-domain. Most probably, the mutation rate will be lower in the C-domain because this gene region has a housekeeping function, coding for the part of the VSG surface protein that is not exposed to the outside environment, unlike the N-domain which codes for the exposed part and thus interacts much more with the unpredictable environment of the host. The advantage of having a mutation rate that is variable across the genome has to do with the evolutionary flexibility it promotes in the face of environmental change (see Moxon *et al.* (1994) for a discussion in relation to bacterial pathogens), while minimizing deleterious effects on fitness.

Importantly, the organization of the VSG archive in subfamilies means that in addition to point mutation, gene conversion between subfamilies and the rest of the archive can act as another diversifying process on the N-domain. However, global gene conversion across the archive serves to preserve and create dynamically a small number of high-identity gene pairs, needed for mosaic formation, so crucial in the chronic stages of infection (Marcello & Barry, 2007b).

Understanding VSG archive diversification requires however more detailed analyses of the types of mutations and their significance for antigenic variation. In particular, following the evolution of expressed VSG sequences in parallel to VSG switching over the course of an infection could provide a deeper insight into the short-term and long-term effects of mutational processes. An interesting comparison would be to examine mutation rates *in vitro* and *in vivo* and identify possible host factors that could enhance or block this process. It seems there are two strategies available to the parasite during

a chronic infection: either new mosaic gene formation from multiple gene conversion events involving silent VSG copies, or expression of a mutated but functional VSG gene that is sufficiently non-cross-reactive with previous variants. Determining which of the strategies is easier to attain by trypanosomes will depend on possible constraints and limitations of either process, and may highlight parallels between trypanosomes and other pathogens (e.g. viruses) that rely on mutation for immune escape. By the complexity and wide range of trypanosome hosts in the field, we expect one mechanism alone to be insufficient to provide the parasite all the flexibility needed for adaptation and infection persistence. It is likely that a combination of mutational mechanisms may be adopted instead.

The high VSG N-terminal domain mutation rates we estimated in our study suggest that perhaps the large number of pseudogenes observed today in the trypanosome antigenic archive may have resulted from such high mutational propensity of the pathogen combined with a high gene duplication rate. Indeed, the process of archive expansion via gene duplication must be very important. Besides prolonging infection in any host, a large antigenic archive serves to increase the antigenic variation potential of the parasite especially in already immune hosts. Although we did not address this process in any of our models, it remains a very interesting topic for further modelling and investigation in the future. For example, one could address VSG gene family size evolution through birth-death-innovation models (Karev *et al.*, 2003). The ultimate explanation for the large size and structure of the VSG antigenic archive today must come from the integration of detailed knowledge about selective and non-selective forces acting on the parasite genome. Processes like point mutation, gene conversion, gene duplication, deletion etc. must all be taken into account. Conceivably, increasing the number of genes may be a strategy to compensate for the high number of deleterious mutations in existing genes. More data are needed to investigate these questions, and it is likely that optima governed by bottom-up genomic processes, similar to those governed by top-down population processes, are found to emerge.

6.4 Data and experiments

Our primary aim was to describe general patterns of trypanosome infections and we did not fit our theoretical within-host model to any concrete infection data. Thus, although we remain confident that this model captures well the most important characteristics of a chronic infection profile and can provide qualitative and quantitative insight into the interplay of different infection processes, we suggest it should be improved further before it can be applied to real-infection scenarios.

In order to advance our understanding of this infectious disease and increase the predictive value and practical usability of our model, we must capture in greater detail host immune response kinetics. More data on different immune mechanisms acting in the host, whether antigen-specific or general, are needed to determine accurately the functional forms of these types of parasite control and their time-dependence. It is important to dissect which model parameters vary between hosts of different types and different host species, and how allometry might be involved to simplify some of these variations. The ultimate aim of our modelling approaches must be to improve the control of trypanosomes in humans, both at the level of the individual and at the level of the population and envision sustainable long-term solutions. Thus, more comprehensive, longitudinal data are needed, that monitor prevalence in the field, and human infection data, that monitor real infections, both parasite loads and host immune status and susceptibility. Such data would help to adapt and parametrize the within-host model, but also inform the construction of between-host transmission models. It is important to find ways to deal with noisy infection data, the presence of drug effects, and account for infection-induced immune suppression. With refined statistical and modelling techniques, we should be able to capture both general patterns of an average infection and individual variation. Crucially, cases of host individuals that are better at controlling the infection must be analyzed thoroughly, to understand the causes and nature of this phenomenon, whether it is related to host resistance or tolerance, and which physiological and genetic host mechanisms might be involved, as well as parasite factors. Because the severity of sleeping sickness is associated with the parasite crossing the blood-brain barrier, it would be interesting to investigate the parasite strains circulating in the host blood immediately prior this transition and check any genetic signatures of virulence that might be shared across different hosts.

Concerning the parasite, many questions remain unanswered. Understanding the differences between slender and stumpy cells and their interaction, in terms of utilization of host resources, stimulation of host immune response, virulence, etc. should be a primary goal. Some studies have moved in this direction (e.g. (McLintock *et al.*, 1993)). There might be possible control strategies that could use these differences to drive infection profiles towards regimes where the host suffers a lighter burden from the disease, even though the parasite might not be completely cleared. As argued by Mathews (2011) generally for vector-borne pathogens, and shown by our theoretical modelling of trypanosome infections, the density-dependent differentiation process between the proliferative and transmissive parasite forms is crucial. The resulting within-host carrying capacity, K, emerged to be important in many aspects of chronic infection. More data is needed to determine what controls K. Is it a host or parasite factor, or a combination of both? As shown by Reuner *et al.* (1997) and others, differentiation may be triggered by a soluble stumpy induction factor (SIF), which accumulates in the culture medium and which, upon reaching a threshold concentration, leads to formation of the stumpy form. Unfortunately, the chemical identity of this factor has not been elucidated so far, mainly due to the complexity of the culture media. However, follow-up experimental studies could address whether it is possible to induce differentiation at lower parasite loads, *in vitro* and *in vivo*, by increasing the number of SIF receptors on parasite cells, for example, or increase stumpy cell lifespan to slow down antigenic variation within a host.

With regards to antigenic variation, longitudinal data from field animals should prove crucial in improving our understanding of this process. Our models dealt primarily with the structural aspects of antigenic variation. Another important aspect has to do with the exact variants that arise, how similar are the sequences across hosts, can we infer the antigenic blocks from infection data, can the connectivity network be built and critical variants within blocks identified? Recently, there has been some progress in this direction for the malaria parasite (Recker *et al.*, 2011). Similar advances for trypanosomes may not be far. If there are recurrent patterns of variant appearance and if they correlate somehow with the development of disease, we might be able to exploit these regularities in vaccine research.

Going down to the genetic level, an interesting avenue for future research is the parametrization of the gravity model and its components. How do genetic features

of two VSG genes determine quantitatively the switch rate between them? How does mosaic gene formation depend on VSG sequence identity and can we quantify the deterministic and stochastic component of this process? Given a particular pseudogene, for example, what is the chance for its expression via recombination-driven mosaic formation? One assumption behind the fitness arguments with regards to the antigenic archive is the heritability of the structure of the archive. But how is the structure really inherited? Are the block sizes roughly conserved across parallel and serial infections, and if yes, what is the genetic signature of such conservation? Alternatively, if the antigenic structure is volatile, can we deal with the mechanisms that generate this structure dynamically? Controlling the number of repeats in VSG flanking regions, inducing mutations, blocking recombination pathways are all possibilities to be explored in this direction.

Finally, quantifying the role of mutation and recombination mechanisms on larger-scale processes such as trypanosome speciation and parasite strain divergence is crucial for our understanding of trypanosome epidemiology and evolution in the field. Recombination has been shown to play a major role in parasite adaptation, virulence and drug resistance in bacteria (Fraser *et al.*, 2007; Hanage *et al.*, 2009), but appears to be a general feature of many other pathogens (Awadalla, 2003). More research is needed to determine the functional significance of recombination for trypanosomes. So far, we have only analyzed diversity at the level of one VSG archive, and used pairwise alignment data derived from one parasite strain. The fresh evidence that is emerging on sexual reproduction mechanisms of parasites (Heitman, 2006), including trypanosomes (Peacock *et al.*, 2011), motivates further investigation of recombination and genetic exchange processes driving the diversification of this parasite at other levels of biological organization.

6.5 Concluding remarks

African trypanosomes are fascinating parasites of enormous complexity. To fully understand them, their diversity, and the disease they cause, it is important to study their dynamics at different spatial and temporal scales, and subsequently unify them. In this thesis we have considered scales ranging from the single variant and single block

dynamics to multiple-block dynamics and chronic infection, parasite fitness within- and across hosts, transmission within and between communities, genetic diversification within VSG subfamilies, across the entire antigen archive and divergence between repertoires. However, it is impossible to answer all questions about this parasite in one thesis. Unfortunately (or fortunately) any model is just *one* model, with its scope and limitations, thus always leaving room for future improvement. I hope the models presented in this thesis are found to be useful. How infection processes, ecological feedbacks and genetic mechanisms interact to generate and maintain diversity in hosts and parasites remains an open question.

Appendix A

Mathematical details for the within-host model

A.1 Parasite dynamics with only differentiation

Here, we calculate the steady states when there is only differentiation-mediated parasite control, assuming $a_i = 0$ for all i. Since we consider the dynamics of an antigenic block, where all variants are symmetric, $V(t) + M(t) = \eta v_i(t) + \eta m_i(t)$ for all time, and each variant shares an equal proportion $(1/\eta)$ of the total parasite load. We have

$$\frac{dv_i}{dt} = rv_i\big(1 - \eta\frac{v_i + m_i}{K}\big), \tag{A.1}$$

$$\frac{dm_i}{dt} = rv_i\eta\frac{v_i + m_i}{K} - \delta_M m_i. \tag{A.2}$$

The two steady states are: the trivial one, $v_i^* = 0, m_i^* = 0$ and the nontrivial one, $v_i^* = \frac{\delta_M K}{\eta(\delta_M + r)}, m_i^* = \frac{rK}{\eta(\delta_M + r)}$. At the trivial steady state, the eigenvalues of the corresponding Jacobian matrix of the linearized system are $\lambda_1 = -\delta_M$, $\lambda_2 = r$, hence the disease–free steady state is always unstable. At the non-trivial steady state the eigenvalues of the Jacobian matrix are $\lambda_1 = 1/2(-\delta_M - \sqrt{\delta_M^2 - 4r^2})$, $\lambda_2 = 1/2(-\delta_M + \sqrt{\delta_M^2 - 4r^2})$. Both have negative real part implying the non-trivial steady state, where $V^* = \frac{\delta_M K}{\delta_M + r}$ and $M^* = \frac{rK}{\delta_M + r}$, is asymptotically stable.

In the case where the total parasite population changes independently of the block size η we have:

$$\frac{dV}{dt} = rV\left(1-\frac{V+M}{K}\right), \tag{A.3}$$

$$\frac{dM}{dt} = rV\frac{V+M}{K}-\delta_M M, \tag{A.4}$$

and total parasite number at steady state remains fixed at K, even as the number of variants changes. Thus $\partial V(t)/\partial\eta = \partial M(t)/\partial\eta = 0$, for all time, in particular also at their respective maxima. Denoting by V_{max} and M_{max} the peak total slender and stumpy populations of a block of variants, we have:

$$\frac{\partial V_{max}}{\partial\eta} = \frac{\partial M_{max}}{\partial\eta} = 0. \tag{A.5}$$

On the contrary, for each individual variant, $v_i(t) = V(t)/\eta$ and $m_i(t) = M(t)/\eta$, thus leading to $\partial v_i(t)/\partial\eta < 0$ and $\partial m_i(t)/\partial\eta < 0$ for all time t. In particular, $v_i(t)$ and $m_i(t)$ are linearly decreasing functions of η, also true at the respective peaks $\hat{v}_i$ (slender cells) and $\hat{m}_i$ (stumpy cells),

$$\frac{\partial\hat{v}_i}{\partial\eta} < 0, \quad \frac{\partial\hat{m}_i}{\partial\eta} < 0 \implies \frac{\partial(v_i+m_i)_{max}}{\partial\eta} < 0. \tag{A.6}$$

The ratio of slender-to-stumpy dominance, shifts from favouring slender cells at the beginning of infection, to $\delta_M/r < 1$ at steady state, favouring stumpy non-dividing cells. To summarize, when there is only parasite differentiation, the total coupling between variants due to density-dependence is maximal. This implies the total parasite load is independent of the block size, while the size of individual variant peaks decreases with η.

A.2 Block size and only host control

The dynamics corresponding to the extreme case where there is no differentiation ($K \to \infty$), thus no stumpy cells, and immune variant specific control is strongly coupled to parasite numbers ($\tau = 0, x = 1$) are given by

$$\frac{dv_i}{dt} = rv_i - da_i v_i, \tag{A.7}$$

$$\frac{da_i}{dt} = c\left(\frac{v_i}{C}\right)(1-a_i). \tag{A.8}$$

Dividing the above equations, we get:

$$\frac{dv_i}{da_i} = \frac{C(r-da_i)}{c(1-a_i)}. \tag{A.9}$$

Now, v_i reaches its maximum when $a_i = r/d$. By integrating eq. (A.9) and substituting $a_i = r/d$, we get the maximum value of each variant v_i, which we denote by $\hat{v}_i$:

$$\hat{v}_i = \frac{C}{c}(d-r)\ln\left(1-\frac{r}{d}\right) + \frac{Cr}{c} + v_i(0). \tag{A.10}$$

If initial conditions are such that $v_i(0)$ is independent of the block size, η, then the dynamics of a single variant are entirely decoupled from those of the other variants, and as a result, individual variant subpeaks and duration of infection are not affected by changes in block size: $\frac{\partial \hat{v}_i}{\partial \eta} = 0$, and the peak total parasite load increases linearly with η:

$$\frac{\partial V_{max}}{\partial \eta} = \frac{\partial \sum_i \hat{v}_i}{\partial \eta} = \frac{\partial(\eta \hat{v}_i)}{\partial \eta} = \hat{v}_i + \eta\frac{\partial \hat{v}_i}{\partial \eta} = \hat{v}_i > 0. \tag{A.11}$$

If instead, the block size limits $v_i(0)$, such that the initial parasite burden released from a block of variants is fixed at V_0, with $v_i(0) = V_0/\eta$ for example, then the size of individual variant peaks decreases with block size: $\frac{\partial \hat{v}_i}{\partial \eta} = -\frac{V_0}{\eta^2} < 0$, whereas the peak parasite load still increases with block size:

$$\frac{\partial V_{max}}{\partial \eta} = \frac{\partial(\eta \hat{v}_i)}{\partial \eta} = \hat{v}_i + \eta\frac{\partial \hat{v}_i}{\partial \eta} = \hat{v}_i - \frac{V_0}{\eta} > 0. \tag{A.12}$$

In summary, parasite control mediated by specific immunity alone results in variant dynamics that are decoupled from each other. The peak total parasite load increases with the block size η, while individual variant subpeaks and block wave duration remain constant.

A.3 Between-block cross-reactivity and η_{crit}

Here we calculate how the critical size of an antigenic block might depend on the size of the previous block in the presence of cross-reactivity. This involves relaxing the assumption that all blocks are of the same size in the antigenic switch matrix. Denote by γ the cross-reactive interference between all the variants in one block (η_{old}

variants) and all the variants in the new block (η_{new}). Assuming that ε is sufficiently small, the two blocks are decoupled, thus when the new block is generated, the variants of the previous block are all in the decline phase (tending to 0) and have a negligible contribution in the total parasite load $V+M$. This implies $V+M \approx \eta_{new}(v_i+m_i)$. However, with the immune responses to all previous variants raised to maximum, when the new block is generated, its variants start to grow following the dynamics:

$$\begin{aligned}
\frac{dv_i}{dt} &= rv_i(1-\frac{V+M}{K}) - dv_i(a_i+\eta_{old}\gamma), \\
\frac{dm_i}{dt} &= rv_i\frac{V+M}{K} - \delta_M m_i - \delta m_i(a_i+\eta_{old}\gamma), \\
\frac{da_i}{dt} &= c(1-a_i)\left(\frac{v_i(t-\tau)+m_i(t-\tau)}{C}\right)^x,
\end{aligned} \tag{A.13}$$

experiencing an additional mortality term due to cross-reactive interference $\eta_{old}\gamma$. Notice that because of this additional mortality, the new variant-specific responses need not reach r/d to initiate clearance, but only $r/d-\eta_{old}\gamma$, thus a lower saturation level. Applying again similar quasi-steady state arguments to the dynamics of the new block of variants, we get that the duration of the non-growth phase of each new variant, once having reached the lower peak $K(1-d\eta_{old}/r)/\eta_{new}$, can be approximated by $T_{non-growth} \approx \tau + T_{r/d-\eta_{old}\gamma}$, where τ denotes the immune delay and $T_{r/d-\eta_{old}\gamma}$ the time it takes for the specific immune response to reach the threshold required for the initiation of parasite clearance. We can neglect the influence of specific immunity against the new variants during their growth phase, thus a_i for the new block starts to grow following the equation:

$$\frac{da_i}{dt} \approx c(1-a_i)\left(\frac{K(1-d\eta_{old}/r)}{C\eta_{new}}\right)^x. \tag{A.14}$$

As a result, the time it takes for a_i to reach $r/d-\eta_{old}\gamma$ is given by:

$$T_{r/d-\eta_{old}\gamma} = \left[\frac{K}{C\eta_{new}}(1-\frac{d\gamma\eta_{old}}{r})\right]^{-x}\frac{\ln(1-\frac{r}{d}-\eta_{old}\gamma)}{c}. \tag{A.15}$$

Based on the same heuristic ($T_{non-growth}=2\tau$) argument as in Section 2.4.3, when $T_{r/d-\eta_{old}\gamma}=\tau$, we find the critical size for the new block with η_{new} variants:

$$\eta_{crit}^{new} = \frac{K}{C\eta_{new}}(1-\frac{d\gamma\eta_{old}}{r})\left[\frac{\ln(1-\frac{r}{d}-\eta_{old}\gamma)}{c\tau}\right]^{-1/x} \tag{A.16}$$

such that for $\eta_{new} < \eta_{crit}^{new}$ host immunity can clear rapidly the new block, while for $\eta_{new} \geq \eta_{crit}^{new}$, density-dependent differentiation controls the dynamics, prolonging the duration of the non-growth phase and delaying parasite decline.

A.4 Immune suppression within a block

We let $A(z)$ denote the degree of impairment to immune growth as a function of antigenic diversity z, and be determined by

$$A(z) = \frac{\phi e^{\alpha z}}{\phi + e^{\alpha z} - 1}, \tag{A.17}$$

with $A_0 = A(0) = 1$, where α is the strength of immune impairment generated by each variant, and ϕ the maximum level of impairment caused by infection. We choose a saturating function to represent the idea that as the number of variants tends to be very large, their deteriorating effect on the immune system of the host must tend to a constant. We can analytically study the simple case of immune suppression in the presence of only host control (Appendix section A.2). For immune suppression within a block, i.e. $A(\eta)$, the equation for the specific immune responses becomes:

$$\frac{da_i}{dt} = \frac{c}{A(\eta)} \left(\frac{v_i(t-\tau) + m_i(t-\tau)}{C} \right)^x (1 - a_i). \tag{A.18}$$

In the only-host-immunity limit ($K \to \infty$), and assuming $\tau = 0$, following the analysis of Appendix A.2, we find the peak variant load is given by:

$$\hat{v_i} = \frac{A(\eta)C}{c} \left[(d-r) \ln \left(1 - \frac{r}{d} \right) + r \right] + v_i(0), \tag{A.19}$$

which, if $v_i(0)$ is independent of η, implies:

$$\frac{\partial \hat{v_i}}{\partial \eta} = \frac{\partial A}{\partial \eta} \frac{C[(d-r)\ln(1-\frac{r}{d}) + r]}{c} = \frac{Ce^{\alpha\eta}\alpha\phi(-1+\phi)[(d-r)\ln(1-\frac{r}{d}) + r]}{c(e^{\alpha\eta} - 1 + \phi)^2} \tag{A.20}$$

which is positive for all η, since $\phi > A_0 = 1$. The above equation confirms that individual peaks $\hat{v_i}$ increase with block size in the presence of immune-suppression, and this dependence attains a maximum at a particular value of η, found by solving $\partial^2 \hat{v_i} / \partial \eta^2 = 0$, giving $\eta^* = \frac{1}{\alpha} \ln(\phi - 1)$, such that for $\eta < \eta^*$, the marginal gain of

each variant peak from increasing block size increases, and for $\eta > \eta^*$, this marginal gain decreases until it approaches a constant. This phenomenon emerges as a result of the fact that as $\eta \to \infty$, $A(\eta) \to \phi$, i.e. tends to a constant. In this limit, we get back the case of an inexhaustible immune response, only this time, with a reduced immune response growth rate c/ϕ. In the presence of within-block immune suppression, it is most advantageous for the parasite to increase antigenic diversity (block size) while $\eta < \eta^*$, because linear increases in η produce more than linear increases in V_{max}.

However, as differentiation-mediated control is introduced ($K \ll \infty$), when the block size increases, the negative feedback exerted by density-dependent regulation predominates over the positive feedback induced by immunosuppression on individual variant peaks and there is a gradual decrease of the marginal gain in peak parasite load from higher η. On the other hand, if already for η small differentiation dominates parasite control, the effects of immunosuppression on the relationship between $\hat{v}_i$ and η are negligible. Generally $\partial \hat{v}_i / \partial \eta < 0$ persists, despite immunosuppression. This is intuitive, since the sensitivity of infection dynamics to features of the immune response (e.g. immune-suppression) can be expected only if the immune response is playing a substantial role in the dynamics.

Appendix B

Genetic mechanisms of antigenic variation

B.1 The gravity model

We borrow the notion of a gravity model from Sen & Smith (1995), related to Newton's gravitational law, and widely used in economic models, and propose it to describe switch rates between two antigen genes. Whatever the genetic mechanisms involved, the stochastic event of switching from expression of one gene to expression of another gene must depend on at least three types of factors: 1) properties of the gene that is being switched away; 2) properties of the gene that is being activated; and 3) mutual relatedness properties of the two genes. Then the rate of gene j switching to gene i can be seen as a function of their individual contributions. One can define the pairwise switch rate as

$$s_{ji} = f\left(R_j, A_i, \phi(d_{ji})\right), \tag{B.1}$$

where R_j ("repulsiveness" of the current gene) describes the intrinsic propensity of a gene to switch away, A_i ("attractiveness" of the new gene) describes the intrinsic probability that a gene is activated, and $\phi(d_{ji})$ is a function of the mutual relatedness between the genes. Inevitably, each of these factors will depend on adaptive molecular

mechanisms acting at the level of the parasite genome.

Repulsiveness of a gene The intrinsic probability of a gene to switch away may depend on a series of genetic factors such as chromosomal location, transcription instability, whether it is an intact gene or a pseudogene, etc. For *P. falciparum* it has been shown for example that *var* genes located in central chromosomal regions had very stable expression patterns, but that subtelomerically located *var* genes easily switched away to alternative *var* loci (Frank *et al.*, 2007). Similarly, chromatin structure may hold some clues about *var* gene silencing in *Plasmodium* (Kyes *et al.*, 2007), or it may be that the expression of a functional gene elicits a feedback signal that reduces switching.

Attractiveness of a gene The intrinsic probability of a gene to be activated may also depend on factors such as the gene locus. In *T.brucei*, there is evidence that VSG genes physically residing in telomeric locations have a higher chance to be activated (Robinson *et al.*, 1999; Van der Werf *et al.*, 1990). Additionally, the number of DNA repeats in gene flanking regions has been suggested to play a role in gene activation (McCulloch & Barry, 1999), with more repeats favoring expression, such as for intact array genes of *T.brucei*. Another factor that may contribute to the attractiveness of a particular gene may be its spatial distance from the expression site: the further away, the lower its chances to be activated (e.g. pseudeogenes in the bacterium *Anaplasma phagocytophilum* (Foley *et al.*, 2009)).

Mutual relatedness While the intrinsic propensities of genes to switch away or to be activated depend on individual gene features, the third component of the gravity model relates to mutual properties of the gene pair. Features such as genetic similarity between the two gene sequences, or their physical distance on parasite chromosomes may play a role in determining this third component. In *T. brucei*, VSG pseudogenes can only be expressed via gene conversion with high-identity functional partners. It has been shown that the genetic distance between the currently expressed gene and the new mosaic VSG strongly affects their pairwise probability of switching, with larger distances acting to reduce switching (Marcello & Barry, 2007a). Plausibly, in the case of *in situ* switches, genes located together in the parasite genome can be expected to display higher pairwise switching. Mechanisms of recombination between gene family members have been reported also for *P. falciparum* (Deitsch *et al.*, 1999) and *Borrelia hermsii* (Donelson, 1995), suggesting that mutual features between the

currently expressed gene and the activated gene must be accounted for in the switch rate.

In summary, for the parametrization of the gravity model, it will be essential to identify and quantify features of genes individually, and their higher-order groupings in the antigenic archive. The current empirical findings on trypanosome antigenic variation motivate us to propose a particular form of the switch matrix, combining both the hypothesized block structure of the genetic architecture of this parasite (Figure 3.5) and the gravity model.

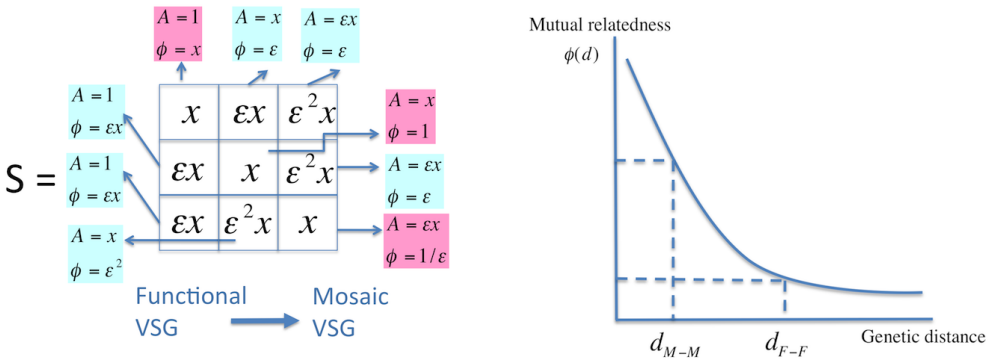
B.1.1 The VSG archive: gravity and blocks

VSG genes are located in telomeric and subtelomeric regions of chromosomes. The primary mechanism in *T.brucei* antigenic variation is recombinational switching, by which silent VSGs from the archive are moved into the expression site (Morrison *et al.*, 2009). There are many such expression sites in the genome of this parasite. Two main types of switching seem to occur: hierarchical switching of intact genes ($\sim 10\%$ of the archive), mediated by the 70bp repeats in their flanking regions, and homology-based switching for pseudogenes ($\sim 90\%$ of the archive), when expressed as mosaics. We propose a gravity model to represent these processes, which assumes f is given by the product between the three independent variables,

$$s_{ji} = R_j A_i \phi(d_{ji}).$$

The probability of switching away in *T.brucei* has been estimated to be in the range $10^{-3} - 10^{-2}$ per cell division and the same for all VSGs (Turner, 1997). For the gravity model, this implies that the "repulsiveness" component R_j should be constant across the archive, $R_j = R$ for all j. Since the jth row sum in the switch matrix equals the overall switch rate of gene j, $\sum_i R A_i \phi(d_{ji}) = R$, which leads to $\sum_i A_i \phi(d_{ji}) = 1$ for all j, imposing a balance between the intrinsic "attractiveness" of genes, and their mutual relatedness ϕ (Figure B.1(a)).

During the course of an infection, first telomeric intact genes on minichromosomes are activated, then subtelomeric array VSGs, and finally subtelomeric pseudogenes, which may have undergone multiple gene conversion events to give rise to mosaics (Morrison *et al.*, 2009). This implies an intrinsic variability in the attractiveness of



(a) The switch matrix

(b) The mutual relatedness function of the gravity model

Figure B.1: a) Schematic representation of the switch matrix. Between-block/within-block average switch ratio is given by $\varepsilon \ll 1$. Each row of the matrix has to sum up to the total switch rate per unit of time, given by $\sigma = 0.01/r\ln(2)$, 0.01 being the probability of switch per parasite division. A power-law function is assumed to govern the switch rate between distant blocks. b) A hypothetical representation of the dependence of pairwise propensity to switch on genetic distance between VSGs. The genetic distance between two mosaic genes, d_{M-M}, is smaller than the genetic distance between two functional genes d_{F-F}, thus the mutual relatedness component ϕ_d is higher in the switch rate between mosaic VSG genes.

genes, decreasing from high to medium to low across these three broad VSG groupings. The molecular mechanisms mediating this hierarchy are related both to VSG locations within the genome, and the relative number and homology of repeats in the flanking regions for intact VSG genes, and a potential hierarchy in the level of dysfunction among pseudogenes.

Given that $\sum_i A_i \phi(d_{ji}) = 1$ for all j, to compensate for the decreasing A_i , the ϕ component across these groups must increase. If $\phi(d_{ji})$ is a decreasing function of pairwise genetic distance, this would require that the average genetic distance between functional VSGs be higher than the average genetic distance between mosaics, a hypothesis thus far supported by empirical findings (Lucio Marcello, PhD thesis).

The molecular mechanisms by which ϕ operates are primarily gene conversion and recombination between closely related genes. Since these genetic processes are favoured by high identity between gene sequences, ϕ plays a negligible role in switch rates of intact genes early in infection, but plays an increasingly major role in switch rates of mosaics at the later stages of infection (Figure B.1(b)).

The general gravity model (Eq.B.1) can be used to generate a switch matrix S with characteristic structure that reflects the genetic architecture of the pathogen (Figure 3.5). For African trypanosomes, the switch matrix likely consists of blocks of genes, highly intra-connected, but with decreasing inter-connections. This block structure generates antigenic variation dynamics that closely resembles real trypanosome infections (Figure 2.2), and represents for this parasite a life-history strategy open to modulation by natural selection.

Appendix C

Details on the Hidden Markov Model

C.1 Mismatch data description

Here, for completeness, we present some characteristics of the mismatches found in our VSG subfamily dataset, including the number of insertions/deletions (Fig.C.1) and types of nucleotide substitutions corresponding to each gene pair (Fig.C.2, Table C.1).

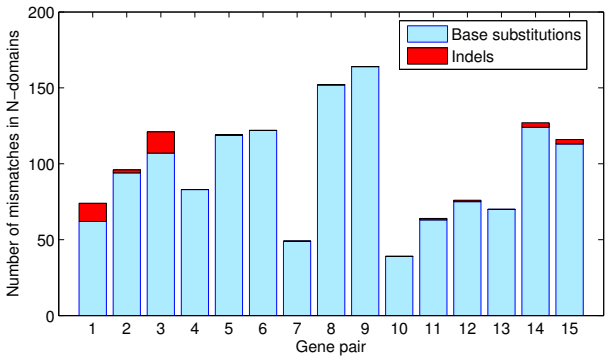


Figure C.1: Nucleotide substitutions exceed the number of indels in our dataset. The low number of indels implies that the diversification rates we infer correspond rigorously to nucleotide substitutions.

Table C.1: The types of mutations found in our VSG dataset. The pairs are listed in the order 1-2,1-3, and 2-3 for each triplet, starting from triplet 1. ∑S and ∑M refer to the sum of nucleotide substitutions and sum of total mismatches on each alignment respectively. *L* denotes the length of the N-domain for each triplet.

Pair	AC	AG	AT	CG	CT	TG	∑S	Indels	∑M	*L*	%Id.
1	18	22	0	11	9	2	62	12	74	1092	93.22
2	23	32	3	16	14	6	94	2	96	1092	91.21
3	24	40	3	20	11	9	107	14	121	1092	88.92
4	28	28	4	12	7	4	83	0	83	1026	91.91
5	30	45	10	18	11	5	119	0	119	1026	88.40
6	25	44	8	24	13	8	122	0	122	1026	88.11
7	16	16	3	4	5	5	49	0	49	1035	95.27
8	32	57	6	24	18	15	152	0	152	1035	85.31
9	37	58	9	25	19	16	164	0	164	1035	84.15
10	11	10	1	5	9	3	39	0	39	1032	96.22
11	16	15	2	6	19	5	63	1	64	1032	93.80
12	21	18	2	9	18	7	75	1	76	1032	92.64
13	18	23	2	6	16	5	70	0	70	1086	93.55
14	21	52	5	18	20	8	124	3	127	1086	88.31
15	19	51	5	13	19	6	113	3	116	1086	89.32

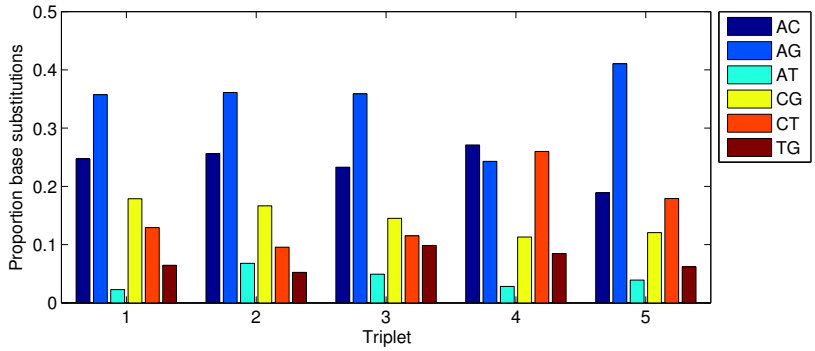


Figure C.2: Nucleotide substitutions in detail. Although we have analyzed only the patterns of pairwise identity between sequences, future studies might address more specific genetic features underlying the same data, for example the bias in certain types of point mutation, or the nucleotide landscape within estimated conversion segments.

C.2 Parameter estimation procedures

C.2.1 The Metropolis-Hastings (MH) Algorithm

The fundamental formula in Bayesian inference methods is Bayes' theorem (Bayes & Price, 1763):

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}, \tag{C.1}$$

where D denotes the observed data and θ denotes model parameters. $P(\theta)$ is the prior distribution on the parameter values, constructed by previous knowledge about the process, and $P(D|\theta)$ is the likelihood of observing the data, given a particular combination of parameter values. $P(\theta|D)$ is the posterior distribution one wants to determine after having observed D, and is the object of all Bayesian inference. Any features of the posterior distribution, such as quantiles, moments, etc, can be expressed in terms of posterior expectations of functions of θ.

When the posterior distributions are hard or impossible to get analytically, MCMC sampling is used to get a representative sample from the posterior. Monte Carlo integration evaluates the expectation of a function $E[f(X)]$ by drawing samples $\{X_t, t =$

$1,..,n\}$ from $\pi(.)$ and then approximating

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t), \tag{C.2}$$

whereby the population mean of $f(X)$ is estimated by a sample mean. When the samples $\{X_t\}$ are independent, laws of large numbers ensure that the approximation can be made as accurate as desired by increasing the number of iterations (sample size n). One way of doing this is to construct a Markov Chain having $\pi(.)$ as its stationary distribution. Constructing such a Markov chain can be done for example via the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953). For the Metropolis-Hastings algorithm, at each time t, the next state X_{t+1} is chosen by first sampling a candidate point Y from a proposal distribution $q(.|X_t)$. The candidate point Y is then accepted with probability $\alpha(X_t, Y)$ where

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right). \tag{C.3}$$

If the candidate point is accepted, the next state becomes $X_{t+1} = Y$. If the candidate is rejected, the chain does not move: $X_{t+1} = X_t$. Notice that if the proposal distribution is symmetric, i.e. $q(X|Y) = q(Y|X)$ for all X and Y, the acceptance probability reduces to:

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)}{\pi(X)}\right), \tag{C.4}$$

where π is the target stationary distribution. An example of a symmetric proposal distribution, which depends on the current point X_t is the multivariate normal with mean X and fixed covariance matrix $N(X, \sigma^2)$. In this case, it is necessary to choose the scale parameter σ carefully, because if σ is too large, a large number of iterations will be rejected and the algorithm will be very inefficient as a consequence. Similarly, if σ is too small, the random walk will accept nearly all proposed moves, but will move around the parameter space very slowly, again leading to inefficiency. It is suggested that σ be tuned so as to achieve an overall acceptance rate in the range $[0.15, 0.5]$ (Gilks *et al.*, 1996). Notice that when the priors are uniform, $\pi(X)$ reduces to the likelihood $P(D|\theta)$. This is our case. As a result, calculating $\pi(X)$ and $\pi(Y)$, at each iteration, can be done using the hidden Markov model function *hmmdecode* in MATLAB (R2010b, Mathworks, Natick, MA) that sums over all possible hidden paths: $P(\mathbf{y}) = \sum_S P(\mathbf{y}, S)$, for a given set of HMM parameters.

Usually many parallel chains from different starting values are monitored. Several statistical measures of convergence and tests can indicate whether convergence has been reached. Two of them are the Gelman-Rubin convergence statistic and the auto-correlation measure (Gilks *et al.*, 1996). The idea is that convergence has been reached when the variance within and between parallel Markov chains is approximately the same. Consider m parallel simulations, each with length n and a scalar summary we want to track ψ, indexed as $\psi_{ij}, j = 1,...,n; i = 1,..,m$. We then compute two quantities, the between-sequence variance B and the within-sequence variances W:

$$B = \frac{n}{m-1}\sum_{i=1}^{n}(\bar{\psi}_i - \bar{\psi})^2, \quad \text{where } \bar{\psi}_i = \frac{1}{n}\sum_{j=1}^{n}\psi_{ij}, \quad \bar{\psi} = \frac{1}{m}\sum_{i=1}^{m}\bar{\psi}_i \tag{C.5}$$

$$W = \sum_{i=1}^{m} s_i^2, \quad \text{where } s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\psi_{ij} - \bar{\psi}_i)^2. \tag{C.6}$$

From the two variance components, we construct two estimates of the variance of ψ in the target distribution. The first one is: $\hat{var}(\psi) = \frac{n-1}{n}W + \frac{1}{n}B$, which is an unbiased estimate of the variance under stationarity, but an overestimate under the more realistic assumption that the starting points are overdispersed. At the same time, for finite n, the within-sequence variance, W, should underestimate the variance of ψ because the individual sequences have not had time to explore the full posterior and, as a result, have less variability. In the limit as $n \to \infty$, both $\hat{var}(\psi)$ and W approach $var(\psi)$ but from different directions. By monitoring the 'estimated potential scale reduction', denoted by the following quantity:

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{var}(\psi)}{W}}, \tag{C.7}$$

we can check the convergence of the Markov Chain by estimating the factor by which the conservative estimate of the distribution of ψ might be reduced. In practice, the simulations run until the values of $\hat{R}$ (the dimension equals the number of parameters) are all less than 1.1 or 1.2. After the burn-in period has been reached, the Markov chains are let to run for a substantial number of other iterations to explore the full posterior, after which the posterior distributions of the parameters and other statistics of interest can be computed.

C.2.2 Alignment decoding

After the most likely transition matrix, T, and probability distributions, Φ_k, have been estimated through MCMC techniques, one can use them to "decode" the observation sequence, i.e. obtain the most likely path of associated hidden states: $S^* = \arg\max_S P(\mathbf{y}, S)$. This path can be found recursively. The Viterbi algorithm that performs this task is based on dynamic programming principles (Forney, 1973). In MATLAB, the function performing this is called *hmmviterbi* and it takes as input a 2 x 2 transition matrix ($TRANS$) and the 2 vectors specifying the emission probabilities ($EMIS$). We use the means of the posterior distributions as input parameters for determining $TRANS$ and $EMIS$, when looking for the most likely S, for each $\mathbf{y}$, corresponding to each alignment.

C.2.3 Pair-correlation function calculation

The pair correlation function (see (Illian *et al.*, 2008) for a general description) is a second-order characteristic of the mismatch point pattern on each alignment, which can capture subtle features of clustering. It measures the density of particles (mismatches, in our case) found at a certain distance from each other. It is usually defined for point patterns on continuous space, but on the basis of its definition, it can be easily extended to discrete-space point patterns, as in our case of M mismatches occurring in an alignment of length L. The formula in one dimension is given by:

$$g(r) = \frac{M(M-1)}{L^2} \sum_{i=1}^{M} \sum_{j \neq i}^{M} \frac{e_h(r - d_{ij})}{L - d_{ij}}, \tag{C.8}$$

where e_h is the Epanecnikov kernel, defined as $e_h(x) = \frac{3}{4h}(1 - \frac{x^2}{h^2})I_{|x| \leq h}$, with h being the bandwidth and I the indicator function. In our case $h = 2$, $r \in \mathbb{Z}$, and $g(r)$ counts how many pairs of mismatches in the alignment lie within discrete distance r from each other. When $g(r)$ is greater than 1, this indicates clustering in the spatial point pattern at that scale, whereas if $g(r) \approx 1$, the pattern is random and the point occurrence follows the Poisson process. The pair correlation functions of our data are plotted in Figure 4.15(a).

C.3 Algorithm validation

In order to show the performance of the Bayesian estimation method, we generated data by simulating the process as in Appendix 4.2.1 and then evaluated the ability of the algorithm in recovering these parameters. The following analysis presents the results of the estimation process. We show validation results for only two models: the simplest model (Model 1) with only 4 parameters, and the most complicated model (Model 4) with 3+N parameters, and one set of parameter values. We look at three aspects of algorithm performance: 1) the evaluation of Bayesian posterior means as estimators of the true values of genetic parameters; 2) the precision of the predicted Bayesian confidence interval in containing the true values of the parameters and 3) the accuracy in the matching between "decoded' and true hidden states.

C.3.1 Model 1: Global fit

We simulated 15 alignments of length $L = 1000$, with parameters $(\lambda_{begin}, \lambda_{end}, \mu, m)$ $= (0.01, 0.02, 0.25, 0.03)$. On average each simulated alignment resulted in 130 next-mismatch distances. We estimated the means of the genetic parameters from the posterior distributions obtained via the Metropolis-Hastings algorithm, applied 10 times on different sets of 15 simulated alignments. The Metropolis-Hastings algorithm was implemented with 3 parallel Monte Carlo Markov Chains. Uniform priors were assumed. The variance of the normal proposal distribution was set to 0.00025 for every parameter. Convergence was generally reached within the first 5000 iterations. This was ascertained by computing the Gelman-Rubin statistic and ensuring that it had reached a value below 1.1 for all parameters. Afterwards the Markov Chains were let to run for another 10 000 iterations, after which the performance of the estimation procedure was evaluated.

The performance of the algorithm was very good. The average posterior means over just 10 runs were: $(\hat{\lambda}_{begin}, \hat{\lambda}_{end}, \hat{\mu}, \hat{m}) = (0.0094, 0.0204, 0.2513, 0.0312)$, in excellent agreement with the 'true' parameter values. The average normalized deviations of the means of the posteriors from the 'true' values of the parameters are less than 15 % for all parameters. Indeed, the more sequences used in each run of the algorithm,

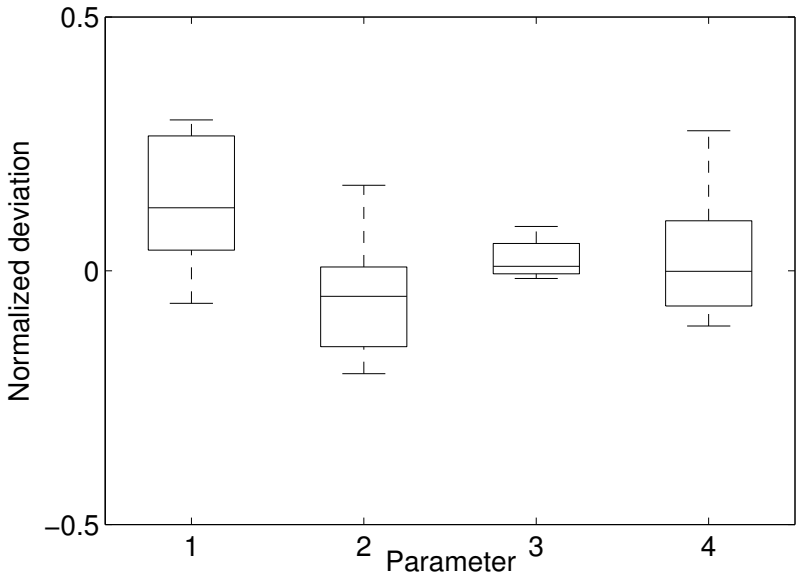


Figure C.3: The error distributions for the inference method applied on 10 sets of artificially constructed data. The deviations of the posterior means (found via the Metropolis-Hastings algorithm) from the 'true' parameter values are divided by the 'true' values (normalized deviations). In the boxplots, the thick horizontal bars show the medians; the box contains the middle half of the data; the whiskers extending the box reach to the most extreme non-outlier (1.5 $\times$ inter-quartile range); outlying points are plotted individually. The order of the parameters is $(\lambda_{begin}, \lambda_{end}, \mu, m)$.

Table C.2: Deviations of Bayesian posterior means from the 'true' values of the model parameters, obtained from the Metropolis-Hastings algorithm. True parameter values: $(\lambda_{begin}, \lambda_{end}, \mu, m) = (0.01, 0.02, 0.25, 0.03)$.

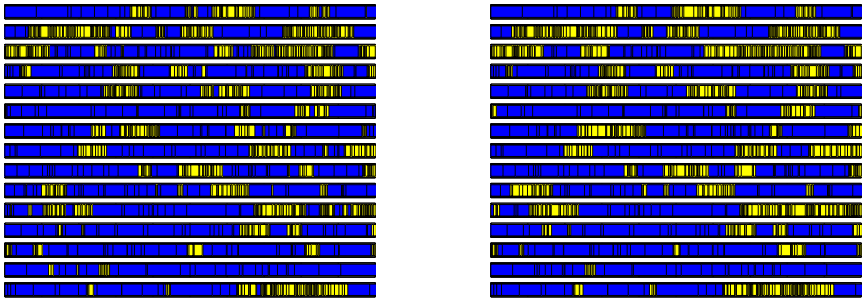
	λ_{begin}	λ_{end}	μ	m
Mean Bias	0.0013	-0.0011	0.0061	0.0006
Root Mean Squared Error	0.0015	0.0019	0.0078	0.0024
Mean Squared Error	3.06×10^{-6}	5.71×10^{-6}	1.14×10^{-4}	1.12×10^{-5}
Norm.Dev.	0.1343	-0.0527	0.0244	0.0187

the better its performance. Similarly, the longer each alignment, the better the performance of the algorithm, because more data are available. In Table C.2 we summarize some results of the estimation procedure. The normalized deviations are shown in Fig. C.3. We see that with as few as 15 alignments and a length of each alignment equal to 1000 (similar to the data we have), the performance is still impressive.

Another aspect of the performance of our parameter estimation procedure can be evaluated by comparing the 'true' states next-mismatch-segments (*within* or *between*) with the inferred states by the decoding algorithm. As can be seen from Figure C.4, the accuracy of the decoding, with parameters extracted via the Metropolis-Hastings algorithm, is very high.

C.3.2 Model 4: Individual ages

Here we generated alignments of different 'ages' by changing the probability of mutation and probability of conversion initiation. In fact, the effective probability of a point mutation per nucleotide will be the baseline value (reference pair) multiplied by the relative divergence time with the given gene pair. Similarly the probability of a conversion event initiation. We assume alignment 1 has age 1, and then infer the following



(a) True states (b) Decoded states

Figure C.4: Model 1 validation. The comparison of the 'true' sequence of hidden states (Simple model) with the most likely decoded sequence. Each horizontal bar corresponds to one alignment from the simulated data. The blue regions represent between-conversion segments, the yellow regions represent within-conversion segments. The parameters used for T and Φ in the decoding were the means of the posteriors obtained from the Metropolis-Hastings Algorithm. The accuracy of decoding was 90.45 % in this case (N=15).

parameters: λ_{end} and μ which are common across all alignments, λ_{begin} and m which are the baseline conversion and mutation probabilities corresponding to alignment 1 (i.e. $A_i = 1$), and the ages of the other alignments $A_i, i = 2,..,N$, relative to the first one.

We simulated the mismatch process on 5 hypothetical alignments of length $L = 1000$ each. We simulated the dataset 50 times, and each time we ran the Metropolis-Hastings algorithm to infer the posterior distributions associated to each parameter. We used uniform priors and multivariate normal proposal distributions with variance 0.000025. The latter yielded a satisfactory acceptance rate of 56%. The parameters used for the simulation and the predicted means obtained through the algorithm are summarized in Table C.3.

For every instance, we ran 3 MCMC chains until convergence to the stationary distribution was reached. These initial iterations were generally of the order of 10000

iterations and were discarded as *burn-in*. Afterwards the chains were let to run for another 30000 iterations and the posteriors were calculated from parameter samples of length 3×30000. We found that the true value of each parameter generally fell within the 5% and 95% predicted confidence bounds for each individual run (Figure C.5) and always within the mean 90% confidence interval over many runs (Figure C.6), confirming a good precision of our method. Furthermore, the overall bias was low (Figure C.7) and the normalized deviations of the posterior means from the true parameter values were generally within 30%. We also checked the accuracy of the decoding procedure, and we obtained as a mean over 50 runs, an algorithm accuracy of 80.5% (figure not shown). These results taken together, indicate the very good performance of the algorithm in extracting the values of model parameters, even with a number of stochastic runs as low as 50. Conceivably, if more stochastic runs are analyzed (e.g. ≥ 1000) the accuracy of the algorithm and its performance can only increase further.

Table C.3: Comparison of algorithm results for Model 4 and true parameter values used for 50 sets of simulated data. E[θ] denotes the average over the 50 posterior means.

Parameter	λ_{begin}	λ_{end}	μ	m	A_2	A_3	A_4	A_5
True θ	0.01	0.02	0.25	0.03	2	3	4	5
E[θ]	0.0126	0.0264	0.2478	0.0323	2.1076	3.1263	4.3148	4.9605

Table C.4: Deviations of the posterior means from the 'true' values for Model 4, obtained through the Metropolis-Hastings algorithm applied on 50 runs with 5 sequences each. 'True' parameters: $(\lambda_{begin}, \lambda_{end}, \mu, m, A_2, A_3, A_4, A_5)$=(0.01, 0.02, 0.25, 0.03, 2, 3,4,5). MB: mean bias; MSE: mean squared error; RMSE: root mean squared error; Norm.Dev: normalized deviations.

	λ_{begin}	λ_{end}	μ	m	A_2	A_3	A_4	A_5
MB	0.0026	0.0064	-0.0022	0.0023	0.1076	0.1263	0.3148	-0.0395
MSE	0.0000	0.0002	0.0001	0.0001	0.4474	0.7638	1.4131	1.2687
RMSE	0.0033	0.0097	0.0079	0.0064	0.5383	0.6627	0.8242	0.9266
Norm.Dev.	0.2606	0.3206	-0.0086	0.0755	0.0538	0.0421	0.0787	-0.0079

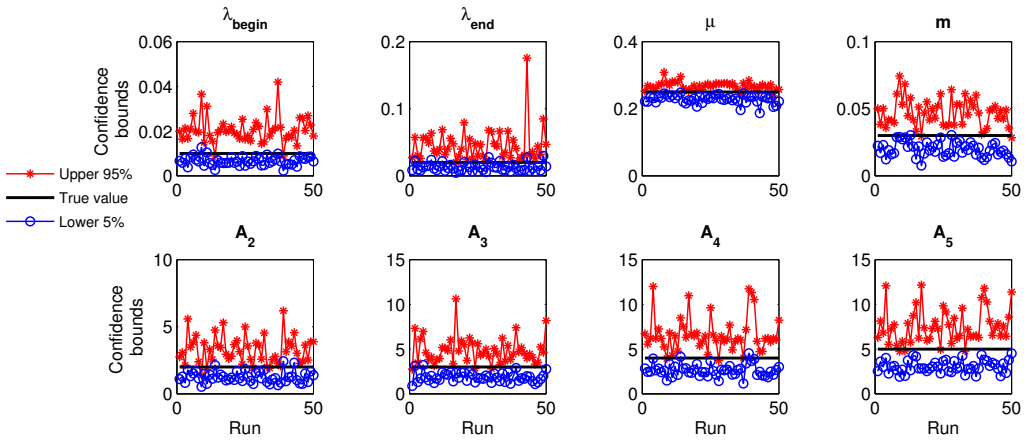


Figure C.5: Bayesian algorithm performance II (precision). Confidence bounds obtained from the posteriors of all parameters of Model 4 for 50 runs of the algorithm. The probabilities for each true parameter value to fall within the 90% predicted confidence interval were: $0.92, 0.84, 0.96, 0.92, 0.86, 0.90, 0.92, 0.92$. Thus, the true values of the parameter lied within the 90% confidence interval in approximately 90% of the cases.

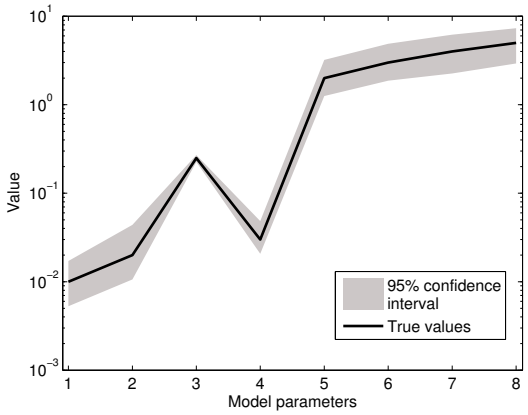


Figure C.6: Mean performance of the *Individual ages* model over 50 runs of the algorithm on 9 simulated alignments. The estimated mean confidence intervals (grey shaded area) contain the true values of the parameters (black line) from which the data was generated, listed in the order: $\lambda_{begin}, \lambda_{end}, \mu, m, A_2, A_3, A_4, A_5$.

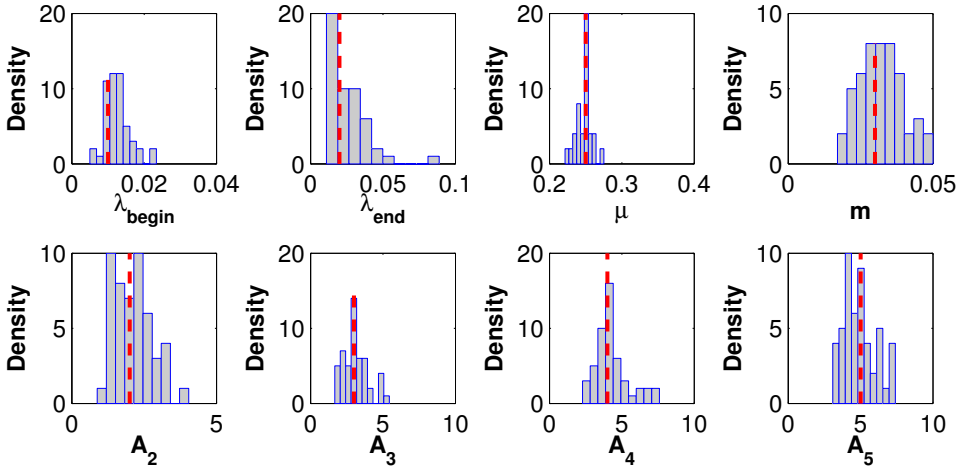


Figure C.7: Bayesian algorithm performance I (bias). The empirical distributions of posterior means (Model 4) obtained over 50 runs of the algorithm vs. the true parameter values (dotted red lines). Posterior means estimated via the Bayesian algorithm generally deviated $\leq$30 % from the true parameter values.

Appendix D

Variance around mean identity

D.1 Mutation-only case

In the model presented in Section 5.2.1, the number of identical aligned nucleotides in one gene pair, denoted by $n(t)$ changes randomly in time. To analyze its dynamics under mutation only, we consider a simple death stochastic process as follows. In an infinitesimal time interval Δt, $n(t)$ can only decrease by 1 unit or remain unchanged. $n(t+\Delta t) = n(t) - 1$ with transition probability $m_0 n(t)\Delta t$, where m_0 denotes the *per-aligned nucleotide per unit of time* probability of undergoing a mutation ($1 \to 0$ change) and equals $2\mu/NL$. Denote by $p_i(t)$ the state probabilities such that $p_i(t) = \text{Prob}[n(t) = i]$. Recall $n(0) = L$. Following classical Markov process theory, one can write the forward Kolmogorov equations in the limit $\Delta t \to 0$:

$$\frac{dp_i}{dt} = m_0(i+1)p_{i+1} - m_0 i p_i, \quad i = 1,..,L-1 \quad p_i(0) = \delta_{iL}. \tag{D.1}$$

Rearranging Eq. D.1 and substituting it into the equation for the probability generating function (p.g.f) $\frac{\partial P(z,t)}{\partial t} = \sum_{i=0}^{\infty} \frac{dp_i}{dt} z^i$, we obtain:

$$\frac{dP}{dt} = \frac{dP}{dz}(-m_0 z + m_0), \quad P(z,0) = z^L. \tag{D.2}$$

Substituting z by e^{θ}, whereby $P(e^{\theta},t) = M(\theta,t)$, we obtain an equation for the moment-generating-function $M(\theta,t)$ via: $dP/dz = 1/z\partial M/\partial\theta$. We have:

$$\frac{\partial M}{\partial t} = \frac{\partial M}{\partial \theta}[-m_0 + m_0 e^{-\theta}], \quad M(\theta,0) = e^{L\theta}. \tag{D.3}$$

The above equations can be solved by the method of characteristics, giving:

$$P(z,t) = [1 - e^{-m_0 t}(1-z)]^L, \quad \text{and} \quad M(\theta,t) = [1 - e^{-m_0 t}(1-e^{\theta})]^L. \tag{D.4}$$

The first moment, corresponding to the *mean* number of identical nucleotides, and the second moment, corresponding to the *variance* of n are given by:

$$\frac{\partial M}{\partial \theta}|_{\theta=0} = Le^{-m_0 t}, \quad \text{and} \quad \frac{\partial^2 M}{\partial \theta^2}|_{\theta=0} - \left(\frac{\partial M}{\partial \theta}|_{\theta=0}\right)^2 = Le^{-m_0 t}(1-e^{-m_0 t}). \tag{D.5}$$

These imply the mean pairwise identity (in the case of pure death process mutation) has the following mean and variance:

$$\bar{h}(t) = e^{-m_0 t} \quad \text{and} \quad Var[h(t)] = \frac{e^{-m_0 t}(1-e^{-m_0 t})}{L}, \tag{D.6}$$

a formula which agrees well with the numerical simulations of the mutation-only process (Figure 5.5). In analogy, we can derive a similar formula for the variance of the combined mutation-conversion process, which is a birth-death process, with m_0 as above, and $c_0 = 2\gamma l_c/LN(N-1)$, and mean pairwise identity given by $\bar{h}(t) = (c_0 + m_0 e^{-(c_0+m_0)t})/(c_0+m_0)$, and variance:

$$Var[h^{M-C}(t)] = \frac{1}{L}\bar{h}(t)(1-\bar{h}(t)). \tag{D.7}$$

However, numerical simulations suggest that this approximation, despite capturing well the time dependence, results in an under-estimation of the actual variance, thus it needs refinements to become accurate. Most probably the higher variance is due to the length of the conversion segments, not being small enough relative to L, and the number of gene pairs, not being large enough to dilute the indirect effects of gene conversion.

References

AGUR, Z., ABIRI, D. & VAN DER PLOEG, L.H. (1989). Ordered appearance of antigenic variants of African trypanosomes explained in a mathematical model based on a stochastic switch process and immune-selection against putative switch intermediates. *Proceedings of the National Academy of Sciences USA*, **86**, 9626–9630. 13, 14, 15, 36, 77

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723. 135

AKSOY, S., GIBSON, W. & LEHANE, M. (2003). Interactions between tsetse and trypanosomes with implications for the control of trypanosomiasis. *Advances in Parasitology*, **53**, 1–83. 27

AKSOY, S., BERRIMAN, M., HALL, N., HATTORI, M. & HIDE, W.E.A. (2005). A case for a glossina genome project. *Trends in Parasitology*, **21**, 107–111. 27

ALIZON, S. & VAN BAALEN, M. (2008a). Acute or Chronic? Within-host models with immune dynamics, infection outcome, and parasite evolution. *The American Naturalist*, **172**, E244–E256. 79, 190

ALIZON, S. & VAN BAALEN, M. (2008b). Transmission-virulence trade-offs in vector-borne diseases. *Theoretical Population Biology*, **74**, 6–15. 83, 94, 188

ALTHAUS, C. & BONHOEFFER, S. (2005). Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. *Journal of Virology*, **79**, 13572–13578. 36

ANDERSON, R. & MAY, R. (1982). Coevolution of hosts and parasites. *Parasitology*, **85**, 411–426. 28, 83

ANDERSON, R. & MAY, R. (1991). *Infectious diseases of humans: Dynamics and Control.*. Oxford University Press, Oxford. 35

ANTIA, R. & LIPSITCH, M. (1997). Mathematical models of parasite responses to host defences. *Parasitology*, **115**, 155–167. 36

ANTIA, R., NOWAK, M.A. & ANDERSON, R. (1996). Antigenic variation and the within-host dynamics of parasites. *Proceedings of the National Academy of Sciences USA*, **93**, 985–989. 13, 17, 18, 36, 77, 187

AWADALLA, P. (2003). The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, **4**, 50–60. 195

BAKER, R. (1992). Modelling trypanosomiasis prevalence and periodic epidemics and epizootics. *Mathematical Medicine and Biology*, **9**, 267–287. 27

BALBER, A. (1972). Trypanosoma brucei: Fluxes of the morphological variants in intact and x-irradiated mice. *Experimental Parasitology*, **31**, 307–319. 12

BARBOUR, A.G. & RESTREPO, B.I. (2000). Antigenic variation in vector-borne pathogens. *Emerging Infectious Diseases*, **6**, 449–457. 1, 4, 82

BARBOUR, A.G., DAI, Q., RESTREPO, B.I., STOENNER, H.G. & FRANK, S.A. (2006). Pathogen escape from host immunity by a genome program for antigenic variation. *Proceedings of the National Academy of Sciences USA*, **103**, 18290–18295. 1, 79, 82, 115

BARRY, J. (1986). Antigenic variation during Trypanosoma vivax infections of different host species. *Parasitology*, **92**, 51–65. xvi, 6, 7, 45, 80, 84, 107, 110

BARRY, J. (1997). The relative significance of mechanisms of antigenic variation in African trypanosomes. *Parasitology Today*, **13**, 212–218. 1, 7, 115, 120, 154

BARRY, J. & MCCULLOCH, R. (2001). Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Advanced Parasitology*, **49**, 1–70. 2, 50, 115

BAUER, B., AMSLER-DELAFOSSE, S., CLAUSEN, P., KABORE, I. & PETRICH-BAUER, J. (1995). Successful application of deltamethrin pour-on to cattle in a campaign against tsetse flies (glossina spp) in the pastoral zone of samorogouan, burkina faso. *Trop. Med. Parasitol.*, **46**, 183–189. 24

BAYES, T. & PRICE, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, **53**, 360–418. 210

BELL, G., PERELSON, A. & PIMBLEY, G. (1978). *Theoretical Immunology. Immunology series*. M.Dekker, New York, 8th edn. 35

BERRIMAN, M., GHEDIN, E., HERTZ-FOWLER, C., BLANDIN, G., RENAULD, H., BARTHOLOMEU, D., LENNARD, N., CALER, E., HAMLIN, N. & HAAS, B. (2005). The genome of the African trypanosome Trypanosoma brucei. *Science*, **309**, 416–422. xvi, 2, 5, 7, 43, 120, 154, 155

BEST, A., WHITE, A., KISDI, E., ANTONOVICS, J., BROCKHURST, M. & BOOTS, M. (2010). The evolution of host-parasite range. *The American Naturalist*, **176**, 63–71. 187

BETRAN, E., ROZAS, J., NAVARRO, A. & BARBADILLA, A. (1997). The estimation of the number and the length distribution of gene conversion tracts from population dna sequence data. *Genetics*, **146**, 89–99. 126, 127

BLUM, M., DOWN, J., GURNETT, A., CARRINGTON, M., TURNER, M., & WILEY, D. (1993). A structural motif in the variant surface glycoproteins of Trypanosoma brucei. *Nature*, **362**, 603–609. 8

BOLZONI, L., DE LEO, G., GATTO, M. & DOBSON, A. (2008). Body-size scaling in an SEI model of wildlife diseases. *Theoretical Population Biology*, **73**, 374–382. 109

BORST, P., RUDENKO, G., BLUNDELL, P. & VAN LEEUWEN, F. (1997). Mechanisms of antigenic variation in African trypanosomes. *Behring Institute Mitteilungen*, **99**, 1–15. 115, 120, 155

BRIGTWELL, R., DRANSFIELD, R. & KYORKU, C. (1991). Development of a low-cost tsetse trap and odour baits for glossina pallidipes and g. longipennis in kenya. *Med Vet Entomol*, **5**, 153–164. 24

BROOKS, S. & GELMAN, A. (1998). General methods of monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics.*, **7**, 434–455. 135

BRUN, R., BLUM, J., CHAPPUIS, F. & BURRI, C. (2010). Human African trypanosomiasis. *The Lancet.*, **375**, 148–159. 5

BULL, P., BUCKEE, C., KYES, S., KORTOK, M., THATHY, V., GUYAH, B., STOUTE, J., NEWBOLD, C. & MARSH, K. (2008). Plasmodium falciparum antigenic variation. mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular Microbiology*, **68**, 1519–1534. 46

CABLE, J., ENQUIST, B. & MOSES, M. (2007). The allometry of host-pathogen interactions. *PLoS ONE*, **2**, e1130. 110

CAPBERN, A., GIROUD, C., BALTZ, T. & MATTERN, P. (1977). Trypanosoma equiperdum: étude des variations antigéniques au cours de la trypanosomose expérimentale du lapin. *Experimental Parasitology*, **42**, 6–13. 7, 40, 45

CARRINGTON, M., MILLER, N., BLUM, M., RODITI, I., WILEY, D. & TURNER, M. (1991). Variant specific glycoprotein of Trypanosoma brucei consists of two domains each having an independently conserved pattern of cysteine residues. *Journal of Molecular Biology*, **221**, 823–835. 8, 10

CHECCHI, F., FILIPE, J. & BARRETT, M. (2008). The natural progression of Gambiense sleeping sickness: What is the evidence? *PLOS Neglected Tropical Diseases*, **2**, e303. 4

CROW, J. & KIMURA, M. (1956). Some genetic problems in natural populations. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, **4**, 1–22. 172

CROW, J. & KIMURA, M. (1970). *An Introduction to Population Genetis Theory*. Harper & Row, New York. 169

DE BOER, R.J. & PERELSON, A. (1994). T cell repertoires and competitive exclusion. *Journal of Theoretical Biology*, **169**, 375–390. 36

DEAN, S., MARCHETTI, R., KIRK, K. & MATHEWS, R. (2009). A surface transporter family conveys the trypanosome differentiation signal. *Nature*, **459**, 213–217. 11

DEITSCH, K., DEL PINAL, A. & WELLEMS, T. (1999). Intra-cluster recombination and Tar transcription switches in the antigenic variation of plasmodium falciparum. *Journal of Molecular and Biochemical Parasitology*, **101**, 107–116. 204

DIEKMANN, O. & HEESTERBEEK, J. (2000). *Mathematical Epidemiology of Infectious Diseases. Model building, analysis and interpretation*. Wiley. 28

DIEKMANN, O., HEESTERBEEK, J. & METZ, J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, **28**, 365–382. 27, 85

DIEKMANN, O., CHRISTIANSEN, F. & LAW, R. (1996). Evolutionary dynamics. *Journal of Mathematical Biology*, **34**, 483–688. 28

DOBSON, A. (2004). Population dynamics of pathogens with multiple host species. *The American Naturalist*, **164**, S64–S78. 109

DONELSON, J. (1995). Mechanisms of antigenic variation in *Borrelia Hermsii* and African trypanosomes. *Journal of Biological Chemistry*, **270**, 7783–7786. 115, 204

DURBIN, R., EDDY, S., KROGH, A. & MITCHISON, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK. 130, 134

EWALD, P.W. (1983). Host-parasite relations, vectors and the evolution of disease severity. *Annual Review of Ecology, Evolution, and Systematics*, **14**, 465–485. 28, 83, 190

FARIKOU, O., NJIOKOU, F., SIMO, G., ASONGANYI, T., CUNY, G. & GEIGER, A. (2010). Tsetse fly blood meal modification and trypanosome identification in two sleeping sickness foci in the forest of southern cameroon. *Acta Tropica*, **116**, 81–88. 27

FENTON, A., LELLO, J. & BONSALL, M. (2006). Pathogen responses to host immunity: the impact of time delays and memory on the evolution of virulence. *Proceedings of the Royal Society B.*, **273**, 2083–2090. 66

FEVRE, E., PICOZZI, K., JANNIN, J., MAUDLIN, I. & WELBURN, S. (2006). Human African trypanosomiasis: epidemiology and control. *Advanced Parasitology*, **61**, 167–221. 5

FISHER, R. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, **42**, 321–431. 169

FISHER, R. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford. 166, 168

FOLEY, J., NIETO, J., BARBET, A. & FOLEY, P. (2009). Antigen diversity in the parasitic bacterium Anaplasma phagocytophilum arises from selectively-represented, spatially clustered functional pseudogenes. *PLOS One*, **4**, e8265. 46, 204

FORNEY, G. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, **61**, 268–278. 146, 213

FRANK, M., DZIKOWSKI, R., AMULIC, B. & DEITSCH, K. (2007). Variable switching rates of malaria virulence genes are associated with chromosomal position. *Molecular Microbiology*, **64**, 1486–1498. 204

FRANK, S. (1999). A model for sequential dominance in African trypanosome infections. *Proceedings of the Royal Society - Biological Sciences*, **266**, 1397–1401. 8, 13, 19, 20, 31, 36, 77, 115, 116, 188

FRANK, S. (2002). *Immunology and Evolution of Infectious Disease*. Princeton University Press. 1, 30, 31, 32

FRANK, S. & BARBOUR, A. (2006). Within-host dynamics of antigenic variation. *Infection, genetics and evolution*, **6**, 141–146. 30

FRANK, S.A. (1996). Models of parasite virulence. *The Quarterly Review of Biology*, **71**, 37–78. 190

FRANK, S.A. & BUSH, R. (2007). Barriers to antigenic escape by pathogens: trade-off between reproductive rate and antigenic mutability. *BMC Evolutionary Biology*, **7**. 112

FRASER, C., W.P., H. & SPRATT, B. (2007). Recombination and the nature of bacterial speciation. *Science*, **315**, 476–480. 195

FUTSE, J., BRAYTON, K., DARK, M., KNOWLES, D. & PALMER, G. (2008). Superinfection as a driver of genomic diversification in antigenically variant pathogens. *Proc. Natl. Acad. Sci. USA*, **105**, 2123–2127. 46

GANDON, S. (2004). Evolution of multi-host parasites. *Evolution*, **58**, 455–469. 84, 112

GANDON, S., VAN BAALEN, M. & JANSEN, V. (2002). The evolution of parasite virulence, superinfection, and host resistance. *The American Naturalist*, **159**, 658–669. 112

GANUSOV, V.V., BERGSTROM, C.T. & ANTIA, R. (2002). Within-host population dynamics and the evolution of microparasites in a heterogeneous host population. *Evolution*, **52**, 213–223. 36

GILCHRIST, M. & SASAKI, A. (2002). Modeling host-parasite coevolution: a nested approach based on mechanistic models. *Journal of Theoretical Biology*, **218**, 289–308. 28, 83, 86

GILKS, W., RICHARDSON, S. & SPIEGELHALTER, D. (1996). *Markov Chain Monte Carlo In Practice*. Chapman and Hall, London, UK. 133, 211, 212

GILLESPIE, D. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, **81**, 2340–2361. 126, 157

GJINI, E., HAYDON, D., BARRY, J. & COBBOLD, C. (2010). Critical interplay between parasite differentiation,. host immunity, and antigenic variation in trypanosome infections. *The American Naturalist*, **176**, 424–439. 33

GOLDBERG, S. (1950). Phd thesis. *Cornell University*. 172

GRAHAM, A., ALLEN, J. & READ, A. (2005). The evolutionary causes and consequences of immunopathology. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 373–397. 118

GRAY, A. (1965). Antigenic variation in a strain of Trypanosoma brucei transmitted by G. morsitans and G. palpalis. *Journal of General Microbiology*, **41**, 195–214. 7, 12, 39, 40

GRENFELL, B., PYBUS, O., GOG, J., WOOD, J., DALY, J., MUMFORD, J. & HOLMES, E. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**, 327–332. 29

HABETEMARIAM, T., RUPPANNER, R., RIEMANN, H. & THEIS, J. (1983). Epidemic and endemic characteristics of trypanosomiasis in cattle: A simulation model. *Preventive Veterinary Medicine*, **1**, 137–145. 27

HAJDUK, S. & VICKERMAN, K. (1981). Antigenic variation in cyclically transmitted Trypanosoma brucei. variable antigen type composition of the first parasitaemia in mice bitten by trypanosome-infected Glossina morsitans. *Parasitology*, **83**, 609–621. 12, 39, 78

HANAGE, W., FRASER, C., TANG, J., CONNOR, T. & CORANDER, J. (2009). Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science*, **324**, 1454, doi:10.1126/science.1171908. 195

HARGROVE, J., SILAS, O., MSALILWA, J. & FOX, B. (2000). Insecticide-treated cattle for tsetse control: The power and the problems. *Med Vet Entomol*, **14**, 123–130. 24, 27

HAYDON, D. & MATHEWS, L. (2007). Introduction. Cross-scale influences on epidemiological dynamics: from genes to ecosystems. *Journal of the Royal Society Interface*, **4**, 763–765. 2

HAYDON, D. & WOOLHOUSE, M. (1998). Immune avoidance strategies in RNA viruses: fitness continuums arising from trade-offs between immunogenicity and antigenic variability. *Journal of Theoretical Biology*, **193**, 601–612. 188

HEITMAN, J. (2006). Sexual reproduction and the evolution of microbial pathogens. *Current Biology*, **16**, R711–25. 195

HIDE, G. & TAIT, A. (2004). Genetics and molecular epidemiology of trypanosomes. *In: Maudlin. I.. Holmes. P., Miles, M. (Eds.) CAB International*, 77–93. 3

HILLIKER, A., HARAUZ, A., REAUME, M., S., G., CLARK, S. & CHOVNICK, A. (1994). Meiotic gene conversion tract length distribution within the rosy locus of drosophila melanogaster. *Genetics*, **137**, 1019–1026. 126, 127

HSIA, R., BEALS, T. & BOOTHROYD, J. (1996). Use of chimeric recombinant polypeptides to analyse conformational, surface epitopes on trypanosome variant surface glycoproteins. *Journal of Molecular Microbiology*, **19**, 53–63. 8

HUTCHISON, O., PICOZZI, K., JONES, N., MOTT, H., SHARMA, R., WELBURN, S. & CARRINGTON, M. (2007). Variant Surface Glycoprotein gene repertoires in Trypanosoma brucei have diverged to become strain-specific. *BMC Genomics*, **8**. 27, 32, 151, 190

ILLIAN, J., PENTINNEN, H., H., S. & STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley and Sons Chichester, NY. 136, 146, 213

ISHI, K., MATSUDA, H., IWASA, Y. & SASAKI, A. (1989). Evolutionarily stable mutation rate in a periodically changing environment. *Genetics*, **121**, 163–174. 188

JACKSON, A., SANDERS, M., BERRY, A., MCQUILLAN, J., ASLETT, M., QUAIL, M., CHUKUALIM, B., CAPEWELL, P., MACLEOD, A., MELVILLE, S., GIBSON, W., BARRY, J., BERRIMAN, M. & HERTZ-FOWLER, C. (2010). The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human African trypanosomiasis. *PLoS Neglected Tropical Diseases*, **4**, e658. 151

JANEWAY, C., TRAVERS, P., WALPORT, M. & SHLOMCHIK, M. (2005). *Immunobiology: the immune system in health and disease 6th ed.*. Garland Science Publishing, New York. 40

KAREV, G., WOLF, Y. & KOONIN, E. (2003). Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, **19**, 1889–1900. 192

KARLIN, S. & TAYLOR, H. (1975). *A first course in stochastic processes. Second ed.*. Academic Press, New York. 126

KEPLER, T. & PERELSON, A. (1995). Modeling and optimization of populations subject to time-dependent mutation. *Proceedings of the National Academy of Sciences USA*, **92**, 8219–8223. 41, 63, 64

KERMACK, W. & MCKENDRICK, A. (1927). A contribution to the mathematical theory of epidemics. *Royal Society of London Proceedings (A)*, **115**, 700–721. 35

KIMURA, M. (1955a). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences*, **41**, 141–150. 173

KIMURA, M. (1955b). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbour Symposium on Quantitative Biology*, **20**, 33–53. 169, 173

KOELLA, J. & ANTIA, R. (1995). Optimal pattern of replication and transmission for parasites with two stages in their life cycle. *Theoretical Population Biology*, **47**, 277–291. 83

KOHAGNE, T.L., MENGUE, P., M'EYI, MIMPFOUNDI, R. & LOUIS, F. (2010). Glossina feeding habits and diversity of species of trypanosomes in an active focus of human African trypanosomiasis in Gabon. *Bull. Soc. Pathol. Exot.*, **103**, 264–271. 27

KOSINSKI, R. (1980). Antigenic variation in trypanosomes: a computer analysis of variant order. *Parasitology*, **80**, 343–357. 13, 14, 36, 77

KRAFSUR, E. (2009). Tsetse flies: Genetics, evolution and role as vectors. *Infection, Genetics and Evolution*, **9**, 124–141. 26

KYES, S., KRAEMER, S. & SMITH, J. (2007). Antigenic variation in *Plasmodium falciparum*: gene organization and regulation of the *var* multigene family. *Eukaryotic Cell*, **6**, 1511–1520. 115, 204

LEVIN, S. & PIMENTAL, D. (1981). Selection of intermediate rates of increase in parasite-host systems. *The American Naturalist*, **117**, 308–315. 28, 83

LIPSITCH, M. & O'HAGAN, J. (2007). Patterns of antigenic diversity and the mechanisms that maintain them. *Journal of the Royal Society Interface*, **4**, 787–802, doi:10.1098/rsif.2007.0229. 82

LUCIANI, F. & ALIZON, S. (2009). The evolutionary dynamics of a rapidly mutating virus within and between hosts: The case of Hepatitis C virus. *PLoS Computational Biology*, **5**, e1000565–. 36

LUCKINS, A. (1972). Effects of x-irradiation and cortisone treatment of albino rats on infections with brucei-complex trypanosomes. *Transactions of the Royal Society of tropical medicine and hygiene*, **66**, 130–139. 12

LYTHGOE, K., L.J., M., READ, A. & BARRY, J. (2007). Parasite-intrinsic factors can explain ordered progression of trypanosome antigenic variation. *Proceedings of the National Academy of Sciences USA*, **104**, 8095–8100. 2, 8, 13, 16, 30, 36, 37, 38, 39, 40, 43, 77, 115

MACKINNON, M.J. & READ, A. (1999). Genetic relationships between parasite virulence and transmission in the rodent malaria *Plasmodium chabaudi*. *Evolution*, **53**, 271–281. 83

MACKINNON, M.J. & READ, A. (2004). Immunity promotes virulence evolution in a malaria model. *PLOS Biology*, **2**. 99

MARCELLO, L. & BARRY, J. (2007a). Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Research*, **17**, 1344–52. xvi, 7, 8, 10, 30, 31, 42, 51, 79, 115, 120, 151, 175, 177, 204

MARCELLO, L. & BARRY, J. (2007b). From silent genes to noisy populations - dialogue between the genotype and phenotypes of antigenic variation. *Eukaryotic Microbiology*, **54**, 14–17. xvii, 2, 9, 30, 46, 71, 120, 155, 191

MATHEWS, K. (2011). Controlling and coordinating development in vector-transmitted parasites. *Science*, **331**, 1149–1153. 11, 118, 194

MAUDLIN, I. & WELBURN, S. (1989). A single trypanosome is sufficient to infect a tsetse fly. *Annals of Tropical Medicine and Parasitology*, **83**, 431–433. 96

MCCULLOCH, R. & BARRY, J. (1999). A role for RAD51 and homologous recombination in Trypanosoma brucei antigenic variation. *Genes Development*, **13**, 2875–2888. 31, 204

MCDERMOTT, J. & COLEMAN, P. (2001). Comparing apples and oranges–model-based assessment of different tsetse-transmitted trypanosomosis control strategies. *International Journal for Parasitology*, **31**, 603–609. 27

MCKENZIE, F. & BOSSERT, W. (2005). An integrated model of *Plasmodium falciparum* dynamics. *Journal of Theoretical Biology*, **232**, 411–426. 187

MCLINTOCK, L., TURNER, C. & VICKERMAN, K. (1993). Comparison of the effects of immune killing mechanisms on Trypanosoma brucei parasites of slender and stumpy morphology. *Parasite Immunology*, **15**, 475–480. 39, 43, 194

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. & E., T. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092, doi:10.1063/1.1699114. 211

MIDEO, N., ALIZON, S. & DAY, T. (2008). Linking within- and between-host dynamics in the evolutionary epidemiology of infectious disease. *Trends in Ecology and Evolution*, **23**, 511–517. 83

MILLER, E. & TURNER, M. (1981). Analysis of antigenic types appearing in first relapse populations of clones of Trypanosoma brucei. *Parasitology*, **82**, 63–80. 50

MILLER, E., ALLAN, L., & TURNER, M. (1984). Topological analysis of antigenic determinants on a variant surface glycoprotein of Trypanosoma brucei. *Journal of Molecular and Biochemical Parasitology*, **13**, 67–81. 8

MILLIGAN, P. & BAKER, R. (1988). A model of tsetse-transmitted animal trypanosomiasis. *Parasitology*, **96**, 211–239. 27

MORRISON, L. (2007). Phd thesis. *University of Glasgow*. 68

MORRISON, L., MAJIWA, P., READ, A. & BARRY, J. (2005). Probabilistic order in antigenic variation of Trypanosoma Brucei. *International Journal for Parasitology*, **35**, 961–972. 7, 12, 30, 40, 42, 45, 51, 55, 115, 155

MORRISON, L., MARCELLO, L. & MCCULLOCH, R. (2009). Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. *Cellular Microbiology*, **11**, 1724–1734. 7, 30, 31, 45, 115, 120, 205

MORRISON, W. & MURRAY, M. (1985). The role of humoral immune responses in determining susceptibility of A/J and C57BL/6 mice to infection with Trypanosoma congolense. *Parasite immunology*, **7**, 63–79. 84

MOXON, R., RAINEY, P., NOWAK, M. & LENSKI, R. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology*, **4**, 24–33. 191

NOWAK, M., MAY, R. & ANDERSON, R. (1990). The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *AIDS*, **4**, 1095–1103. 13, 16, 17, 36, 57, 187

OHTA, T. (2010). Gene conversion and evolution of gene families: An overview. *Genes*, **1**, 349–356. 121

PARHAM, P., ADAMS, E. & ARNETT, K. (1995). The origins of hla-a,b,c polymorphism. *Immunology Reviews*, **143**, 141–180. 149

PAUL, R., BONNET, S., BOUDIN, C., TCHUINKAM, T. & ROBERT, V. (2007). Aggregation in malaria parasites places limits on mosquito infection rates. *Infection, Genetics and Evolution*, **7**, 577–586. 96

PEACOCK, L., FERRIS, V., SHARMA, R., SUNTER, J., BAILEY, M., CARRINGTON, M. & GIBSON, W. (2011). Identification of the meiotic life cycle stage of Trypanosoma brucei in the tsetse fly. *Proceedings of the National Academy of Sciences USA*, doi:10.1073/pnas.1019423108. 27, 195

PERELSON, A. & WIEGEL, F. (2004). Some scaling principles for the immune system. *lmmunology and Cell Biology*, **82**, 127–131. 110

PERELSON, A. & WIEGEL, F. (2009). Scaling aspects of lymphocyte trafficking. *Journal of Theoretical Biology*, **257**, 9–16. 110

PETERS, R. (1983). *The ecological implications of body size.*. Cambridge studies in ecology, Cambridge University Press, New York. 109, 118

PICOZZI, K., FEVRE, E., ODIIT, M., CARRINGTON, M., EISLER, M., MAUDLIN, I. & WELBURN, S. (2005). Sleeping sickness in Uganda: a thin line between two fatal diseases. *British Medical Journal*, **331**, 1238–1242. 5

RADWANSKA, M., GUIRNALDA, P., DE TREZ, C., RYFFEL, B., BLACK, S. & MAGEZ, S. (2008). Trypanosomiasis-induced b cell apoptosis results in loss of protective anti-parasite antibody responses and abolishment of vaccine-induced memory responses. *PLoS Pathogens*, **4**, e1000078. 12, 75, 107

RAPIN, N., KESMIR, C., FRANKILD, S., NIELSEN, M., LUNDEGAARD, C., BRUNAK, S. & LUND, O. (2006). Modelling the human immune system by combining bioinformatics and systems biology approaches. *Journal of Biological Physics*, **32**, 335–353. 35

RECKER, M. & GUPTA, S. (2006). Conflicting immune responses can prolong length of infection in Plasmodium falciparum malaria. *Bulletin of Mathematical Biology*, **68**, 821–835. 73

RECKER, M., NEE, S., BULL, P., KINYANJUI, S., MARSH, K., NEWBOLD, C. & GUPTA, S. (2004). Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. *Nature*, **429**, 555–558. 15, 16, 71

RECKER, M., BUCKEE, C., SERAZIN, A., KYES, S., PINCHES, R., CHRISTODOLOU, Z., SPRINGER, A., GUPTA, S. & NEWBOLD, C. (2011). Antigenic variation in Plasmodium falciparum malaria involves a highly structured switching pattern. *PLOS Pathogens*, **7**, e1001306, doi:10.1371/journal.ppat.1001306. 46, 79, 115, 194

REECE, S., ALI, E., SCHNEIDER, P. & BABIKER, H. (2010). Stress, drugs and the evolution of reproductive restraint in malaria parasites. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 3123–3129. 118

REEDER, J. & BROWN, G. (1996). Antigenic variation and immune evasion in Plasmodium falciparum malaria. *Immunology and Cell Biology*, **74**, 546–554. 30

REGOES, R.R., NOWAK, M.A. & BONHOEFFER, S. (2000). Evolution of virulence in a heterogeneous host population. *Evolution*, **54**, 64–71. 112

REUNER, B., VASSELLA, E., YUTZY, B. & BOSHART, M. (1997). Cell density triggers to stumpy differentiation of Trypanosoma Brucei bloodstream forms in culture. *Molecular and Biochemical Parasitology*, **90**, 269–280. 11, 43, 80, 194

ROBINSON, N., BURMAN, N. & MELVILLE S.E. AMD BARRY, J. (1999). Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes. *Molecular And Cellular Biology*, **19**, 5839–5846. 40, 50, 79, 204

ROGERS, D. (1988). A general model for the African trypanosomiases. *Parasitology*, **97**, 193–212. 22, 27

ROGERS, D. & RANDOLPH, S. (1991). Mortality rates and population density of tsetse flies correlated with satellite imagery. *Nature*, **351**, 739–741. 26, 27

SASAKI, A. (1994). Evolution of Antigen Drift/Switching: Continuously Evading Pathogens. *Journal of Theoretical Biology*, **168**, 291–308. 13, 18, 19, 57, 187

SASAKI, A. & HARAGUCHI, Y. (2000). Antigenic drift of viruses within a host: A finite site model with demographic stochasticity. *Journal of Molecular Evolution*, **51**, 245–255. 78

SASAKI, A. & IWASA, Y. (1991). Optimal growth schedule of pathogens within a host: switching between lytic and latent cycles. *Journal of Theoretical Population Biology*, **39**, 201–239. 83, 188

SAVILL, N. & SEED, J. (2004). Mathematical and statistical analysis of the Trypanosoma brucei slender to stumpy transition. *Parasitology*, **128**, 53–67. 14, 43, 77, 189

SAWYER, S. (1989). Statistical tests for detecting gene conversion. *Molecular and Biological Evolution*, **6**, 526–538. 122

SCHIMD-HEMPEL, P. (2008). Parasite immune evasion: a momentous molecular war. *Trends in Ecology and Evolution*, **23**, 318–326. 1

SEED, J. (1978). Competition among serologically different clones of Trypanosoma brucei gambiense in vivo. *Protozoology*, **25**, 526–529. 12, 36, 77

SEED, J. & WENCK, M. (2003). Role of the long slender to short stumpy transition in the life cycle of the African trypanosome. *Kinetoplastid Biology and Disease*, **2**. 189

SEN, A. & SMITH, T. (1995). *Gravity models of spatial interaction behavior*. Springer, Berlin. 186, 203

SHANNON, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, **30**, 50–64. 69

SIMARRO, P.P., JANNIN, J. & CATTAND, P. (2008). Eliminating human african trypanosomiasis: Where do we stand and what comes next? *PLOS Medicine*, **5**, e55. xvi, xvii, 4, 21, 22, 25, 26

SIMARRO, P.P., DIARRA, A., POSTIGO, J., FRANCO, J. & JANNIN, J. (2011). The human african trypanosomiasis control and surveillance programme of the world health organization 20002009: The way forward. *PLOS Negl.Trop.Dis.*, **5**, e1007. xvii, 23

SONG, G., HSU, C., RIEMER, C. & MILLER, W. (2011). Evaluation of methods for detecting gene conversion in clusters. *BMC Bioinformatics*, **12**, s45. 122

SPIEGELHALTER, D., BEST, N., CARLIN, B. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, **64**, 583–616. 135

TAYLOR, K. (1998). Immune responses of cattle to African trypanosomes: protective or pathogenic? *International Journal of Parasitology*, **28**, 219–240. 3, 12, 25, 84

THON, G., BALTZ, T., GIROUD, C. & ELSEN, H. (1990). Trypanosome variable surface glycoproteins: composite genes and order of expression. *Genes and Development*, **4**, 1374–1383. 7, 120

TIMMERS, H., DE LANGE, T., KOOTER, J. & BORST, P. (1987). Coincident multiple activations of the same surface antigen gene in Trypanosoma brucei. *Journal of Molecular Biology*, **194**, 81–90. 79

TURCHIN, P., OKSANEN, L., EKERHOLM, P., OKSANEN, T. & HENTTONEN, H. (2000). Are lemmings prey or predators? *Nature*, **405**, 562–565. 70

TURNER, C. (1997). The rate of antigenic variation in fly-transmitted and syringe-passaged infections of trypanosoma brucei. *FEMS Microbiology Letters*, **153**, 227–231. 6, 205

TURNER, M. (1999). Antigenic variation in Trypanosoma brucei infections: an holistic view. *Journal of Cell Science*, **112**, 3187–3192. 2, 7, 30

TURNER, M. & BARRY, J. (1989). High frequency of antigenic variation in Trypanosoma brucei rhodesiense infections. *Parasitology*, **99**, 67–75. 36, 41, 43, 77

TURNER, M., ASLAM, N. & DYE, C. (1995). Replication, differentiation, growth and the virulence of Trypanosoma brucei infections. *Parasitology*, **111**, 289–300. 43

TYLER, K., HIGGS, P.G., MATHEWS, K. & GULL, K. (2001). Limitation of Trypanosoma brucei parasitaemia results from density-dependent parasite differentiation and parasite killing by the host immune response. *Proceedings of the Royal Society B: Biological Sciences*, **268**, 2235–2243. 14, 39, 40, 43, 77

VALE, G., LOVEMORE, D., FLINT, S. & COCKBILL, G. (1988). Odour-baited targets to control tsetse flies, glossina spp. (diptera: Glossinidae), in zimbabwe. *Bulletin of Entomological Research*, **78**, 31–49. 24

VAN DEN BOSSCHE, P., KY-ZERBO, A., BRANDT, J., MARCOTTY, T., GEERTS, S. & DE DEKEN, R. (2005). Transmissibility of Trypanosoma brucei during its development in cattle. *Trop. Med. Int. Health*, **10**, 833–839. 87, 96

VAN DER WERF, A., VAN ASSEL, S., AERTS, D., STEINERT, M. & PAYS, E. (1990). Telomere interactions may condition the programming of antigen expression in Trypanosoma brucei. *The Embo Journal*, **9**, 1035–1040. 204

VOLTERRA, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, **118**, 558–560. 35

VREYSEN, M., SALEH, K., ALI, M., ABDULLA, A. & ZHU, Z. (2000). Glossina austeni (diptera: Glossinidae) eradicated on the island of unguja, zanzibar, using the sterile insect technique. *J. Econ. Entomol.*, **93**, 123–135. 24

WELBURN, S., FEVRE, E., COLEMAN, P., ODIIT, M. & MAUDLIN, I. (2001). Sleeping sickness: a tale of two diseases. *Trends in Parasitology*, **17**, 19–24. 3

WELBURN, S., FEVRE, E., COLEMAN, P. & MAUDLIN, I. (2004). The trypanosomiases. *In: Maudlin. I. Holmes. P., (Eds.) CAB International.*, 219–232. 3

WESTESSON, O. & HOLMES, I. (2009). Accurate detection of recombinant breakpoints in whole-genome alignments. *PloS Computational Biology*, **5**, e1000318. 122

WHO (2006). Human African Trypanosomiasis (sleeping sickness): epidemiological update. *Weekly Epidemiology Record*, **81**, 69–80. 3, 21, 22

WILLETT, K. (1963). Some principles of the epidemiology of human trypanosomiasis in Africa. *Bulletin WHO*, **28**, 645–652. 24, 26

WOOLHOUSE, M., J.W., H. & MCNAMARA, J. (1993). Epidemiology of trypanosome infections of the tsetse fly Glossina pallidipes in the Zambezi Valley. *Parasitology*, **106**, 479–485. 27

WRIGHT, S. (1930). Review of Fisher. *Journal of Heredity*, **21**, 349–356. 173

WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97–159. 166, 168

WRIGHT, S. (1945). The differential equation of the distribution of gene frequencies. *Proceedings of the National Academy of Sciences. USA*, **31**, 382–389. 169