IMPERIAL COLLEGE, LONDON

The Determination of the Basis of HLA Class I Associated Protection in HTLV-I Infection

by

Aidan MacNamara

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Medicine
Department of Immunology

July 2010

Declaration of Authorship

I, Aidan MacNamara, declare that this thesis titled, 'The Determination of the Basis of HLA Class I Associated Protection in HTLV-I Infection' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

“Satan delights equally in statistics and in quoting scripture...”

H. G. Wells

Abstract

Aim

The aim of this work was to investigate what constitutes a protective cellular immune response to HTLV-I infection, focusing on understanding the relationship between HLA class I and the outcome of infection. Initially, we conducted an analysis of our patient cohort to identify robust associations between HLA class I type and infection outcome. We then identified and optimized the most accurate epitope prediction software and applied these techniques to identify the constituents of a protective HLA class I restricted response. In related projects, we identified viral factors that impact on the efficiency of the CD8$^+$ T cell response and also investigated the role of KIR:HLA genotype in the outcome of HTLV-I infection.

Results

We produced a more accurate method of epitope prediction that we call Metaserver. This method was used to predict HTLV-I epitopes that bind to the HLA class I alleles of our patient cohort. We then developed a statistical method that demonstrated targeting the HTLV-I protein HBZ is beneficial in terms of disease status and proviral load. In the related analysis of the CD8$^+$ T cell response to HTLV-I infected cells, we produced a model measuring the lysis of infected cells as a function of antigen expression and showed a positive relationship between antigen expression and lysis efficiency. For our analysis of KIR genotype, we found no relationship between KIR:HLA interactions and disease status or proviral load in HTLV-I infection.

Conclusion

We conclude that CD8$^+$ T cells play a central role in the control of HTLV-I and that CD8$^+$ cells specific to HBZ, not the immunodominant protein Tax, are the most effective. We suggest that HBZ plays a central role in HTLV-I persistence. The statistical method we used to quantify the specificity of HLA class I binding to viral epitopes is applicable across all pathogens, even where data are sparse.

Acknowledgements

I would like to thank my supervisor, Becca Asquith, for her advice, time, support and encouragement throughout my PhD and for producing a great environment for learning. I would also like to thank Charles Bangham for his help and support and his infectious enthusiasm for HTLV-I research and science in general.

Many thanks to Ulrich Kadolsky. As a computational biologist, with the emphasis on biology, I was fortunate to be sitting next to a real computer scientist, or else everything would have taken a good deal longer. I would also like to thank my collaborators within the Department of Immunology, namely Tarek Kattan and Aileen Rowan. It was very enjoyable working together to produce research that combined experimental and theoeretical approaches. I would also like to thank Graham Taylor and Silva Hilburn for their help with my research and everyone in the department for their general help and support.

I am grateful to the Wellcome Trust for funding this work.

Finally, I would like to thank Jaydee, Noah and my family for their patience with all those hours spent in front of the computer!

Contents

List of Figures

List of Tables

Abbreviations

AC	**A**symptomatic **C**arriers
ATL	**A**dult **T**-cell **L**eukaemia/Lymphoma
AUC	**A**rea **U**nder **C**urve
CNS	**C**entral **N**ervous **S**ystem
CSF	**C**erebrospinal **F**luid
CTL	**C**ytotoxic **T** **L**ymphocyte
dsDNA	**D**ouble **S**tranded **D**eoxy**r**ibo**n**ucleic acid
ELISpot	**E**nzyme-**L**inked Immuno**spot** Assay
FACS	**F**luoresence **A**ctivated **C**ell **S**orting
FCS	**F**oetal **C**alf **S**erum
HAM/TSP	**H**TLV-I **A**ssociated **M**yelopathy / **T**ropical **S**pastic **P**araparesis
HBV	**H**epatitis **B** **V**irus
HBZ	**H**TLV-I **bZ**IP-Factor
HCV	**H**epatitis **C** **V**irus
HIV	**H**uman **I**mmunodeficiency **V**irus
HLA	**H**uman **L**eucocyte **A**ntigen
IFN	**I**nter**f**ero**n**
Ig	**I**mmuno**g**lobulin
KIR	**K**iller Cell **I**munnoglobulin-like **R**eceptor
LTR	**L**ong **T**erminal **R**epeat
MHC	**M**ajor **H**istocompatibility **C**omplex
ORF	**O**pen **R**eading **F**rame
PBMC	**P**eripheral **B**lood **M**ononuclear **C**ells
SD	**S**tandard **D**eviation

TCR	**T**-**C**ell **R**eceptor
TNF	**T**umour **N**ecrosis **F**actor
WHO	**W**orld **H**ealth **O**rganisation

Symbols

Symbol	**Name**	**Unit**
c	Rate of increase of tax expression	Tax$^+$CD4$^+$ / 24hrs
c_1	Rate of increase of Tax$^{\text{low}}$ expression	Tax$^{\text{low}}$CD4$^+$ / 24hrs
c_2	Rate of increase of Tax$^{\text{high}}$ expression	Tax$^{\text{high}}$CD4$^+$ / 24hrs
y	Proportion of Tax$^{\text{low}}$CD4$^+$ cells	%Tax$^{\text{low}}$CD4$^+$cells
w	Proportion of Tax$^{\text{high}}$CD4$^+$ cells	%Tax$^{\text{high}}$CD4$^+$cells
ϵ	Rate of killing of Tax$^+$CD4$^+$ cells per CD8$^+$	Tax$^+$CD4$^+$ / CD8$^+$ / 24hrs
ϵ^{low}	Rate of killing of Tax$^{\text{low}}$CD4$^+$ cells per CD8$^+$	Tax$^{\text{low}}$CD4$^+$ / CD8$^+$ / 24hrs
ϵ^{high}	Rate of killing of Tax$^{\text{high}}$CD4$^+$ cells per CD8$^+$	Tax$^{\text{high}}$CD4$^+$ / CD8$^+$ / 24hrs

Chapter 1

Aim and Overview

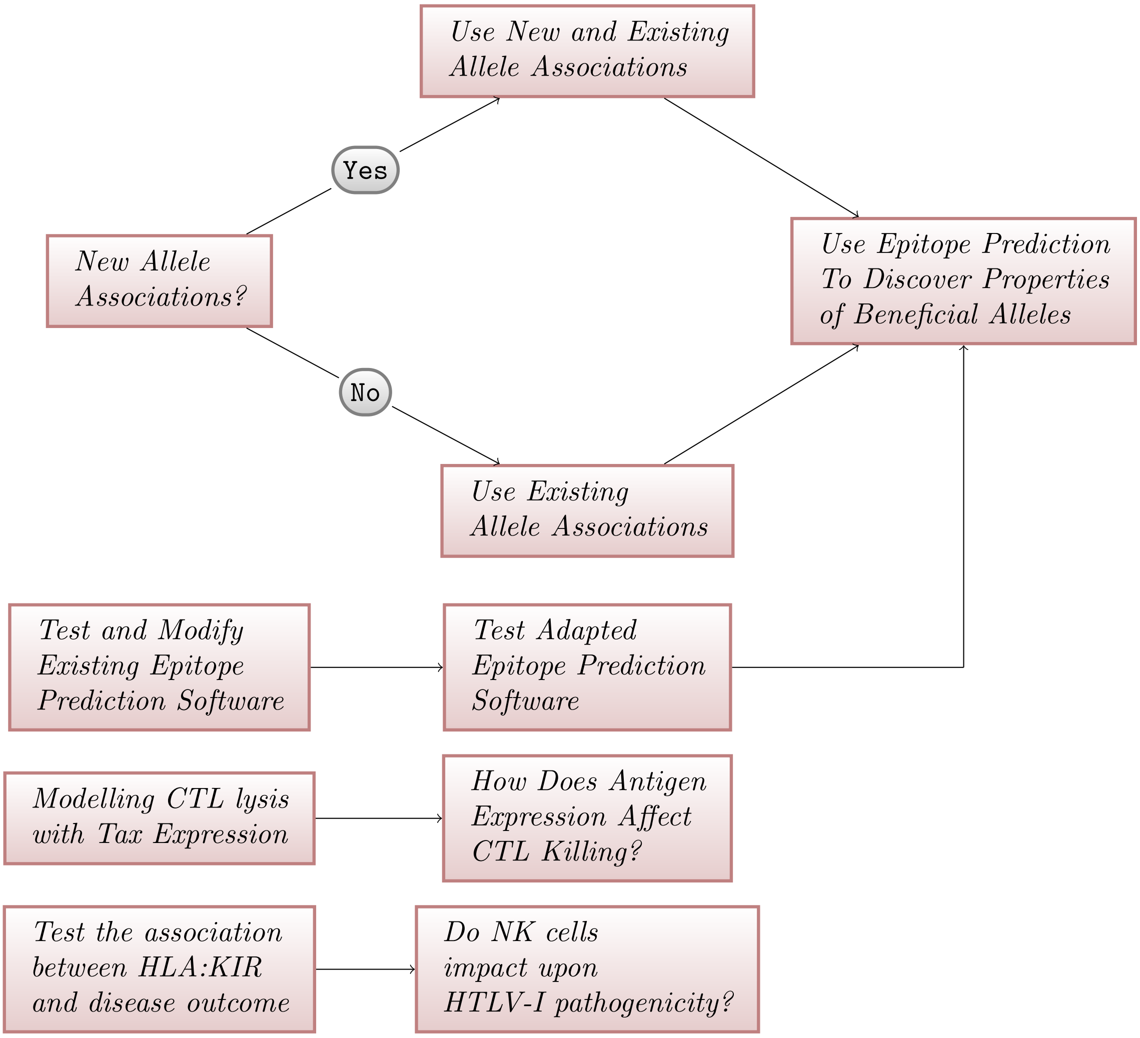
1.1 Introduction

Host genetic factors, including the HLA class I genotype, are major determinants of susceptibility to infectious disease in humans. However, it is currently a difficult task to demonstrate a direct link between the host immune response and the outcome of viral infections in either human or animal populations [1]. Human T-lymphotropic virus-I (HTLV-I) is a persistent retrovirus that infects 10-20 million people worldwide. The virus is endemic in the Caribbean, Japan and parts of Africa. Most infected people remain healthy, but 1-2% develop a progressive paralytic myelopathy (HTLV-I associated myelopathy/tropical spastic paraparesis; HAM/TSP) and a further 2-3% develop an aggressive T cell leukaemia/lymphoma. The reasons for these different outcomes is unknown. What is known is that the risk of HAM/TSP is determined, in part, by the host's HLA class I alleles.

Cells that have become infected with a virus are recognized by the host immune system because they display fragments of the pathogen bound to HLA class I molecules on the infected cell surface. Different people have a diverse range of shapes that make up the HLA class I molecules, owing to their different alleles. Thus, the molecules bind to different parts of the pathogen proteome and present this peptide (or epitope in this context) to CD8$^+$ cytotoxic T lymphocytes (CTLs). Once CTLs recognize the HLA-peptide complex, they are capable of destroying the infected cell by the release of lytic granules containing cytotoxic effector proteins. This results in the destruction of the target cell by apoptosis. An effective CTL response has been shown to confer protection against viral infection, such as HIV [2] and HTLV-I [3]. The effectiveness of the response varies between individuals and part of this variation is thought to be due to differences in the host genotype.

The ultimate aim of the project is to increase our understanding of why some HLA class I molecules are better than others at eliciting a more effective immune response. This would increase our knowledge of a key part of the immune system and specifically the design and implementation of improved vaccines.

1.2 PhD Design


Use New and Existing Allele Associations
Yes
New Allele Associations?
Use Epitope Prediction To Discover Properties of Beneficial Alleles
No
Use Existing Allele Associations
Test and Modify Existing Epitope Prediction Software
Test Adapted Epitope Prediction Software
Modelling CTL lysis with Tax Expression
How Does Antigen Expression Affect CTL Killing?
Test the association between HLA:KIR and disease outcome
Do NK cells impact upon HTLV-I pathogenicity?

1.2.1 Identifying Alleles Associated with Risk or Prevention in HTLV-I Infection

The work of Jeffery *et al.* [1] produced evidence that a number of HLA class I alleles are associated with either disease risk or protection in HTLV-I infection. Added to this, results showing that class I heterozygosity is associated with significantly lower proviral loads [4] would suggest that the protective effect of the HLA haplotype extends to a range of alleles.

Hence, my initial task was to reanalyze a database of individuals from Kagoshima, Japan, who had been infected with HTLV-I and displayed symptoms of HAM/TSP (see Section 2.2.1: HAM/TSP description) or remained asymptomatic. Chapter 3 details the progression of this work.

1.2.2 Epitope Prediction

There are relatively few experimentally confirmed HTLV-I epitopes for HLA class I alleles compared to HIV. Therefore, in order to test the protective properties of specific HLA class I alleles, it was necessary to use epitope prediction software to predict what epitopes these alleles bind to. The aim of this section was to test the accuracy and predictive power of a number of web-based prediction servers. The starting point was NetCTL v1.2, an integrated web-based prediction method that used information pertaining to proteasomal cleavage, TAP and HLA-peptide binding in epitope prediction. We tested and modified this method, in conjunction with other epitope prediction software, to produce a novel method of epitope prediction that we used for the purpose of discovering the HTLV-I epitopes of "beneficial" or "detrimental" alleles. The details of this work are in Chapter 4.

1.2.3 The Properties of HLA class I Alleles Associated with Disease Outcome

Combining the two strands of research, Section 1.2.1 and Section 1.2.2, gave us the ability to predict HTLV-I epitopes for each of the HLA class I alleles contained within the Kagoshima database. Hence, we were able to test what properties of these epitopes were associated with disease risk and proviral load. For instance, we asked the question, "do alleles associated with protection from disease bind to specific regions of the HTLV-I proteome?". The details of this work are in Chapter 5.

1.2.4 Modelling CTL Efficiency in Terms of Tax Expression

In the work detailed above we analyzed the host genetic factors that determine the efficiency of the CD8$^+$ T cell response. We then extended this work to investigate the CD8$^+$ T cell response itself in terms of its properties and dynamics.

In collaboration with experimentalists within the Department of Immunology, we examined the efficiency of CTL-mediated lysis. We tested the hypothesis that the lysis of infected target cells may depend on the expression level of the viral protein Tax in the target cell. This was based on experimental data that showed target cells expressing a higher level of Tax per cell may be killed quicker by CD8$^+$ cells. The analysis consisted of a series of lytic assays, followed by the development of models of Tax expression dynamics and the rate of killing of target cells by CD8$^+$ cells. This data is shown in Chapter 6.

1.2.5 HLA Class I Alleles and KIR Genotype

One of the main roles of HLA class I molecules is to present antigen to CD8$^+$ T cells. However, HLA-peptide complexes are also recognized by natural killer (NK) cells via their killer cell immunoglobulin (Ig)-like receptors (KIRs). NK cells are critical components of the innate immune system that have direct involvement in the anti-viral immune response. Disease association studies have shown that the interaction between KIR and HLA class I can be protective or detrimental to disease progression in a number of viral infections. In Chapter 7, we tested the hypothesis that KIR-HLA interactions are predictive of disease status in HTLV-I infection.

Chapter 2

Introduction

One of the most important contributions to human health has been vaccination. From the success of Jenners and Pasteurs vaccines against smallpox and chicken cholera, through to the global campaign for the eradication of polio and the widespread immunization against potentially fatal childhood diseases, vaccination has been a vital component of preventative health care. However, the threat of emerging diseases such as avian influenza, as well as current epidemic diseases such as HIV/AIDS, malaria and tuberculosis, has ensured that vaccine development remains a vital component of biomedical research.

One course of vaccination development that has shown recent promise is the identification and utilization of peptide epitopes that stimulate protective immunity. This technique takes advantage of the adaptive immune response to foreign proteins, such as viruses, where pieces of these proteins called epitopes are recognized by the antigen-specific receptors of the immune system (e.g. T-cell receptors, antibodies). Hence, the goal of the vaccine is to safely expose the immune system to pathogenic epitopes to induce an immune response. There are a number of challenges associated with this: identifying what the pathogenic epitopes should be, designing effective delivery of the epitopes when they are found. And increasingly, understanding the complexity of how this works.

The human T-cell lymphotropic virus type 1 (HTLV-I) was the first human retrovirus discovered and its associated diseases: ATL (Adult T-cell leukaemia/lymphoma), HAM/TSP (HTLV-I associated myelopathy/tropical spastic paraparesis) and other chronic inflammatory diseases cause considerable global morbidity and mortality. This chapter gives an overview of the pathogenesis and treatment of the virus and demonstrates the relevancy of my work to understand the basis of an effective immune response towards

HTLV-I infection. Hence, a greater understanding of the targets (epitopes) of HTLV-I specific CD8$^+$ T cells may lead to a vaccine against this widespread debilitating virus.

2.1 HTLV-I

2.1.1 Virology

HTLV-I is a type C particle-like onco-retrovirus, and its discovery was first reported in 1980 when a retrovirus was successfully isolated from a T-lymphoblastoid cell line (HUT 102) established from a patient with a cutaneous T-cell lymphoma [5]. This discovery was the first formal proof that human retroviruses exist and suggested their aetiological role in human cancer, a hypothesis that had been proposed decades before [6]. The diploid genome consists of 2 identical positive single-stranded RNA molecules each of 9032 bases associated in a complex. It contains the typical retroviral genes of *gag*, *pol* and *env*, and in addition genes encoding regulatory proteins such as *tax* and *rex*.

Gag comprises the structural polypeptides p15 (nucleocapsid), p19 (matrix) and p24 (capsid). Env encodes the envelope protein which is cleaved into the surface glycoprotein gp46 (SU) and the transmembrane protein p21 (TM). Pol encodes the genes for reverse transcriptase, integrase and RNaseH. The rest of the genome contains unique accessory genes in four open reading frames (ORFs) of the pX region of the viral genome, as well as a negative strand product, HTLV-I bZIP-factor (Figure 2.1). The regulatory proteins encoded by pX ORFs III and IV, Tax and Rex, respectively, have been extensively characterized. Tax is a *trans*-acting transcriptional activator. It is of central importance in the dynamics of HTLV-I infection as it is thought to be one of the first to be expressed in the viral life cycle and is a promiscuous transcriptional transactivator. It transactivates both its own LTR and many of those of the infected host cell. It is also central to the host's immune response to the virus as it is the dominant target antigen for the CTL response [7–11]. Rex is an essential shuttle protein required for nuclear export of unspliced and incompletely-spliced viral RNAs [12].

Open reading frames I and II are less well known. Both ORFs are alternatively spliced, producing the proteins Rof (p27$^{\mathrm{I}}$) and p12$^{\mathrm{I}}$ for ORF I and Tof (p30$^{\mathrm{II}}$) and p13$^{\mathrm{II}}$ for ORF II. It was thought that ORFs I and II did not significantly affect viral replication. While the expression of mRNAs for these proteins is well-documented *in vitro* and *ex vivo*, their detection in infected cells has remained elusive. However, a body of evidence suggests that these proteins may be essential for viral persistence. p12$^{\mathrm{I}}$ localizes to cellular endomembranes, particularly the ER and expression in virally infected cells could result in decreased expression of MHC class I on the cell surface, thereby protecting infected cells from CTL recognition [13]. Recent findings suggest that p30^{II} functions as a post-transcriptional regulator of Tax/Rex mRNA and may also modulate the expression of viral and cellular genes [13]. In the presence of Tax, p13^{II} is stabilized and localizes to the nucleus. It has been reported that p13^{II} induces Tax degradation and inhibits its transcriptional activity, thereby decreasing viral replication [14].

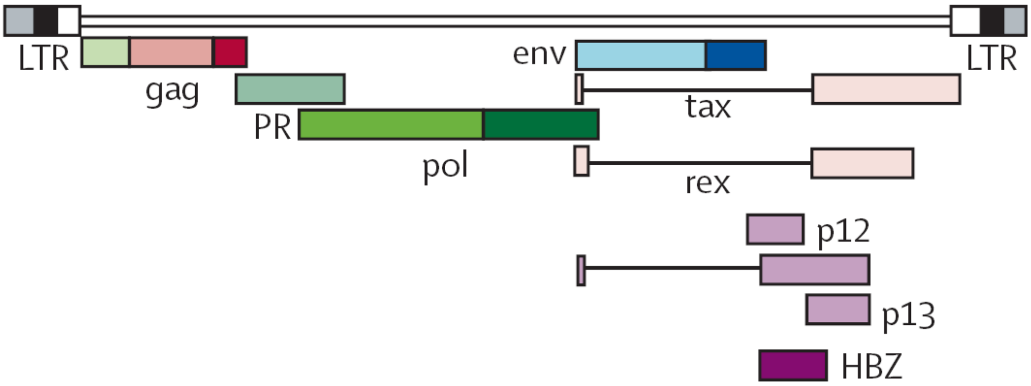


FIGURE 2.1: The genomic organisation of HTLV-I, taken from [15]

It is believed that humans have been exposed to HTLV-I for thousands of years [16–18] and HTLV-I viral DNA has been detected in Andean mummies 1,500 years old [19, 20]. Even though the virus has been in contact with humans for this amount of time, HTLV-I isolates from Japan, Africa, the Caribbean Basin and the Americas show high sequence conservation (0.5 to 4%) [21]. HTLV-I has been classified into three major lineages known as the Cosmopolitan, Central African and Melanesian groups [21–23]. There is further subdivision of the Cosmopolitan group into four subgroups based on LTR sequencing; these are the (A) Transcontinental, (B) Japanese, (C) West African and (D) North African [24, 25]. Generally, a single viral genotype is found in any one location but in Kagoshima, Southern Japan, both Cosmopolitan A (Transcontinental) and B (Japanese) coexist because of its location between Honshu Island (Cosmopolitan B) and Okinawa (Cosmopolitan A) [26, 27].

The virus infects T-cells, with CD4$^+$ and CD45RO$^+$ T-lymphocytes being the main targets for infection [28, 29]. CD8$^+$ T-cells have also been shown to act as as reservoir for the virus *in vivo* [30, 31]. HTLV-I can spread directly between lymphocytes across a specialized, virus-induced cell-cell contact - a 'viral synapse' [32]. The cellular receptor for HTLV-I has not been identified, despite intensive efforts over many years. However, its presence has been demonstrated by virus-induced cell fusion experiments, leading to syncytium formation [33, 34]. It has been mapped to chromosome 17 (17q region) [35] but it is possible that this site encodes not the putative receptor but just an essential cofactor.

2.1.2 Epidemiology

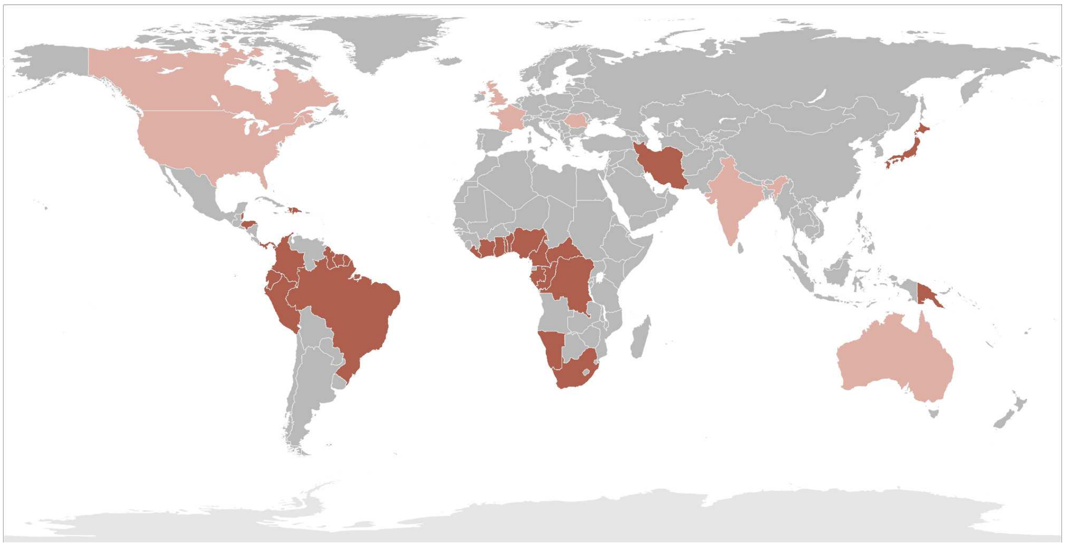


FIGURE 2.2: Countries with endemic HTLV-I, defined as prevalence between 1 and 5% in some populations, are shown in dark red. Countries with reports of low prevalence (less than 1% in some groups), due mainly to immigration from endemic areas, are shown in light red. It should be noted that HTLV-I endemic areas do not correspond exactly to the country boundaries shown in the map. For example in Brazil, Japan and Iran, HTLV-I is limited to residents of certain areas of each country. Data is from [36].

The virus is endemic to a number of geographically distinct regions across the world (Figure 2.2). In the Caribbean, 3-4% of the population are seropositive for HTLV-I, in Africa the virus is detectable along an increasing gradient from north to equatorial Africa and in Japan, several regions have high incidences of seropositive individuals [37]. It can also be found in northern Iran, southern India and the aboriginal peoples of northern Australia. Other populations include immigrants from these endemic areas, as well as sporadic cases of HTLV-I among white Europeans with no identifiable risk factors. Overall, it is estimated to infect between 10 and 20 million people worldwide [6]. It is a chronic infection which remains asymptomatic in the majority of cases. There is a maximum seroprevalance of 35% in Okinawa, Japan [38].

The principal modes of transmission are transfer of infected CD4^{+} lymphocytes from mother to child in breast milk, sexual transmission (especially from infected men to women via semen) and via inoculation/transfusion of infected blood. Cell-free blood products have a negligible risk because of the paucity of free virus particles present in plasma. Infection in endemic regions occurs mainly through breast-feeding, via the transfer of infected lymphocytes in the milk [39]. However, infection can occur during the peri-natal period [40] and transplacental transmission has been documented but is not thought to be common [41]. Male to female transmission is roughly 4 times higher than the converse [42] and the risk of infection is increased in the presence of genital ulceration, high proviral loads and high antibody titres [43, 44]. The probability of seroconversion following transfusion of infected blood products is 50-60% with a median time to conversion of 51 days [45, 46]. Screening transfusion blood for HTLV-I is now routine in Brazil, Japan, the UK and the USA. There is an increasing prevalence of HTLV-I among intravanous drug users in both Europe and the USA [6].

2.2 HTLV-I Associated Diseases

Most HTLV-I-infected people remain healthy, but between 1-2% will develop HAM/TSP, another 2-3% develop ATL and a small number develop other less well defined inflammatory disorders. The factors deciding these outcomes are not fully understood but they are two distinct pathologies and the pathogenesis of the two appear to be very different. It has been thought that the occurrence of the two syndromes in the same person is not seen any more frequently than would be expected by chance. However, a high frequency of co-presentation of HAM/TSP and ATL has recently been reported in Bahia, Brazil [14].

2.2.1 HTLV-I Associated Myelopathy/ Tropical Spastic Paraparesis (HAM/TSP)

The first descriptions of a myelopathy of unknown origin in tropical areas go back to the 19th century [47]. The association with HTLV-I was recognized independently in the Caribbean (as TSP) and in Japan (as HAM) in 1985-1986 [48, 49]. The lifetime risk for developing HAM/TSP is between 2-7%, except for Japan where it is estimated to be 0.25% [50, 51].

2.2.1.1 Pathology

The main pathological feature of HAM/TSP is a chronic inflammation of the white and grey matter of the spinal cord. Mononuclear cells, mainly T cells, cause perivascular cuffing and infiltrate the parenchyma [15]. The damage is concentrated in the white matter of the lower thoracic spinal cord, which causes the spastic paraparesis in the lower limbs [52]. There is a possibility that the lesions in the central nervous system could be the consequence of a genuine anti-HTLV-I reaction. This is based on the observations that HAM/TSP patients have a higher proviral load, a higher production of proinflammatory cytokines in response to viral peptides (such as IFN-γ and TNFα)

and a higher frequency of HTLV-I specific CD8$^+$ T-cells compared to asymptomatic carriers [53–56]. Polymorphism in the TNFα promoter and the chemokine gene SDF-1α have also been shown to influence the risk of HAM/TSP [57]. Other evidence of an immunopathological reaction in the central nervous system is the observation of infected T-cells within the spinal cord lesions and the accumulation of Tax-specific CD8$^+$ T-cells in the cerebrospinal fluid [58, 59]. There is also a possibility that cross-reactivity between HTLV-I antigens and tissue antigens could be involved in the pathogenesis. This is based on a contentious finding that patients with HAM/TSP appear to develop antibodies to human neurons but not to systemic organs [60]. Additionally, auto-antibodies against other nuclear and perinuclear human brain proteins cross-reacting with different HTLV-I epitopes have been found in the serum of HAM/TSP patients [61]. However, since inflammatory T-cells, rather than antibodies, seem to cause the tissue damage, autoreactivity at the level of the T-cell receptor may be more likely.

2.2.1.2 Presentation

The commonest presenting features in descending order are gait disturbance ($\sim 2/3$ cases), urinary dysfunction ($> 1/3$ cases), then numbness of the lower legs, constipation, lumbar back pain and hand tremors. The lower limbs are usually affected to a much greater degree than the upper limbs. The spasticity and associated upper motor signs can be very severe. Low back pain or ache is very common and affects most at some time during the course of the disease. The spectrum of disease progression is very variable, ranging from minimal gait disturbance maintained over the patients lifetime to severe, very rapid progression and even death (rare). In one study cohort from Columbia, after a mean period of 14.4 years (range of 1 - 30 years), 34% could walk unaided, 40% required a walking aid, and 26% required a wheelchair (otherwise bed-bound) [62].

2.2.1.3 Treatment

Therapies targeting the immune response have been considered for the treatment of HAM/TSP. Corticosteroids have been shown to be of some benefit [63] and interferon-β1a reduced HTLV-I mRNA load. However, the proviral load remained unchanged and there was only a slight improvement in motor function [64] The combination of two nucleoside analogues (zidovudine and lamivudine) has been evaluated in a randomised, double-blind, placebo-controlled study including 16 HAM/TSP patients. After up to 12 months of follow-up, there were no significant changes in proviral load and no clinical improvement was observed [65]. Long term treatment studies have been formulated with with valproic acid (VPA). VPA is a lysine deacetylase inhibitor and is postulated to work by activating viral gene expression and exposing virus-infected cells to the immune system, thus reducing proviral load. It has been shown to be safe but does not seem to alleviate the conditions of HAM/TSP [14]. Additional strategies that have been proposed include minocycline (an antibiotic that inhibits monocyte/macrophage activation), humanized mikβ1 (a monoclonal antibody against CD122, the β subunit shared by IL2 and IL15) and the immunosuppressant cisclosporin [14].

2.2.2 Adult T-cell Leukaemia/Lymphoma (ATL)

ATL was first described in the 1970s when the observation of haematological malignancies did not fit previous pattern descriptions [66]. It is a malignancy of CD4$^+$ post-thymic T-cells in which the HTLV-I provirus is integrated.

2.2.2.1 Pathology

The regulatory protein Tax induces abnormal growth of infected T-cells through several pathways [66]. Tax promotes the transcription of its own proviral genome, but it also promotes transcription of cellular genes, including cytokine (e.g. interleukin-2), cytokine receptor (interleukin-2Ra), and anti-apoptotic genes. By binding to other protein complexes, Tax represses the transcription of genes that are important in negative control of the cell cycle, in activation of apoptosis, and in DNA repair. Tax also binds and inhibits proteins directly involved in tumour suppression and DNA repair. Finally, Tax causes cells to bypass normal cell-cycle checkpoints [66]. The net effect of all these activities of Tax is that T cells are rushed into and through the mitotic phase without checking for chromosomal abnormalities. Genetic damage that would normally be repaired accumulates and apoptotic cell death does not occur even in cells with severely damaged DNA. In these circumstances, T cells can accumulate DNA mutations, resulting in transformation and monoclonal outgrowth of a truly malignant cell. In addition to these genetic changes, epigenetic changes such as DNA methylation may have an important role in leukaemogenesis [67].

2.2.2.2 Presentation

There are several types of HTLV-I induced ATL: acute, lymphotamous, chronic and smouldering [68]. Amost all patients with ATL present with lymphadenopathy (enlargement of the lymph nodes) and/or splenomegaly (enlargement of the spleen). ATL can also affect the lungs, gastrointestinal tract, and central nervous system; involvement of other organs is uncommon [68]. Hypercalcaemia is an important complication: it occurs in up to 70% of patients and is often accompanied by lytic bone lesions. ATL patients are immunosuppressed and opportunistic infections, such as *Pneumocystis jirovecii* pneumonia, cryptococcus meningitis, and disseminated herpes zoster are, therefore, frequent [69]. Liver dysfunction is another complication. The diagnosis of ATL is usually based on morphological analysis. Flower cells (i.e. pleomorphic, atypical lymphoid cells with basophilic cytoplasm and convoluted nuclei) are indicators of acute or lymphomatype ATL. This must be confirmed by clonal integration of HTLV-I provirus in the host genome.

2.2.2.3 Treatment

Acute ATL is very aggressive and highly refractory to treatment. Strategies that show an improvement over conventional chemotherapy in the treatment of ATL include Interferon-α with zidovudine, intensive chemotherapy and allogenic haematopoietic stem cell transplantation [67]. In fact, it is essential not to provide general chemotherapy (CHOP) to first line presenting ATL patients because this treatment selects for a tumor clone with mutated p53 [70]. Nevertheless, the median survival of patients with acute, lymphomatous, and progressing chronic ATL remains low: less than 1 year in most reports [67]. Further improvements could include bortezomib (a proteasome inhibitor), anti-CD52 antibody, proapoptotic agents and consolidation with arsenic and IFNα [14].

2.2.3 Other Conditions Associated with HTLV-I

HTLV-I has been associated with other inflammatory syndromes. In a Japanese cross-sectional study and a US cohort study, the prevalence and the incidence of arthritis were found to be higher among HTLV-I-infected patients than among uninfected individuals [71, 72]. Tax transgenic mice have also developed an arthritis that is pathologically similar to human rheumatoid arthritis [73, 74].Tax has been shown to stimulate the proliferation of synovial cells in vitro [75]. Hence, Tax, released by HTLV-I-infected cells in vivo, could have a part in the pathogenesis of arthropathy.

Reports from Japan have shown that HTLV-I infection is more frequent in patients with uveitis of unknown origin than in the general population [76].The prognosis of HTLV-I-associated uveitis is good: spontaneously, the disease resolves within weeks and recovery is even faster with topical or systemic corticosteroid treatment. However, more than 90% of cases recur within 3 years; the mean interval between episodes is 16 months [77].

Strongyloides stercoralis is an intestinal nematode of tropical regions that can replicate within the human host, an unusual characteristic among helminths. HTLV-I infection is associated with increased susceptibility to *S. stercoralis* infection and a weak Th2 response is characteristic of this co-infection. As a result, the rate of parasite killing decreases and the rate of autoinfection increases [78]. In rare cases this can result in the fatal Strongyloides hyperinfection syndrome [79].

2.3 Pathogenesis of HTLV-I

2.3.1 Background

1-2% of HTLV-I infected subjects develop HAM/TSP and 2-3% develop ATL. The factors deciding these outcomes of infection are not understood.

2.3.2 Genotype

Compared with HIV, HTLV-I is relatively stable in terms of sequence variation and mutation rate. However, the effect of mutation on the pathology of HTLV-I is still considered a possible variant in disease outcome. As a result, HTLV-I has been the subject of a range of studies looking at mutation and variability and how this affects the immune response to the virus. The majority of these studies have focused on *tax*, as it is a dominant target for the CD8$^+$ immune response [11]. Furukawa *et al.* [27] found phylogenetic subgroups in the *tax* gene, one of which was associated with an increased risk of HAM/TSP. This result followed on from a number of studies from Niewiesk *et al.* that focused on Tax expression.

Initially, Niewiesk *et al.* found that general *tax* sequence variability (and not the presence of a specific sequence) was significantly greater in healthy seropositive individuals, compared to those presenting HAM/TSP [80]. This was followed by results showing that amino acid substitutions occurring in known Tax epitopes abolished T cell recognition. These substitutions were also associated with the allele *HLA-A2* and reduced the transactivation function of Tax [81]. However, it was then found that this distinction in the mutation rate of *tax* between healthy individuals and those with HAM/TSP could only be seen with proviral *tax* sequences, but not with cDNA [82]. Kubota *et al.* looked at synonymous and nonsymonymous *tax* mutations in *HLA-A*02* HAM/TSP patients to detect positive selection pressures [83]. They found pressures on three of six CTL epitopes tested, suggesting that CTLs eliminate infected cells *in vivo* and also demonstrating that variant viruses do not accumulate. Once again, this reinforced the observation that Tax is functionally constrained in terms of mutations. Although research on Tax has predominated, some work on other proteins has also been completed.

Furukawa *et al.* showed that sequence variation in p12 may be associated with different outcomes to HTLV-I infection [84]. The Rex protein was also examined and shown to have strong functional constraints on amino acid variation [85].

2.3.3 The CTL Response to HTLV-I

The CTL response plays a central role in deciding the outcome of viral infections. It has been shown through evidence accrued from host and viral genetics, gene expression microarrays and assays of T-cell phenotype and function that individual differences in the efficiency of the virus-specific CTL response strongly determine the outcome of infection with the human retroviruses HTLV-I and HIV-I. From this evidence, it is now believed that differences in the anti-viral CTL efficiency at the single-cell level are responsible for variation in the efficacy of the host response to viruses.

Perhaps the strongest evidence that the CTL response is instrumental in controlling HTLV-I infection comes from the association of certain MHC class I alleles and protection from disease. Studies of HTLV-I genotype show significant associations between class I alleles (HLA-A*02 and HLA-Cw*08) and a reduced proviral load, which would implicate the CTL response as a positive factor[1, 4]. The hypothesis that extends from these results is that HLA-A*02 and HLA-Cw*08 restricted CTLs are more efficient at killing HTLV-I infected cells. Conversely, HLA-B*54 restricted CTLs, which have been associated with increased proviral load [1], would produce a less efficient response.

The frequency of virus-specific CTL has been used to demonstrate the efficiency of the CTL response in HTLV-I infection with different conclusions. There is evidence that the frequency of HTLV-I-specific CD8$^+$ T cells differs little among patients with widely differing proviral load [10]. However it has also been reported that HTLV-I-specific CTL frequency was positively correlated with proviral load [86]. These contradictory results demonstrate the difficulty of using CTL frequency as a marker of viral control in a chronic infection: since CTL proliferate in response to antigen, the frequency of CTL is both a cause and an effect of the viral load. Hence, a more effective metric of CTL efficiency is a measurement variable that accurately reflects an efficient CTL response.

CD8$^+$ cell attributes such as T-cell receptor avidity, specificity, and cell maturation state may all affect the ability of CD8$^+$ cells to control a viral infection. However, Asquith *et al.* devised a combined measure by measuring the rate at which naturally, endogenously infected cells were cleared by autologous CD8$^+$ cells *ex vivo* [87]. This antiviral efficacy can be summarised by Equation 2.1:

$$\frac{dy}{dt} = c - \epsilon yz \tag{2.1}$$

where y is the proportion of CD4$^+$ cells expressing Tax, c is the rate of increase of Tax expression, ϵ is the CD8$^+$ cell-mediated antiviral efficacy and z the proportion of lymphocytes that are CD8$^+$. This approach yielded the conclusions that there was a significant negative correlation between the per-CD8$^+$ cell lysis rate and the proviral load, in both ACs and HAM/TSP patients. Also, the percentage of between-individual variation observed in the proviral load that was attributable to variation in the lysis rate parameter was about 35% [87]. From this data, it was predicted that CTL lysis would reduce the life expectancy of a virus-expressing target cell from the normal 30 days (for a memory phenotype, CD4$^+$CD45RO$^+$ T cell) to between 1 and 10 days. This was confirmed by measurement of infected T-cell turnover rate *in vivo* by the metabolic labeling of lymphocytes with deuterated glucose [88].

The functional avidity is the concentration of antigen that is required to elicit the half-maximal effector response (usually cytokine) in CD8$^+$ T cells. It has been widely used as a marker of the responsiveness or sensitivity of CTL to cognate antigen. In HTLV-1 infection, Kattan *et al.* [89] found that avidity was correlated with per-CD8$^+$ lytic activity, measured by the CD8-dependent elimination of Tax$^+$ cells. The use of CD107a staining of CD8$^+$ T cells as a marker of the recent degranulation activity of HTLV-I-specific CD8$^+$ T cells has also shown differences between HAM/TSP patients and ACs; HAM/TSP patients produced a greater frequency of specific CD8$^+$ T cells but less CD107a staining per cell than ACs [90].

Taken together, this data emphasizes the role of CTL in HTLV-I control. It is clear that the HTLV-I-specific CTL response plays a critical role in limiting the replication of HTLV-I, the proviral load, and the risk of HAM/TSP. However, any understanding of the role of the CTL response in HTLV-I infection must acknowledge the seemingly detrimental effects of this response to the host (Section 2.2.1.1).

2.3.4 The Antibody Response to HTLV-I

Anti-Gag antibodies are the first specific antibodies to appear in response to the infection in the first 2-3 months. Anti-Env antibodies can then be detected, along with anti-Tax antibodies in 50% of infected people [46, 91]. Anti-HTLV-I antibody titres correlate with the provirus load and can be extremely high. It is currently unknown whether these antibodies play a part in protection against HTLV-I infection, against disease or are involved in the pathogenesis of disease.

Levin *et al.* [60] have described a putative autoantigen; neuronal heterogeneous nuclear ribonuclear protein-A1 (hn-RNP-A1), which stained brightly with IgG from HAM/TSP patients and not ACs. These IgG were also found to cross-react with HTLV-I Tax protein and stained human Betz cells specifically. Furthermore, the authors infused these antibodies onto rat brains and showed inhibition of neuronal activity. Thus, they concluded that HAM/TSP is an autoimmune disease, with molecular mimicry between an HTLV-I antigen and a self one causing the generation of cross-reacting antibodies and subsequent neurological disease.

2.3.5 Other Immune Responses to HTLV-I

T_{reg} cells are defined as $CD4^+$ T cells that inhibit immunopathology or autoimmune disease *in vivo* [92]. The subset that has been studied with respect to HTLV-I is characterized by the expression of the glycoproteins CD4 and CD25, as well as the transcription factor Foxp3 [92]. Their role in HTLV-I infection is not yet fully understood but different T_{reg} responses have been associated with both ATL and HAM/TSP. In terms of HAM/TSP, a number of studies have found that in T_{reg} cells infected with the virus, both mRNA and protein expression of Foxp3 were lower in HAM/TSP patients compared to healthy carriers [93, 94]. This has lead to the hypothesis that defects in T_{reg} expression as a result of viral infection could cause the chronic inflammatory response characteristic of HAM/TSP [95]. However, this conclusion remains very uncertain because HTLV-I strongly induces expression of CD25 (a marker of T_{reg} cells) and it is therefore inappropriate to use $CD25^+$ as part of the definition of T_{reg} cells in HTLV-I infection (C. Bangham, pers. comm.).

In terms of ATL, several studies have shown the expression of Foxp3 in the tumour cells of a subset of patients with ATL [96]. Yano *et al.* demonstrated these cells continue to act as regulatory T cells and that their proliferation may be the cause of the severely immunocompromised state of ATL patients [96].

The natural killer (NK) cell response to HTLV-I has received less attention, partly because of the difficulty in identifying NK cells in terms of their surface markers and the existence of NK cell subsets [3]. However, an association has been found between the low frequency of $CD3^+$ NK cells and patients with HAM/TSP [97–99], suggesting a role for this subset of NK cells in disease progression. Other data on lymphocyte gene expression also indicated that high levels of expression of certain genes involved in NK cell-mediated lysis were associated with low proviral load of HTLV-I [100]. This would suggest that, along with $CD8^+$ CTLs, NK cells are part of the cytolytic lymphocytes that reduce HTLV-I proviral load.

The CD4$^+$ (helper) T cell response has been difficult to study as Tax in infected CD4$^+$ T cells produces effects (IFNγ production, T-cell proliferation), which are also the basis of CD4$^+$ T cell response assays [3]. Hence, the presence of HTLV-I would interfere with any analysis on CD4$^+$ T cell response. However, using a modified assay [101], it has been demonstrated that the response is predominantly IFNγ-producing cells. Also the frequency of IFNγ producing CD4$^+$ T cells was between 10 and 25 times greater in HAM/TSP patients compared with asymptomatic carriers [55]. From this information, it is likely that these cells contribute to the chronic inflammatory response seen in HAM/TSP.

2.4 Epitope Prediction

As mentioned in Chapter 1, Section 1.2.2 it was necessary to use epitope prediction software to predict the HTLV-I peptides that bind to different MHC class I alleles. This type of prediction software uses a range of mathematical methods to recognize the small number of pathogenic peptides that can bind to MHC class I and hence elicit a CTL immune response.

Of the large number of peptides that can be derived from a pathogen only a small minority, approximately 1 in 2,000, elicits a CTL response [102]. This limitation in the number of peptides that are immunogenic is conferred by three main constraints: the requirement for peptide cleavage and transport, the requirement for MHC-peptide binding and the requirement for CTL recognition. By far the most stringent of these is the requirement for MHC-peptide binding, because only 1 in 200 peptides binds a specific MHC molecule with sufficient affinity to elicit an immune response [102]. Further selection is largely due to the limitations of peptide processing and transport. In these processes, individual peptides are produced from the precursor polypeptides by proteasomal cleavage of the polypeptide, which can be followed by N-terminal trimming by other peptidases. This is followed by the transport of the peptides from the cytosol to the endoplasmic reticulum, mediated by the TAP complex. Further N-terminal trimming may occur before the peptide binds to the MHC molecule. The requirements of processing and transport eliminate approximately 80% of potential epitopes [102]. Finally, T cell specificity, i.e. the requirement for T cell receptor binding of the MHC-peptide complex, further halves the number of presented peptides that elicit a response. The probability of each of these steps is determined by the polypeptide sequence, amongst other factors [103].

The identification of T cell epitopes is of vital importance in the design of vaccines and understanding of the immune system [104–107]. However, given the scarcity of epitopes, experimentally screening all possible peptides for each MHC allele (e.g. by

IFNγ ELISpot) is time consuming, expensive and inefficient. One way to improve the efficiency of the identification process is to first use theoretical algorithms to predict which peptides are more likely to be epitopes and then experimentally screen this much smaller, selected list of peptides. This method is widely used [108–112] and has been applied in a number of studies to identify potential vaccines [113, 114]. The use of theoretical methods to ”pre-screen” peptides is of particular importance in the case of emerging infections such as avian influenza [115] where rapid vaccine development would be vital. This approach also underpins a large bio-preparedness initiative coordinated by the Large-Scale Antibody and T Cell Epitope Discovery Program [105], which intends to foster development of immune-based therapeutics for emerging and reemerging pathogens including potential bioterrorism agents. More generally, epitope prediction algorithms are being increasingly used to understand the CTL response. For example, in the case of HIV-I infection, algorithms have been used to confirm which epitope mutations are likely to confer escape from a CTL response [116] and to understand why some MHC class I alleles are associated with slow rates of disease progression [117].

A range of computational algorithms have been developed to predict CTL epitopes in pathogen protein sequences. Since the most selective requirement for a peptide to be immunogenic is the ability of the peptide to bind to the MHC molecule, most prediction methods focus on this stage of the pathway. As a general rule, information gained from experimental binding assays is used to train the algorithm until it is efficient at predicting novel MHC-peptide complexes. The algorithms that are used vary in complexity and accuracy. Some can be trained to recognize peptide motifs that are required for binding to a particular MHC molecule [118], others use a weight-matrix method to identify amino acids that occur at a higher-than-expected frequency at specific epitope positions [119–121]. However, the most accurate methods available use logistic regression [122] and, more generally, artificial neural networks [103, 123].

Artificial neural networks (ANNs) take into account, in addition to the identity of each amino acid residue, the interactions between adjacent amino acids in a potential epitope. In summary, an ANN for a particular MHC molecule is trained to recognize associated inputs (a peptide sequence) and outputs (the binding affinity for that sequence with the MHC molecule) [124]. Once an ANN is trained for a particular molecule, it can predict the binding affinity of novel peptide sequences.

NetCTL [103] and NetMHC [119, 124, 125] are two of the most accurate prediction methods currently available [126]. NetMHC uses ANNs for a number of alleles to predict MHC molecule-peptide binding affinities. NetCTL, as well as using the same ANNs to predict MHC-peptide binding, also utilizes information about the proteasomal cleavage of the input peptide sequence, and its ability to bind to TAP. NetCTL or NetMHC will predict a score (either integrated or simply a binding affinity, respectively) for every overlapping nanomer peptide sequence in an input sequence to each MHC molecule for which the method has an ANN. Henceforth, we refer to the trained prediction algorithm for each MHC class I allele as an ”allelic predictor”.

Chapter 3

Identifying Alleles Associated with Disease Status and Proviral Load

3.1 Introduction

Jeffery *et al.* [1, 4] demonstrated the protective effects of the MHC class I alleles A*02 and Cw*08 in terms of disease status and a reduced proviral load in asymptomatic carriers of HTLV-I. It was also shown that B*5401 is associated with a greater risk of HAM/TSP and with a higher proviral load in HAM/TSP patients. Added to this, results showing that class I heterozygosity is associated with significantly lower proviral loads [4] would suggest that the protective effect of the HLA haplotype extends to a range of alleles.

Using the same Kagoshima database as previous studies [1, 4], our initial task was to reanalyze this data in an attempt to broadly classify the HLA class I alleles contained within the cohort into 'detrimental', 'beneficial' and 'undefined' groups, according to their associations with disease risk and proviral load. This design would increase the power of the next stage of our study - understanding the functional basis of these associations though analysis of the HLA class I alleles' HTLV-I epitopes.

Hence, the aim of this chapter was to analyze the HLA class I repertoire of the Kagoshima cohort in terms of its association with proviral load and disease risk, using classical and novel nonparametric statistical methods.

3.2 Methods

3.2.1 The Kagoshima Dataset

All HAM/TSP and AC subjects for this study were of Japanese ethnic origin, and resided in Kagoshima prefecture (1988 population: 1.7 million), southern Kyushu, Japan, where the seroprevalence of HTLV-I infection in adults is approximately 10% [49, 127]. The estimated prevalence of HAM/TSP in the HTLV-I positive population is less than 1% [50]. For the purposes of this study, 230 cases of HAM/TSP were compared with 202 randomly selected HTLV-I seropositive asymptomatic blood donors (asymptomatic carriers - ACs) from the Kagoshima Red Cross Blood Transfusion Service. The diagnosis of HAM/TSP was made according to World Health Organisation (WHO) criteria [64].

3.2.2 Disease Risk

The Yates χ^2 test has been used to test the relationship between disease risk and the presence of an allele. The test takes as its input a matrix (table 3.1) and examines the null hypothesis that the observed frequency of alleles in each population (HAM/TSP and asymptomatic carriers) is the same as the expected frequency. The Yates correction applied in each case is used to prevent overestimation of statistical significance for small amounts of data.

	A$^+$	**A$^-$**
D	a	b
H	c	d

Table 3.1: The input matrix for the χ^2 test, where D = disease, H = health, A$^+$ = positive for protective allele and A$^-$ = negative for protective allele.

3.2.3 Proviral Load

We used the Mann-Whitney U test to examine the null hypothesis that the presence of a single allele had any effect on proviral load. This analysis was performed separately for HAM/TSP patients and asymptomatic carriers (ACs) as there is a very strong association between HAM/TSP and high proviral load.

A novel ranking test was also formulated to examine the robustness of each allele's association with proviral load. For both groups (HAM/TSP and ACs), the following was performed:

- The individuals in each group were randomly assigned to two separate populations.

- For each of the two populations, the alleles were ranked in terms of the median proviral load associated with that allele (i.e. the median proviral load of all individuals who possessed that specific allele). This random assignment was reiterated 1,000 times.
- The result of this was 2,000 rank positions for each allele in terms of median proviral load.
- The median rank and confidence intervals were then compared.

3.3 Results

3.3.1 Proviral Load

Figure 3.1 shows the results of the initial analysis of multiple Mann-Whitney *U* tests for each allele.

Figure 3.3 shows the results of the robustness of rank measure in terms of proviral load. Any allele showing a narrow confidence interval is demonstrating a robust rank in the face of random sampling from the proviral load associated with it. For example, in the HAM/TSP results, we can be confident in the designation of HLA-A*03 as a 'good allele' in terms of proviral load, owing to its position on the x-axis and the narrowness of the confidence interval. From this data, alleles were designated as positive or negative according to the non-overlapping of their confidence intervals (Table 3.2).

3.3.2 Disease Risk

Figure 3.2 shows the alleles ranked in terms of disease risk. These results show an obvious overlap with previous research (the significant results of A*02, B*54 and Cw*08) and the possibility of other candidate alleles (B*48).

3.4 Discussion

Previous studies have clearly demonstrated the protective effect of *Cw*08* and *A*02* in terms of proviral load in aymptomatic carriers and disease risk [1, 4]. This would suggest a protective effect of a strong CTL response. The finding that heterozygosity also resulted in a significantly reduced proviral load suggested the presence of other protective alleles. We reanalysed the Kagoshima Cohort to look for any other allele

Names	P HAM/TSP	HAM/TSP Effect	P AC	AC Effect	P χ^2	χ^2 Effect	HAM/TSP Rank	AC Rank
A01	0.0000	NA	0.1786	protective	0.0000	NA		protective
A02	0.3003	protective	0.0161	protective	< 0.0001	protective		
A03	0.1464	protective	0.7909	protective	0.5261	protective	protective	
A11	0.3999	protective	0.9211	protective	0.3635	detrimental		
A24	0.4787	detrimental	0.3681	detrimental	0.1474	detrimental		
A26	0.5197	protective	0.0437	detrimental	0.8559	detrimental		detrimental
A30	0.0568	protective	0.0000	NA	0.0000	NA	protective	
A31	0.7096	detrimental	0.9037	protective	0.2681	detrimental		
A32	0.2803	protective	0.1786	protective	0.5365	protective	protective	protective
A33	0.1514	detrimental	0.3850	protective	0.2995	detrimental	detrimental	
B07	0.0230	protective	0.0963	detrimental	0.1035	detrimental		
B13	0.7524	protective	0.5186	detrimental	0.4508	detrimental		
B15	0.2597	protective	0.2601	protective	0.1281	protective		
B27	0.5412	detrimental	0.4263	protective	0.8839	protective	detrimental	
B35	0.1039	protective	0.4310	detrimental	0.5841	protective		
B37	0.0000	NA	0.1786	protective	0.0000	NA		protective
B38	0.1893	protective	0.0000	NA	0.0000	NA	protective	
B39	0.3045	detrimental	0.7824	detrimental	0.6011	protective		
B40	0.4786	detrimental	0.5709	protective	0.1104	protective		
B41	0.0000	NA	0.7562	detrimental	0.0000	NA		detrimental
B44	0.2169	protective	0.0733	protective	0.4486	protective		
B46	0.9813	detrimental	0.7261	protective	0.4279	protective		
B48	0.7705	protective	0.4001	protective	0.0263	protective		
B51	0.3067	detrimental	0.7686	detrimental	0.3218	detrimental		
B52	0.6364	detrimental	0.3742	detrimental	0.5040	protective		
B54	0.0034	detrimental	0.8505	detrimental	0.0008	detrimental	detrimental	
B55	0.6298	detrimental	0.3022	protective	0.5622	detrimental		
B56	0.1452	detrimental	0.4664	protective	0.7439	detrimental	detrimental	
B57	0.0000	NA	0.7432	detrimental	0.0000	NA		detrimental
B58	0.3457	detrimental	0.8068	detrimental	0.3748	detrimental		
B59	0.9265	detrimental	0.9980	protective	0.5947	protective		
B67	0.9677	protective	0.6816	detrimental	0.8012	protective		
C01	0.0716	detrimental	0.2828	protective	0.1416	detrimental	detrimental	
C03	0.1176	protective	0.4690	detrimental	0.1659	protective		
C04	0.7755	detrimental	0.4697	detrimental	0.8566	detrimental		
C05	0.2803	protective	0.0438	protective	0.5261	protective	protective	protective
C06	0.0161	protective	0.1786	protective	0.7092	detrimental	protective	protective
C07	0.3218	protective	0.1859	detrimental	0.2001	detrimental		detrimental
C08	0.2647	protective	0.0466	protective	0.0029	protective		
C12	0.8432	protective	0.6302	detrimental	0.9447	protective		
C14	0.2828	detrimental	0.9638	protective	0.2323	detrimental		
C15	0.0170	protective	0.9644	detrimental	0.1904	protective	protective	

TABLE 3.2: The summary of allele association statistics. The first 4 columns give the P values and the direction of effect for the Mann-Whitney U tests of proviral load in both HAM/TSP patients and ACs for alleles in the Kagoshima cohort (Figure 3.1). For each allele, the proviral load of individuals with and without the allele was compared. The next 2 columns show the significance and direction of disease risk associated with the allele in question (Figure 3.2). The last 2 columns show the significant results of the robustness of rank measure (Figure 3.3). For both HAM/TSP patients and ACs, alleles are described as 'positive' if their upper confidence limit does not overlap with the lower confidence limit of the 'detrimental' alleles (and conversely for the detrimental alleles).

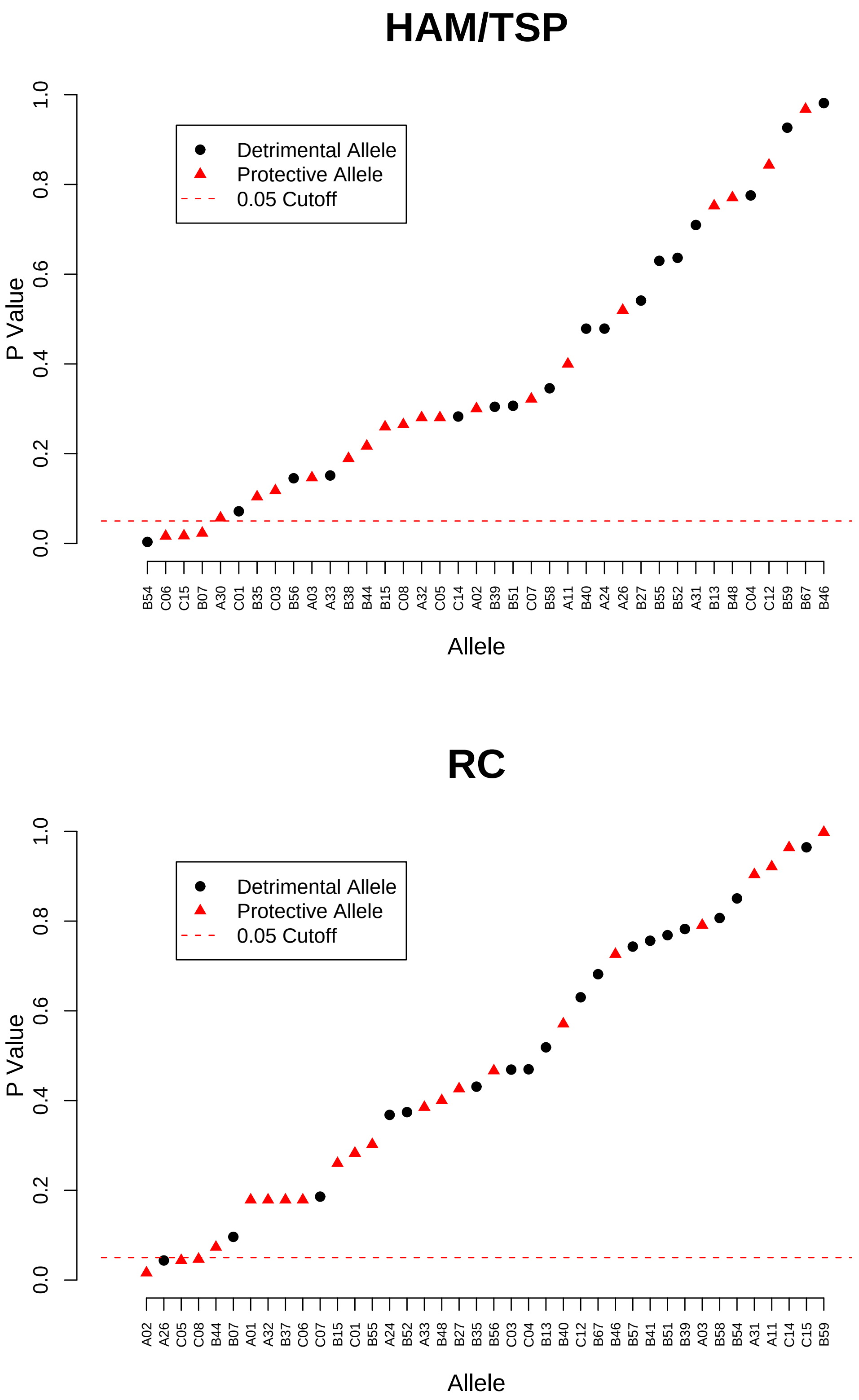


FIGURE 3.1: The allele population of the HAM/TSP and AC individuals ranked by their Mann-Whitney *U*-test *P* values, as described in Section 3.2.3. The black circle indicates that the association is negative (the presence of the allele is associated with a higher proviral load) and the red triangle indicates a positive association.

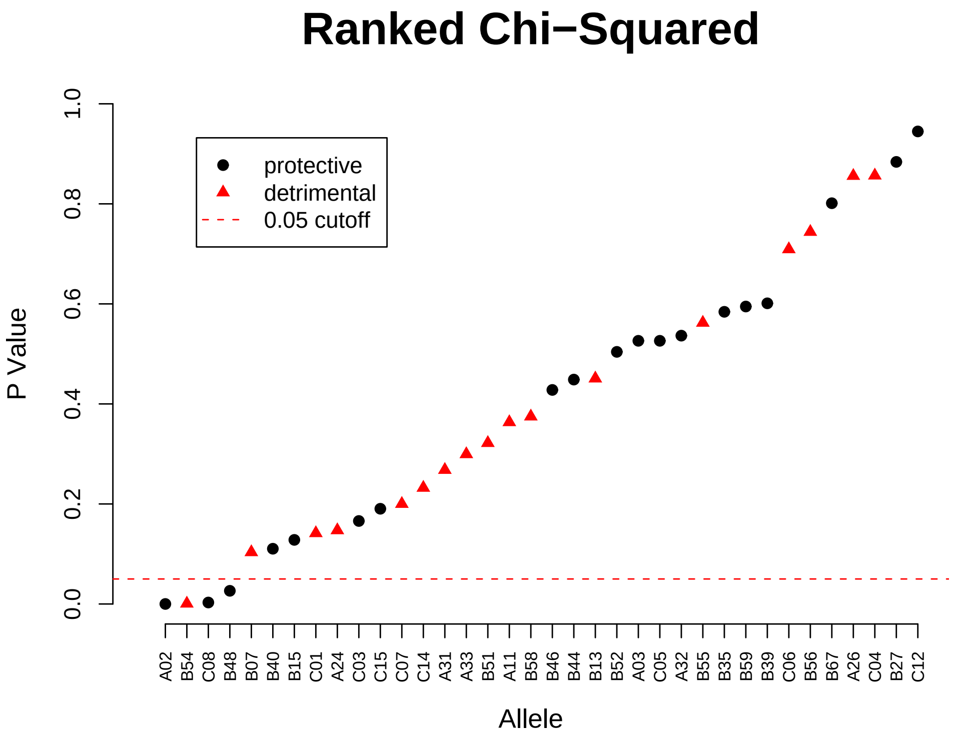
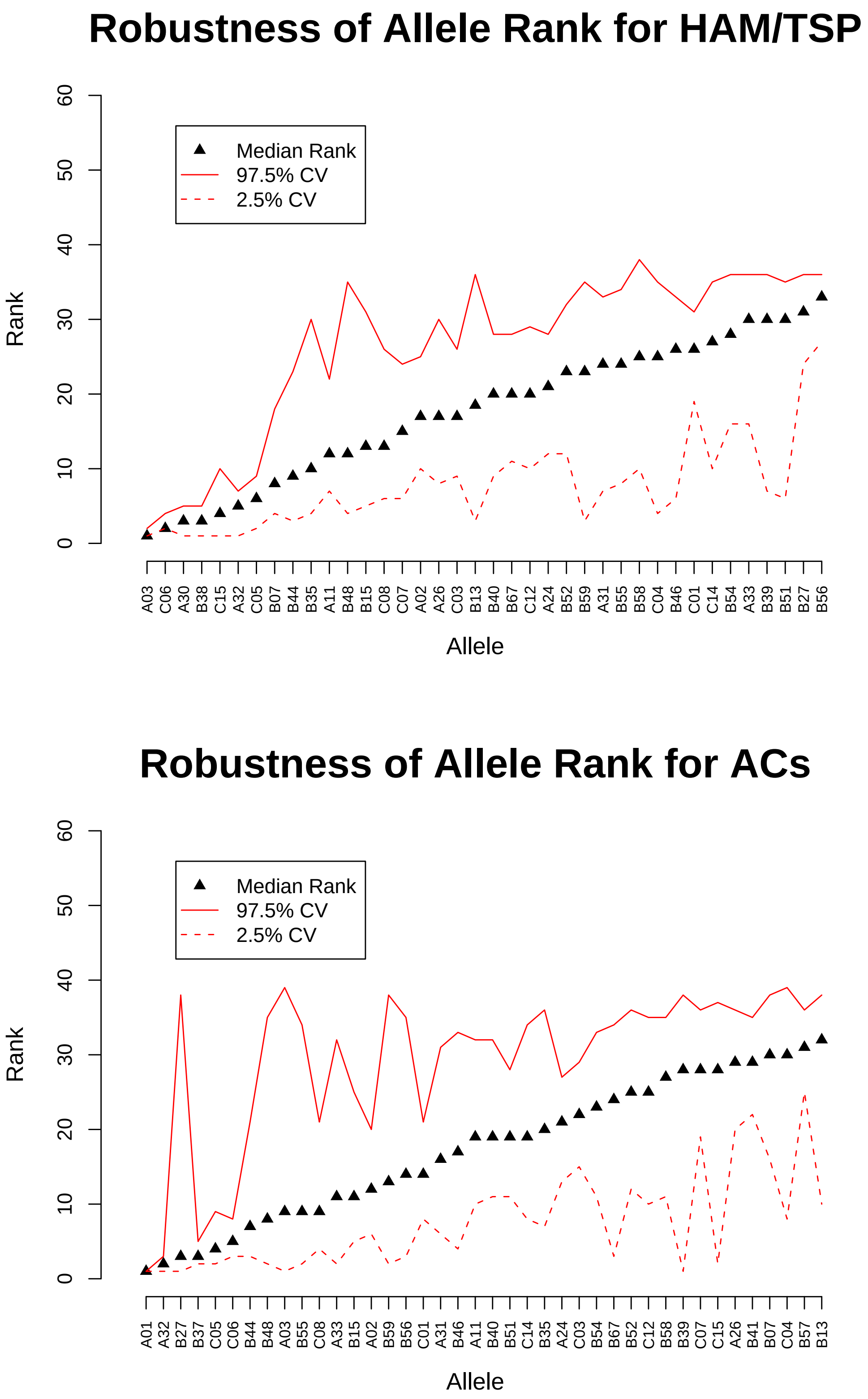


FIGURE 3.2: The result of the Yates χ^2 analysis of disease risk for all alleles in the Kagoshima population. The black circles indicate a protective effect of an allele, the red trangle a detrimental effect. The red dotted line represents significance at $P = 0.05$.

effects with the goal of assembling the alleles of the cohort into larger sets of protective and detrimental alleles in order to give our epitope analysis greater power.

Table 3.2 gives the summary of results for the tests of association between each allele in the Kagoshima Cohort against disease risk and proviral load. Using this combination of tests, we looked for any alleles that were significantly associated with either detrimental or protective outcomes, excluding those already known. From this analysis, no new alleles could be significantly linked with either disease status or proviral load. There are multiple reasons why this could happen: certain alleles may not have been represented at a high enough frequency in the Kagoshima cohort. More specifically, each individual expresses 6 co-dominant MHC class I alleles, which makes it a statistically difficult task to understand each allele's effect at the individual level. For example, the effect of possessing both detrimental and beneficial alleles may be additive or dominant in either direction. We decided at this point to proceed with our definition of beneficial and detrimental alleles limited to *Cw*08/A*02* and *B*54* respectively. By using epitope prediction software each allele could be clearly defined in terms of their epitope properties - a definition that takes into account funtionality and would be a more nuanced method



Robustness of Allele Rank for HAM/TSP

Robustness of Allele Rank for ACs

FIGURE 3.3: The allele rankings and robustness, as described in Section 3.2.3. Each point gives the alleles median rank over 2,000 iterations. The area between the solid dot and dotted lines gives the 95% confidence interval of the value

of separating alleles compared to their association with disease risk and proviral load. This analysis follows in Chapter 4 and Chapter 5.

Chapter 4

Rescaling in T-cell Epitope Prediction

4.1 Introduction

An ideal method to test hypotheses about the protective effect of MHC Class I alleles against disease risk and proviral load in HTLV-I would be to establish experimentally the HTLV-I peptides that bind to these protective and detrimental alleles. Unfortunately, very few MHC Class I epitopes have been experimentally confirmed for HTLV-I, unlike, for example, HIV [128]. Given the scarcity of available epitope information, it was necessary to use epitope-prediction software to predict the HTLV-I peptides that binded to the alleles of the Kagoshima Cohort. Before beginning the analaysis of HTLV-I, it was necessary to test the accuracy of these prediction methods outlined in Section 2.4, namely NetMHC v3.0 and NetCTL v1.2. During the course of this initial testing, our attention was drawn to a normalisation procedure - rescaling - that is used to compare the predicted binding affinities across different allleles. From this data, we wanted to test the hypothesis that rescaling predicted binding affinities results in a loss of allele-specific information and ultimately produces less accurate results in defining the epitope repertoire of HTLV-I.

In order to make the prediction values comparable between each MHC molecule, it is recommended that the MHC-peptide binding affinity scores are rescaled [129]; this is explicitly implemented in NetCTL. The method of rescaling involves obtaining the predicted binding affinities of 500,000 random natural peptides for each MHC allelic predictor. From these affinities, a rescale value is calculated, defined as the binding affinity that is the threshold for the top 1% of total binding affinities. The rescaled affinity is then defined as the predicted affinity score divided by this rescale value [103].

Hence, from this calculation, all alleles are predicted to bind the same number of high-affinity peptides. One pragmatic reason for rescaling is to correct for any discrepancies between the allelic predictors that resulted from inconsistent training data (e.g. data that came from different sources), by assuming that all alleles should bind the same number of epitopes (C. Keşmir, pers. comm.). Additionally, there are biological arguments for believing that different alleles should bind similar numbers of epitopes. It has been postulated that the opposing constraints of effective pathogen recognition but tolerance of self would result in a very narrow range of optimal promiscuity for viable MHC class I molecules. A narrow range of promiscuity would also be predicted as a direct outcome of effective tapasin-dependent peptide optimization in the endoplasmic reticulum [130–132].

However, we will present evidence in this chapter that in correcting for differences between the allelic predictors, information is being lost that reflects true biological variation between MHC molecules and, by extension, differences in their ability to bind to peptide sequences. We show that, for both qualitative and quantitative measures of binding, rescaling impairs rather than improves allelic predictor performance. This is of importance for vaccine design and to understand the nature of the CTL response. In particular, crucial between-allele variations in binding affinity and preference which may contribute to differences in the outcome of infection are likely to be obscured by rescaling.

4.2 Methods

4.2.1 Prediction Method Outputs

In order to test the effect of rescaling on epitope prediction accuracy, we used two web-based prediction methods, NetCTL v1.2 [103] and NetMHC v3.0 [119, 124, 125]. NetCTL is an integrated method that uses information pertaining to TAP and protein cleavage in its predictions, together with MHC binding. The output is combined by rescaling the MHC binding result and adding this to the weighted scores for TAP and protein cleavage. NetCTL has allelic predictors for 12 different class I alleles that are chosen to be representative of each of 12 supertypes; hence it has 12 different rescaling factors.

NetMHC v3.0 simply predicts MHC-peptide binding, using ANNs to predict binding affinities for 43 MHC molecules. In order to test the effect of rescaling, it was necessary to produce rescale values for each of the 43 allelic predictors. This was performed as in NetCTL; 500,000 unique random nonamers were obtained from the proteome of

Mycobacterium tuberculosis, their binding affinity was predicted and the rescale value (top percentile) was found for each allelic predictor. We also performed this calculation with 500,000 random natural peptides to test for the possibility of error from bias in amino acid usage in *Mycobacterium tuberculosis*. There was no significant difference in the rescale values obtained using these two different sources (Figure 4.2).

In summary, we tested two sets of rescaling values: those obtained from NetCTL v1.2 and those that we calculated using NetMHC v3.0.

4.2.2 Datasets

Epitope datasets were constructed from sources detailed below. In each case, the prediction methods were tested by their ability to detect these epitopes amongst the full set of overlapping nonamers derived from the proteins that contained the epitopes. The full set of nonamers will contain a small number of known epitopes and the remainder will be ‘non-epitopes’. Of course, this set of non-epitopes could include epitopes that have not been experimentally verified. However, the majority (see Section 4.1) would be non-binders with the corresponding MHC molecule. Added to this, the labelling of epitopes as ‘non-epitopes’ impact on both rescaled and non-rescaled calculations equally. Previous research has also shown that this property of the ‘non-epitope’ set did not produce significantly different results [122]. Each respective set of experimentally defined epitopes was denoted the positive dataset and the set of non-binding (or unknown) peptides was denoted the negative dataset.

4.2.2.1 The SYF^1 Dataset

The SYF1 dataset is a supertype dataset derived from SYFPEITHI [118] and is identical to that used in the original paper for NetCTL [103]. Each epitope in SYF^1 was experimentally verified to bind to one of 10 MHC class I supertypes [133]. The resulting dataset consisted of 148 epitope-supertype pairs. The corresponding negative dataset was obtained by concatenating the SwissProt entry proteins from which each of the epitopes was derived. The length of the concatenated protein sequence was 78,259 amino acids. The ROC curve (Section 4.2.3) was generated using a negative set of $\big((78,259 \times 10) - 148\big) = 782,442$ nonamers and a positive set of 148 nonamers. The positive set of SYF^1 is available in Appendix A, Table A.1.

4.2.2.2 The Lanl^{661} Dataset

Experimentally defined epitopes in HIV-I were extracted from the HIV Molecular Immunology Database [134]. In total, 1,618 CTL epitopes were found that were bound by human MHC molecules. However, this set was highly redundant; the epitope lengths were variable and a large number of epitopes differed only by mutations within the sequence. Also, resolution of their MHC typing varied from 2 to 4 digits. To correct for this variability, a number of changes were made to the MHC allele-epitope list. Firstly, all MHC alleles were defined to two digits. Secondly, variant epitopes binding the same allele were discarded. Finally, as the prediction software only produced binding predictions for nonamer epitopes, all epitopes that were not 9 amino acids long were removed from the list.

In summary, it was possible to test 41 of the 43 allelic predictors for MHC molecules in NetMHC v3.0. The positive set consisted of 661 epitopes, defined in terms of start and end positions relative to the HIV reference strain HXB2 (Appendix A, Table A.4) and a matching MHC type to 2 digits. The input protein sequence to NetMHC contained 3,000 overlapping nonamers that covered the proteome from which the whole positive set of epitopes was derived. The total 'negative set' for the ROC analysis was $\big((3,000 \times 41) - 661\big) = 122,339$ nonamers, and a positive set of 661 nonamers. The positive set of Lanl^{661} is available in Appendix A, Table A.3.

4.2.2.3 The Lanl^{179} Dataset

The Lanl^{661} dataset was modified for testing with NetCTL. From these 661 epitopes, a total of 179 bound to the 12 alleles for which NetCTL has allelic predictors. The input sequence to NetCTL contained 3,000 overlapping nonamers. For this experiment, the negative set consisted of $\big((3,000 \times 12) - 179\big) = 35,821$ nonamers, and a positive set of 179 nonamers. The positive set of Lanl^{179} is available in Appendix A, Table A.2.

4.2.3 ROC Curves

ROC curves give a visual measure of the accuracy of a prediction method. The threshold at which the prediction method identifies a peptide as being an epitope varies along the length of the curve. Each point on the curve gives the fraction of true positive epitopes found as a function of the number of false positive 'epitopes' at that threshold. Hence, setting a strict threshold for epitope detection will result in high specificity (correct predictions) but low sensitivity (missing a high proportion of true binders). The area under the ROC curve gives the AUC (Area under Curve) measurement. In order to test for significant difference between ROC curves, we conducted the bootstrapping analysis detailed in [135]. Briefly, using bootstrapping with replacement, 100 replicates were formed from each dataset and the resulting non-rescaled and rescaled whole AUC values were compared using a paired *t*-test.

4.2.4 Other Measurements of Performance

Using the 2 epitope datasets, HIV^{216} and SYFPEITHI^{863}, and the same methods from [136], we repeated 3 of the measurements described in that paper for the rescaled and non-rescaled results of NetCTL v1.2. For the Rank measure, we analysed the proteins from which each epitope was derived. For each protein, we calculated the rank of the epitope amongst all overlapping 9-mers using rescaling and non-rescaling scoring methods for all alleles. We then analysed these ranks to see which method ranked the epitopes higher. For the second method, we measured the specificity of both rescaling and non-rescaling at predefined sensitivities. Finally, we measured the sensitivity among the top 5% top-scoring peptides, again for the rescaled and non-rescaled binding affinities.

4.2.5 Other Data Sources

The training data for NetMHC v3.0 is available at the Immune Epitope Database and Analysis Resource (IEDB). An independent set of experimental epitope-allele binding affinities was obtained from IEDB by selecting all experimental data that did not originate from the laboratories of Sette *et al.* or Buus *et al.* (the training data originated from these two sources).

4.3 Results

4.3.1 The Effect of Rescaling on Qualitative Epitope Prediction

ROC curves were used to analyse the effects of rescaling on epitope prediction. Both NetCTL v1.2 and NetMHC v3.0 were tested and 3 datasets were used (Figure 4.1 and Table 4.1). In each case, rescaling resulted in a significant loss of performance (bootstrap test: $P < 0.001$).

In NetCTL v1.2, the TAP and cleavage scores are combined with the rescaled MHC binding score to produce a combined score for each submitted nonamer. In order to test how NetCTL performed without rescaling, it was still necessary to divide the MHC binding score by a rescaling value so the weightings of the TAP and cleavage score were

ROC Curve	Colour	Method	Dataset	Rescaling	AUC 30%	Bootstrap P Value
Figure 4.1 A	Black Solid Red Dashed	NetCTL v1.2 NetCTL v1.2	SYF^1 SYF^1	No Yes	0.949 0.937	$P < 0.001$
Figure 4.1 B	Black Solid Red Dashed	NetMHC v3.0 NetMHC v3.0	SYF^1 SYF^1	No Yes	0.932 0.905	$P < 0.001$
Figure 4.1 C	Black Solid Red Dashed	NetMHC v3.0 NetMHC v3.0	Lanl^{661} Lanl^{661}	No Yes	0.944 0.937	$P < 0.001$
Figure 4.1 D	Black Solid Red Dashed	NetCTL v2.1 NetCTL v2.1	Lanl^{179} Lanl^{179}	No Yes	0.933 0.918	$P < 0.001$

Table 4.1: The summary statistics and details of each ROC curve from Figure 4.1.

still applicable and accurate. By averaging over all rescaling values and dividing the MHC binding value by this number, rescaling differences were "averaged out" and it was still possible to use the extra information from the TAP and cleavage predictions.

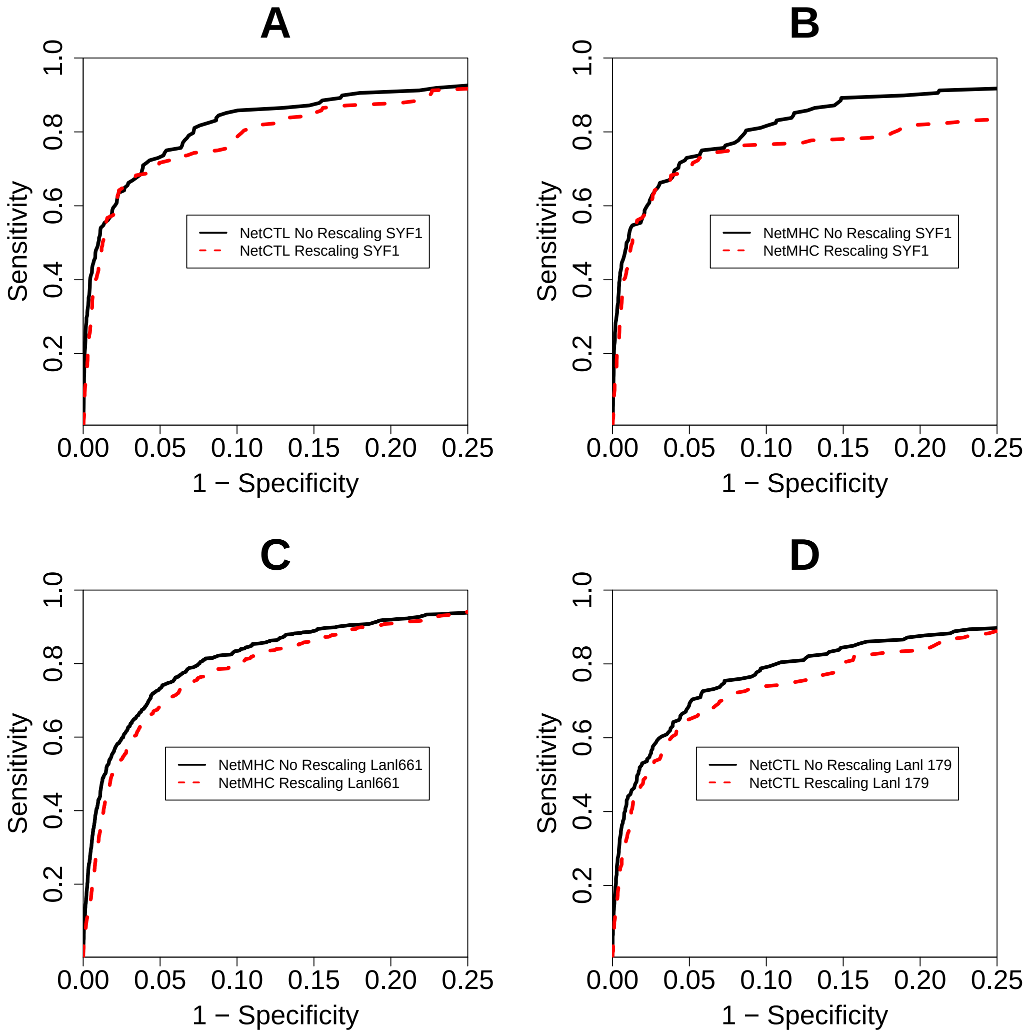


FIGURE 4.1: Each graph shows the ROC curves using different combinations of datasets and prediction methods (see Table 4.1). A uses NetCTL with the SYF^1 dataset, B uses NetMHC with the SYF^1 dataset, C uses NetMHC with the Lanl^{661} dataset and D uses NetCTL with the Lanl^{179} dataset. The x-axis has been scaled to show the region of importance (the AUC with high specificity values). The rescaled results (red dashed line) are compared against non-rescaled (black solid line). Table 4.1 gives the statistics for each graph.

4.3.2 Comparison of Rescale Values

We calculated rescale values based on the predicted binding to 500,000 peptides selected at random from *Mycobacterium tuberculosis*. To check that the source of the peptides did not alter our conclusions we randomly selected 500,000 natural peptides from the

Swiss-Prot database [137] and produced the top percentile re-scaling values for each allele from these peptides. Figure 4.2 compares these values to the re-scaling values we previously used, which were derived from non-overlapping peptides from *Mycobacterium tuberculosis*. As can be seen from Figure 4.2, the two sets of rescale values are strongly positively correlated ($R^2 = 0.9563$, $P < 0.001$). Repeating our ROC curve analysis of rescaled and non-rescaled predictions from Figure 4.1 C using these new rescale factors (Figure 4.3) gives very similar results to those reported here. Consequently, whether we calculate the rescale factors using random natural peptides from *Mycobacterium tuberculosis* or on random natural peptides from a range of proteins, our conclusions remain unchanged.

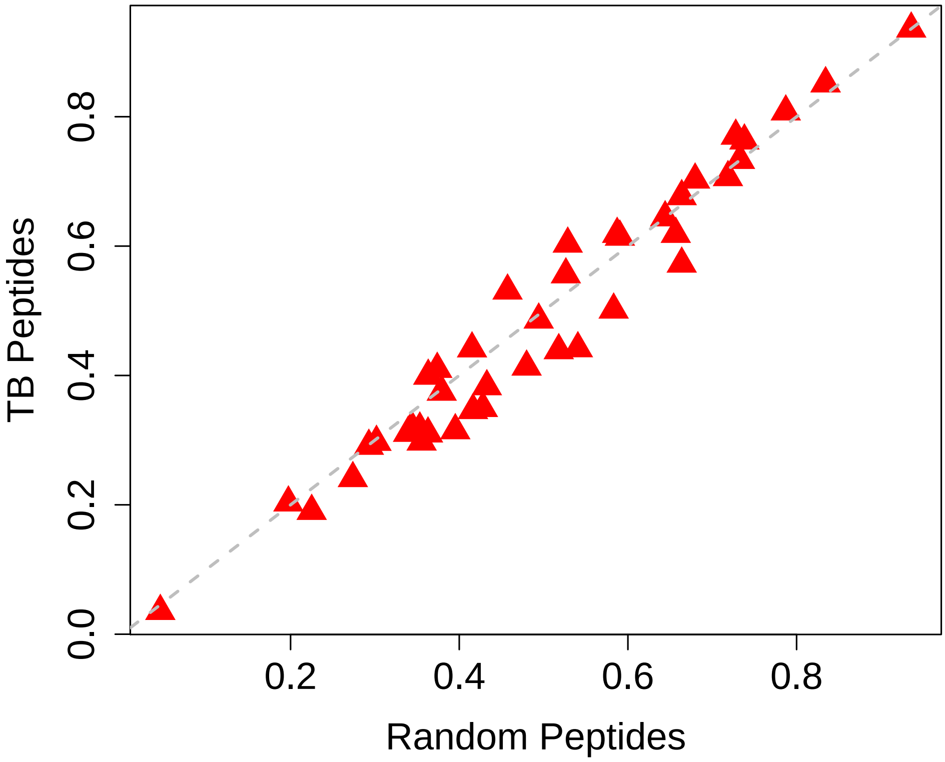


Figure 4.2: The relationship for the top percentile rescaling values for each allele between random natural peptides and non-overlapping peptides from Mycobacterium tuberculosis. There was no significant difference between the two measures (Mann-Whitney paired test, $P = 0.1181$) and the data was strongly correlated ($R^2 = 0.9563$, $P < 0.001$).

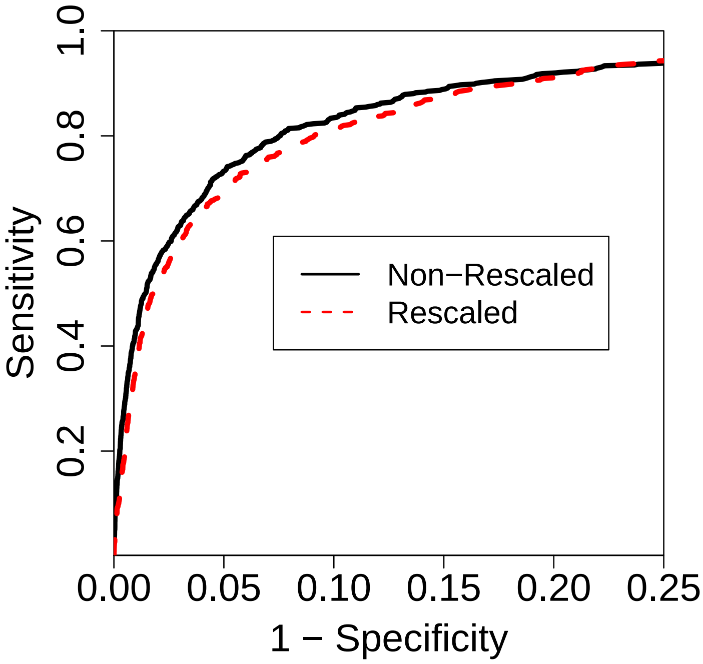


FIGURE 4.3: The ROC analysis of the Lanl^{661} dataset. The rescale values used are derived from random natural peptides, as opposed to peptides originating from *Mycobacterium tuberculosis*. The difference remains significant between the two curves (bootstrap test: $P < 0.001$).

4.3.3 Variation in Rescale Values as a Function of Accuracy

One possible explanation for why rescaling has a detrimental impact on prediction is that there may be a positive correlation between rescale factor and allelic predictor accuracy (Morten Nielsen, pers. comm.). To check this hypothesis we calculated the AUCs for each NetMHC v3.0 predictor using the Lanl^{661} dataset and plotted this against the corresponding rescale factor, the results of which are shown in Figure 4.4. This shows no evidence of a correlation between rescaling values and the AUC values ($R^2 = 0.0068$, $P = 0.606$).

Consequently, it is unlikely that a correlation between rescale values and AUC values explains our findings. However, certain alleles like B0801 do have both a low rescale value and a low AUC. To double check that these poor accuracy predictors were not causing the inaccuracies in rescaled predictions we repeated our ROC curve analysis for Lanl^{661} without the low accuracy predictors (those with an AUC value below 0.9; namely A6801, A6802, B3501, B0702, B0801, B0802 and B4501). In the remaining, reduced subset of predictors there was even less evidence for a correlation between AUC and rescale factor ($R^2 = 0.0007$, $P = 0.887$). For this subset of predictors the accuracy

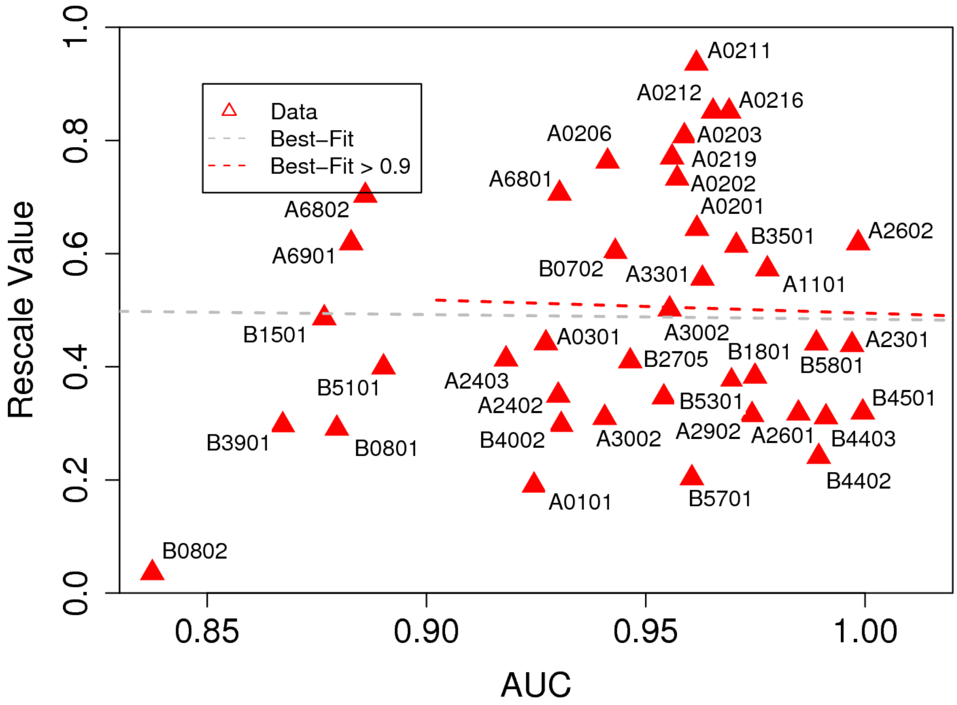


Figure 4.4: The relationship between AUC and rescale value. There is no evidence for a correlation of AUC and rescale value for the whole set of allele predictors ($R^2 = 0.0068$, $P = 0.606$), nor for the subset of predictors with an AUC > 0.9 ($R^2 = 0.0007$, $P = 0.887$). This analysis used the Lanl^{661} epitope dataset.

was still significantly better if rescaling was not applied (Figure 4.5; bootstrap test: $P < 0.001$) and comparable to the ROC curve analysis using the full set of alleles (Figure 4.1 A).

Therefore, we believe there is no evidence to support the hypothesis that the reason rescaling is detrimental is because there is a correlation between rescale factors and AUC.

4.3.4 Other Measurements of Performance

We used 3 other metrics [136] to compare predictive performance with and without rescaling.

1. The rank of known epitopes was compared with non-epitopes from the same protein for both rescaled and non rescaled predictions. From Figure 4.6, it can be seen that the non-rescaled results produced significantly more accurate results for both epitope datasets (paired Wilcoxon ranked sum test, $P < 0.001$).

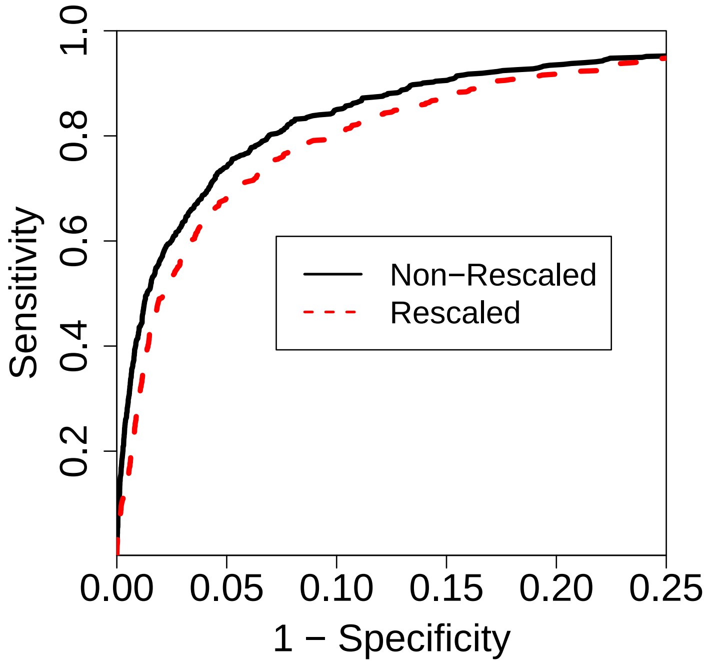


FIGURE 4.5: The result of the ROC curve analysis, using the Lanl^{661} dataset and excluding any alleles (7 in total) that had an AUC < 0.9 from Figure 4.7 (bootstrap: $P < 0.001$).

2. Predicted binding affinities that had not been rescaled produced improved results compared to predictions that had been rescaled at given sensitivities using the epitope datasets from [136] (Table 4.2).

Sensitivity	No Rescaling	Rescaling	Epitope Set
0.3 0.6 0.8	0.995 0.987 0.921	0.989 0.977 0.891	HIV^{216}
0.3 0.6 0.8	0.998 0.991 0.974	0.997 0.991 0.973	SYF^{863}

TABLE 4.2: The specificity of non-rescaled and rescaled results at specified sensitivity values. Epitope datasets are taken from [136].

3. Predicted binding affinities that had not been rescaled also showed an improvement over rescaled affinities when comparing the total number of true epitopes found among the top 5% of peptides predicted to bind (Table 4.3), again using the epitope datasets from [136].

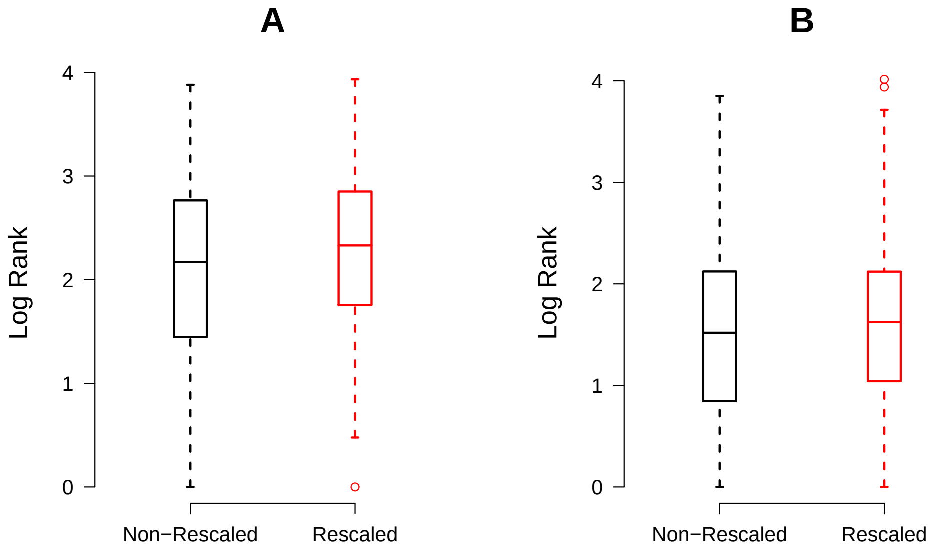


FIGURE 4.6: **(A)** A box plot showing the summary statistics of the ranks ($\log_{10}$) of each of the 216 epitopes in the HIV dataset among all overlapping 9-mer in the epitopes' source proteins. The ranks of the epitopes were significantly lower for non-rescaled scores compared to rescaled scores (Paired Wilcoxon ranked sum test, $P < 0.001$). The non-rescaled scores produced a higher rank for 170 epitopes and rescaled scores for 24 epitopes. **(B)** The same analysis using 863 epitopes from the SYFPEITHI dataset. The ranks of the epitopes were significantly lower for non-rescaled scores compared to rescaled scores (Paired Wilcoxon ranked sum test, $P < 0.001$). The non-rescaled scores produced a higher rank for 474 epitopes and rescaled scores for 369 epitopes.

Epitope Set	No Rescaling	Rescaling
SYF^{863}	0.885	0.877
HIV^{216}	0.718	0.690

TABLE 4.3: The fraction of the total number of epitopes in the 2 epitope datasets among the top 5% of predicted binding affinities.

4.3.5 The Effect of Rescaling on Quantitative Predictions of Binding Affinities

Using 2 sets of experimentally-derived epitope-allele binding affinities, we also showed that the correlation between predicted and experimental affinities was weaker with rescaling than without.

A set of 128 experimentally-derived epitope-allele binding affinities was extracted from the Immune Epitope Database and Analysis Resource [126]. This set of epitopes was known to have no involvement in the training of any of the allelic predictors in NetMHC v3.0 or NetCTL v1.2.

The relationship between the rescaled / non-rescaled predicted binding affinities and the experimental binding affinities was investigated (Figure 4.7). Rescaling resulted in a significantly larger error (the difference between predicted and experimental affinity) compared to predicted binding affinities that were not rescaled ($P < 0.001$). Although rescaling would naturally result in a larger quantitative error, additionally, the correlation between predicted and experimental affinities was weaker with rescaling than without (rescaled: $P < 0.001$, Spearman's $\rho = 0.40$; not rescaled: $P < 0.001$, Spearman's $\rho = 0.51$).

The analysis was repeated using a second experimental dataset. This second dataset came from the Sette and Buus laboratories and included the experimental data used to train NetMHC and NetCTL. The results obtained using this second dataset were very similar to those obtained using the first, independent dataset (Figure 4.8).

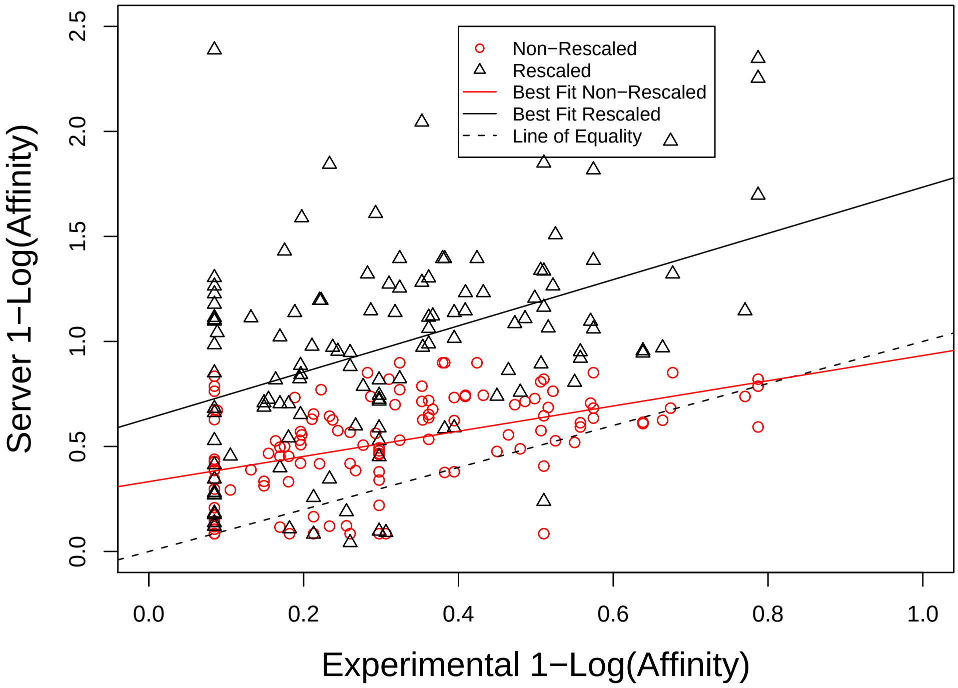


FIGURE 4.7: The experimental binding affinities for 128 epitopes were obtained from IEDB [126] and converted to a log scale ($1 - \log_{50000}(\text{affinity})$). These epitopes were then tested using NetMHC v3.0 to produce 2 sets of predicted binding affinities; rescaled or non-rescaled. The predicted scores were also converted to a log scale ($1 - \log_{50000}(\text{affinity})$) and the non-parametric Spearman's ρ was used to calculate the correlation between experimental and predicted data.

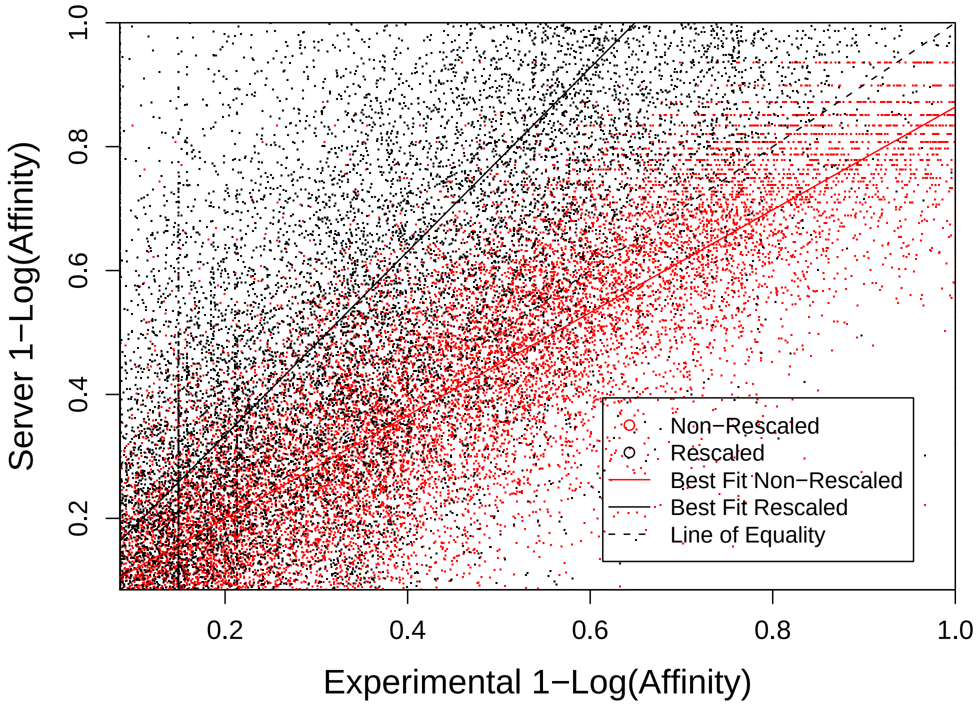


FIGURE 4.8: The experimental binding affinities for 29,336 epitopes were obtained from IEDB and converted to a log scale ($1 - \log_{50000}(\text{affinity})$). These epitopes were then tested using NetMHC v3.0 to produce 2 sets of predicted binding affinities; rescaled or non-rescaled. The predicted scores were also converted to a log scale ($1 - \log_{50000}(\text{affinity})$) and the non-parametric Spearman's ρ was used to calculate the correlation between experimental and predicted data. The correlation between non-rescaled predicted affinities and experimental data showed a P value of < 0.001 (Spearman's $\rho = 0.877$). Rescaled predicted affinities and experimental data gave a P value of < 0.001 (Spearmans $\rho = 0.816$). The absolute difference between the 2 best-fit lines and the line of equality was calculated and it was shown that the non-rescaled values were significantly closer to the line of equality (Wilcoxon Paired Signed Rank Test; P value < 0.001).

4.3.6 The Effect of Negative Data Volume

As explained in Section 4.2.2, we multiplied the negative set of each dataset by the number of allele predictors being tested. This was to mirror an analysis where one would check every possible peptide-allele pair of a pathogen-MHC class I interaction when searching for potential epitopes. A possible argument was that the resultant positive/negative ratio in our datasets was unrealistically low with such a high proportion of negative data. To counter this, the SYF^1 dataset was modified to contain 148 positive epitopes and 78,111 negative peptides, which gave a positive/negative ratio of 0.2%, a figure close to the estimated 1% of all natural peptides that would bind to a given MHC molecule [103]. In order to test the difference in prediction accuracy between rescaled and non-rescaled predicted affinities with this reduced dataset, each of the 78,111 negative peptides was randomly paired with 1 of the 10 supertype predictors (see Section 4.2.2.1) and the predicted binding affinity was found for each of theses pairs. Rescaling again resulted in a significant loss of performance (bootstrap test: $P < 0.001$, Figure 4.9).

Related to this result, we performed a general analysis of the effect of positive and negative set size on the AUC values obtained in ROC curve analysis. As this is outside the scope of this chapter, this is explained in Appendix A.3.

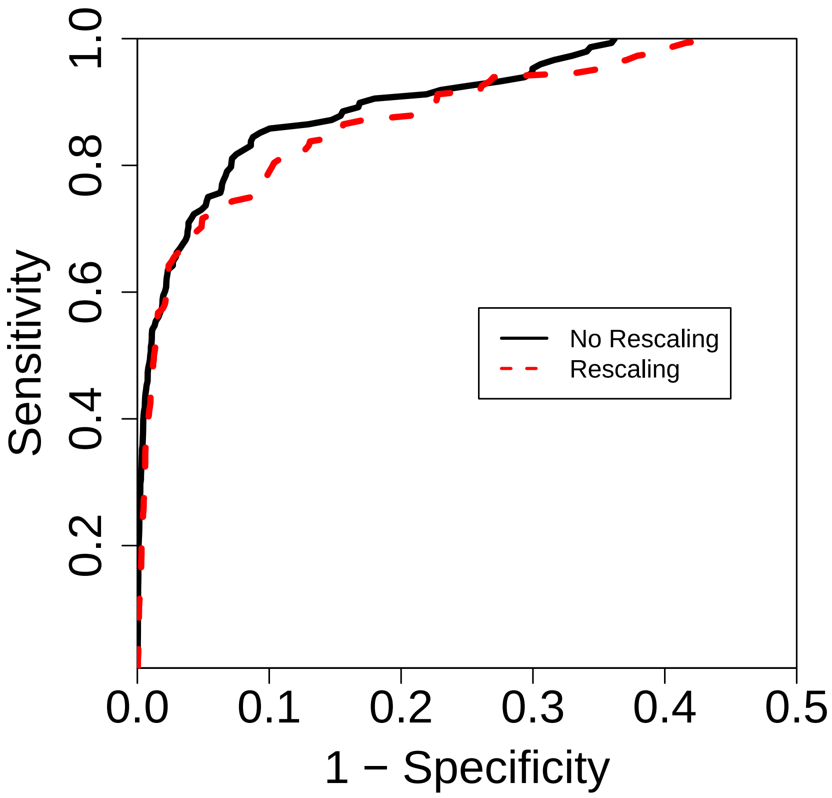


FIGURE 4.9: The result of the ROC curve analysis, using the SYF^1 dataset, with a negative set of 78,111 peptides. The difference remains significant between the two curves (bootstrap test: $P < 0.001$).

4.3.7 Is Rescaling Necessary to Maintain Low Variation in Sensitivity?

Another argument that could be made for rescaling is that, in its absence, those allele predictors with a lower accuracy would have lower sensitivity i.e. fewer epitopes would be detected from those particular alleles (Morten Nielsen, pers. comm.). We tested this hypothesis using the SYF^1 dataset. A score threshold value for rescaled and non-rescaled affinity values was identified at a specificity of 0.95 for the complete dataset

(0.2068 for non-rescaled affinity values and 0.4330 for rescaled affinity values, using $1 - \log_{50000}(\text{affinity})$). Next the sensitivity and specificity values per supertype allele were calculated at these threshold values. The result is shown in Figure 4.10. There is no evidence for a large decrease in the variation in allelic sensitivity upon rescaling. The ranges of sensitivities are identical with or without rescaling and the standard deviations are very similar (0.1526 non-rescaled, 0.1528 rescaled), if anything, slightly higher upon rescaling.

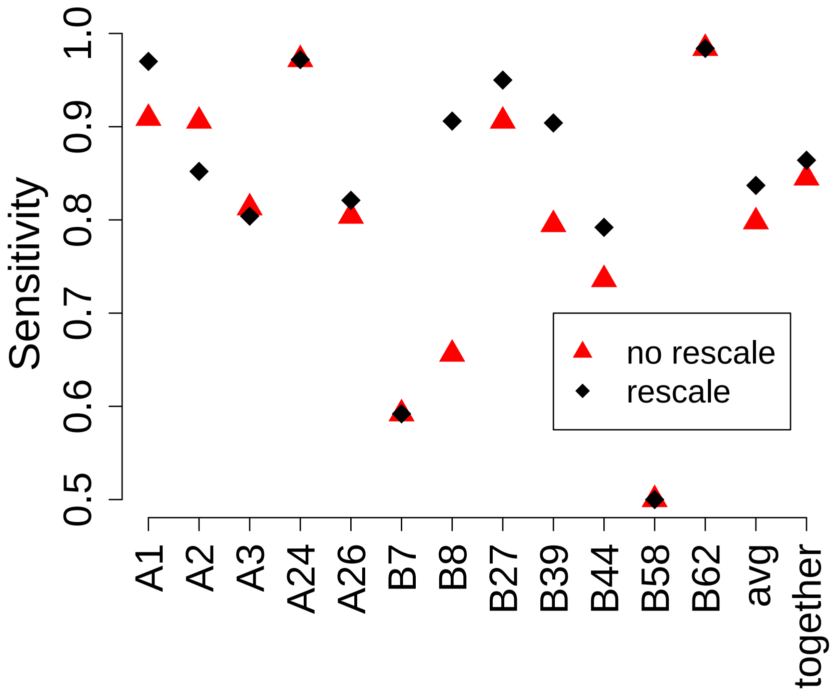


FIGURE 4.10: The effect of rescaling on sensitivity. For each supertype, the sensitivity was calculated when rescaling and not rescaling the predicted binding affinities. The final two results give the average sensitivities across all supertypes ('avg') and the sensitivities of all supertypes measured together ('together').

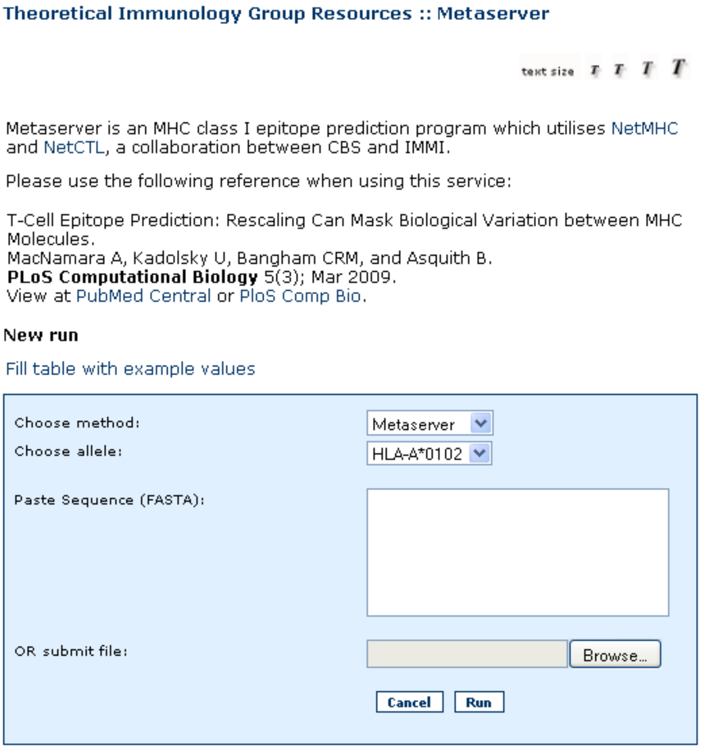
4.4 Producing Metaserver

These results demonstrated a significant improvement in accuracy over NetMHC v3.0 and NetCTL v2.1 when classifying epitopes across a number of HLA class I alleles. Therefor, we decided to produce our own web-based prediction server based on these results, which we called Metaserver.

Metaserver is summarised in Equation 4.1. In this calculation the NetMHC estimated binding affinity is still combined with a rescale value in order to take advantage of the NetCTL-specific information relating to the processing (TAP and cleavage) of the peptide before it binds to HLA class I. However, the rescale value for Metaserver is the same across all alleles and is an average of the rescale values for each individual allele. In summary, Metaserver takes binding information from NetMHC, TAP and cleavage information from NetCTL, but removes the assumption that each HLA class I binds the same number of peptides by 'averaging out' the rescaling of the predicted binding affinity.

$$\text{Metaserver Epitope Score} = \text{NetMHC Binding Affinity}/\ \text{`Averaged' Rescale Value} + w_1 * \text{NetCTL TAP} + w_2 * \text{NetCTL Cleavage} \tag{4.1}$$

w_1 and w_2 above represent the weightings that are applied to the TAP and cleavage prediction scores respectively. In all calculations, $w_1 = 0.05$ and $w_2 = 0.15$. Metaserver uses a Perl LWP script to access the web servers NetMHC and NetCTL and the website itself is written in PHP (Figure 4.11).



Theoretical Immunology Group Resources :: Metaserver

Metaserver is an MHC class I epitope prediction program which utilises NetMHC and NetCTL, a collaboration between CBS and IMMI.

Please use the following reference when using this service:

T-Cell Epitope Prediction: Rescaling Can Mask Biological Variation between MHC Molecules.
MacNamara A, Kadolsky U, Bangham CRM, and Asquith B.
PLoS Computational Biology 5(3); Mar 2009.
View at PubMed Central or PloS Comp Bio.

New run

Fill table with example values

Choose method: Metaserver

Choose allele: HLA-A*0102

Paste Sequence (FASTA):

OR submit file:

Figure 4.11: A screenshot showing the Metaserver web-based resource. The site is hosted at:
http://linuxwebdev.cc.ic.ac.uk/theoreticalimmunology/trans/metaserver.php

4.5 Discussion

Rescaling is, in theory, a sound approach to improving epitope prediction and in particular comparability of predictions obtained using different allelic predictors. However, using a number of different measures of accuracy, in the context of two commonly used prediction methods, we have demonstrated that rescaling actually impairs rather than improves predictive performance and comparability. We suggest that rescaling predicted affinities results in a loss of information that outweighs any advantage gained in correcting for differences in training data.

The first approach used ROC curve analysis and showed clear differences between rescaling and non-rescaling. The ROC curve gives a graphical representation of how well the prediction method ranks true epitopes among a set of non-binding peptides. Or to use an analogy, how efficient it is at finding the epitopic needle in a haystack of random peptides. From Figure 4.1, it is clear that rescaling across all allelic predictors results in a performance loss in terms of how well the method ranks its peptides by binding affinity; that is, rescaling impairs intra-allelic comparisons. This loss could be demonstrated using epitope data from a number of sources (SYFPEITHI, the HIV Molecular Immunology Database) and with two different methods of prediction (the combined approach of NetCTL v1.2 and NetMHC v3.0). This effect of rescaling would be detrimental to any studies screening across a number of alleles for possible epitopes (such as [115]). The effect of this performance difference can be gauged from Figure 4.1 A. In order to identify correctly 85% of the epitopes the percentage of false positives detected was 9% and 15%, for non-rescaled and rescaled methods respectively. To put this result into context, the viral protein NS1 from the H5N1 strain of Avian Influenza A consists of 221 overlapping nonamers. To screen this protein for potential epitopes, 33 epitopes would need to be experimentally checked for each MHC molecule of interest if rescaled predictions were used, as opposed to 20 for the non-rescaled predictions (providing 85% epitope coverage was sufficient).

Added to the significant results from the ROC curve analysis, we also demonstrated the positive effect of removing rescaling in terms of the correlation with experimental data (Figure 4.7) and also in terms of per-protein and sensitivity analysis (Figure 4.6 and Table 4.2 and Table 4.3). Taken together, these results strongly demonstrate the improvement in accuracy of removing the condition of rescaling when comparing predictions between alleles.

There has been little research on the variation in ‘stickiness’ among MHC molecules, i.e. whether some MHC class I molecules are capable of binding to a greater number of epitopes than others. The binding motifs for MHC-peptide binding vary across the range of alleles, but the assumption made for rescaling is that each molecule would bind to the same number of peptides out of a large random selection. Estimates based upon mass spectrometry suggest that over 2,000 peptides are associated with HLA-A2.1 and -B7 and it is speculated that the actual total could be over 10,000 per MHC molecule [138]. However, it is not known how this number varies between molecules. It has been postulated that the twin constraints of effective pathogen recognition but tolerance of self would result in a very narrow range of promiscuity for viable MHC class I molecules [130]. Contrary to this, recent research has shown that this range may be wider than initially envisaged [139] and our results suggest that there is considerable inter-allelic variation in promiscuity.

This data may also be informative regarding optimization of peptide cargo in the endoplasmic reticulum (ER). We would argue that peptide optimization is the biological interpretation of rescaling: alleles have similar numbers of epitopes because peptides with a lower binding affinity are replaced in the ER. We know that optimisation cannot be complete because otherwise every allele would just present one epitope: the one with highest affinity. However, it seems likely that there is a degree of optimization [131, 132]. The observation that rescaling gives worse predictions may put a bound on how much optimisation is occurring. Allied to this, it has been observed that the release of an MHC class I molecule from the peptide-loading complex with a suboptimal peptide takes precedence over the prolonged detention of the MHC class I molecule in the complex until an optimal peptide comes along [131]. Hence, peptide optimization acts to reduce inter-allelic variation and promiscuity results from inter-allelic variation in allele-peptide affinity. However, this peptide optimization is limited by time and is not complete and hence, we note this variation in promiscuity across different alleles.

In summary, we suggest that much of the observed variation between allelic predictors reflects genuine biological information which should not be discarded as experimental noise and that rescaling is based on an unjustified assumption: that all alleles bind the same number of peptides. Removing this assumption, we have demonstrated a significantly improved predictive performance. These conclusions are important both for studies that use prediction methods to understand the CTL response and for T cell epitope discovery programs where avoiding rescaling could save a large amount of experimental effort, ultimately leading to improved vaccine implementation.

In the context of our work on HTLV-I, this research allowed us to fully test NetMHC v3.0 as software to predict HTLV-I epitopes. NetMHC v3.0 demonstrated high accuracy identifying experimentally verified epitopes from HIV (Lanl^{179} and Lanl^{661}), human and other viral sources (SYF^{1}). Chapter 5 follows on from this verificiation to test NetMHC v3.0 in the context of HTLV-I and and hence predict HTLV-I epitopes. From this information, we could test hypotheses relating to protective and detrimental effects of MHC class I alleles.

Chapter 5

The Prediction of Disease From Peptide Binding Affinities

5.1 Introduction

The application of epitope prediction software has almost exclusively been used to identify specific epitope - MHC class I complexes i.e. identifying epitopes from a predefined protein that bind to a particular MHC class I molecule [140]. To the best of our knowledge, this study was the first to use this software to define the epitope repertoire of a virus (HTLV-I) across a large cohort of individuals, in order to identify the epitope properties of a successful CD8$^+$ T cell response. The possibility of success depended on the accuracy of the epitope prediction software used. In Chapter 4, we tested this software and produced our own implementation (Metaserver) of two widely available web-based epitope prediction servers, NetMHC and NetCTL. We showed that Metaserver was able to predict HIV epitopes with high sensitivity (Section 4.3). These results allowed us to bring forward this software to use with HTLV-I and test our hypotheses concerning the role of HLA class I in infection.

Most HTLV-I-infected individuals have a strong, chronically activated CD8$^+$ T cell response to HTLV-I and it is unclear why this fails to eradicate the virus. Furthermore, there is evidence for both protective effects [1, 87, 100] and pathogenic effects [7, 86, 141, 142] of HTLV-I specific CD8$^+$ T cells. As in all viral infections, the attributes of a protective antiviral response *in vivo* are unknown, although specificity for the viral protein Tax is a strong candidate. There are good reasons to believe that a Tax-specific CD8$^+$ response [143] may be particularly protective. Firstly, Tax is the immunodominant HTLV-I antigen in this response [8, 11]. Secondly, HLA-A*02, which is associated with protection in southern Japan [1], binds several Tax epitopes [10], notably Tax 11–19, which is bound unusually strongly [144]. Thirdly, Tax is one of the first HTLV-I proteins to be expressed and it has been shown, for HIV-I infected cells *in vitro*, that CD8$^+$ T cells specific to early viral proteins are particularly effective in viral control [145]. Finally, it has been shown that the selective pressure exerted on Tax is higher in asymptomatic carriers than in those that have developed HAM/TSP [80].

5.1.1 How can CD8$^+$ cell protective efficacy be quantified?

Measurements of CD8$^+$ cell frequency, phenotype, function and specificity are informative but, because antigen load influences each of these factors, it can be difficult to ascertain if a particular immune profile is the cause or effect of good pathogen control [146–149]. An alternative approach is host genotype analysis. Polymorphisms in immune-related genes, particularly the HLA class I genes, have been associated with outcome in *Plasmodium falciparum*, *Mycobacterium tuberculosis*, HIV-I, HTLV-I and Hepatitis B Virus infection. The benefit of a genotypic analysis is that the direction of causality is unequivocal; the drawback is that, in common with all 'omics' approaches to identify biomarkers, mechanistic insight is limited. Provided linkage disequilibrium can be ruled out, class I associations imply that the protective effect is mediated by CD8$^+$ T or NK cells. However, why one particular allele should be protective remains unclear and so provides no information about how to manipulate the immune response to enhance protection.

Hence, the aim of this section was to develop a method to test the hypothesis that the effectiveness of an individuals HTLV-I-specific response and thus their proviral load and HAM/TSP risk was determined by the epitope binding properties of their HLA class I alleles. Specifically, we focused on the question, "does strong binding to peptides from a specific HTLV-I protein benefit the host?". We used a novel ranking measure to define the relationship between HLA class I and HTLV-I proteins. This approach is generally applicable to all pathogens, including those in which few epitopes have been identified experimentally.

5.2 Methods

5.2.1 Epitope Prediction

We used two different algorithms to predict HLA class I epitopes: Metaserver and Epipred. Figures based on Metaserver predictions are in the main text, the corresponding figures for Epipred are in supplementary information.

5.2.1.1 Metaserver

Metaserver is described in detail in Section 4.4. Briefly, it is a combination of two web-based prediction methods that use artificial neural nets, NetCTL v1.2 [103] and NetMHC v3.0 [124, 125]. Metaserver combines the two methods and removes a normalising assumption (which maintains that all alleles bind the same number of peptides) to produce a technique that shows improved accuracy in epitope prediction [150] and predicts epitopes for 43 HLA molecules.

5.2.1.2 Epipred

In order to validate our results, we used a second, independent method of epitope prediction [122]. Epipred uses a logistic regression model that is trained on all available data across all HLA class I alleles and then specified for an individual allele.

5.2.2 Epitope Prediction - Allele Coverage

Metaserver provided coverage of 84% of the total count of A/B alleles in the Kagoshima cohort.

The missing alleles are: A0207, A0210, A2603, A3201, B1301, B1501, B1508, B1511, B1518, B2704, B3701, B3802, B4005, B4006, B4601, B4801, B5201, B5501, B5504, B5601, B5603, B5605, B5705, B5901, and B6701.

We were able to obtain predictions for {A0207, A0210}, A2603 and {B4005, B4006} to a resolution of 2 digits by combining the predictions of other A02*, A26* and B40* predictors according to their frequency in Kagoshima. For example, to obtain a 2-digit "A02" predictor that could be used in place of A0207, Equation 5.1 was used:

$$\text{A02 Binding Affinity} = \frac{\sum_{i=3^{rd}/4^{th}digit}^{n} (\text{Binding Affinity: A02i} * \text{Freq A02i})}{\sum_{i=3^{rd}/4^{th}digit}^{n} \text{Freq A02i}} \tag{5.1}$$

A02i being the set of n 4-digit A02 alleles in the Kagoshima cohort for which we have predictors.

5.2.3 Prediction Quality

The accuracy of epitope prediction algorithms has increased to such an extent that the correlation between predicted binding affinities and measured binding affinity is as strong as the correlations of measurements between different laboratories [135]. The specificity of epitope predictors has been tested by predicting a set of CTL epitopes and subsequently verifying CD8$^+$ T cell responses against these epitopes experimentally. Using this technique has yielded true-positive (correctly predicted) estimates of 62-80% [151]. Using the more direct approach of mass spectrometry to determine HLA-peptide binding yielded a true positive rate of greater than 98% [152]. Additionally, we verified the prediction software we used (Metaserver and Epipred) for HTLV-I peptides (Section 5.2.6).

5.2.4 The Rank Measure

Both prediction methods that we use produce a score for each peptide-HLA that represents the binding strength of that complex. In theory this score would allow us to compare predicted binding affinities between alleles. However, between allele comparisons can be problematic. Firstly, within-allele comparisons (i.e. predictions for different peptides to the same allele) are thought to be more comparable than predictions between alleles [103]. Secondly, whether or not a normalisation procedure should be applied for between-allele comparisons is still being debated in the community [150]. To avoid the potential problem of between-allele comparisons we used the rank measure technique introduced by Borghans *et al.* [117] in which she quantified the strength of peptide-HLA class I binding for peptides from a particular protein by ranking the strength of binding of peptides from the protein of interest to the allele amongst the strength of binding of peptides from the entire proteome to that allele. Specifically, we split each protein in the HTLV-I reference sequence into overlapping nonamers offset by a single amino acid. Using the epitope prediction software, a predicted binding affinity score was calculated for each of these peptides to each HLA allele of interest. For each allele we ranked all nonamers from the proteome from the strongest to weakest predicted binding scores. This produced a list of rank values for each protein to that particular allele that quantified the binding relationship between that allele and the protein (an example is given in Table 5.1). Additionally, we repeated all calculations simply using the raw predicted affinity score rather than the rank measure. All of our conclusions were replicated (Table 5.12).

	A*02		B*54		C*08	
1	Gag	TPKDKTKVL	Tax	LPTTLFQPA	Tax	YLYQLSPPI
2	Pol	PADPKEKDL	Pro	LPVIPLDPA	Tax	LLFGYPVYV
3	Rof	RPPPAPCLL	Env	FPFSLLVDA	Pol	ALLGEIQWV
4	P12	RPPPAPCLL	Pol	MPVFTLSPV	Pol	SLISHGLPV
5	Gag	NANKECQKL	Rof	LPITMRFPA	Pol	FQPYFAFTV
6	Gag	ANNPQQQGL	P12	LPITMRFPA	Gag	FMQTIRLAV
7	Gag	GAPPNHRPW	Pro	LPFRTTPIV	Pol	LTYDAVPTV
...	...	...	...	...	...	...
3389	P12	LLLFLLPPS	Tax	DNDHEPQIS	Tax	DNDHEPQIS

Table 5.1: An example of the rank method used to measure the targeting of specific HTLV-I proteins by HLA class-I alleles. The predicted binding affinities for every overlapping nonamer peptide in the HTLV-I proteome ($N = 3389$) was derived for each allele of interest. These were then ordered from strongest to weakest binding. Then, for each protein, the associated rank values were taken as a measure of the strength of binding of that protein by that allele. In the table above, the alleles previously associated with disease outcome and proviral load are shown along with the ordered HTLV-I peptides that bind to that allele (1 being the strongest, 3389 being the weakest). The strongest binders from Pol, for example, would be as follows: {Cw*08, 2}, {B*5401, 4}, {A*02, 3}.

5.2.5 Independence of Ranks

The rank method of Borghans *et al.* in its original form [117] assumes that the predicted ranks are independent. We were concerned that the binding of the top 8 peptides from a protein to an allele may not be independent of one another because the strength of the strongest binder provides information about the strength of the second highest binder. For this reason, apart from Figure 5.3, which we also verified by an independent method, only the top rank for each protein-allele pair was used. This data is shown in Appendix B, Table B.6.

5.2.6 Experimental Quantification of HLA Class I - Peptide Binding

The REVEAL$^{\text{TM}}$ HLA-peptide binding assay (ProImmune Ltd., Oxford, UK) was used to quantify peptide-HLA binding. For each allele-peptide combination that was tested, assembly of peptide-HLA complexes was quantified by ELISA with a conformation-dependent anti-HLA antibody. Samples of assembling peptide-HLA complexes were taken at a defined time point and snap-frozen in liquid nitrogen prior to analysis. The assembly for each peptide-HLA complex was then compared against a positive control peptide for that allele as the percentage of assembled peptide relative to that control. We selected four HLA class I alleles and 50 HTLV-I peptides for each allele. The allele choice was based on allele frequency in the Kagoshima database and included 2 A alleles and 2 B alleles as well as alleles for which we knew that the epitope prediction tended to be poor. The 50 HTLV-I nonamer peptides for each allele were selected to represent a range of predicted binding affinities, from weak to strong binding peptides. They originated from 4 HTLV-I reference strain proteins: Tax, HBZ, Gag and Polymerase.

5.2.7 Protective versus Detrimental Alleles

Method for Results Section 5.3.2.

Due to allele coverage (see Section 5.2.2), it was necessary to use Metaserver for A*0201 and B*5401 and Epipred for Cw*0801. As the rank values were derived for each allele separately, it was acceptable to use different prediction methods for each allele in this case. The ranks of the strongest binding 8 peptides from each protein to the alleles A*0201 and Cw*0801 (16 rank values) were compared against the ranks of the strongest binding 8 peptides to the allele B*5401 (8 rank values). A Wilcoxon-Mann-Whitney test was performed for each protein to test for differences between the two sets of rank values. Table 5.2 shows an example of this calculation.

Count	Cw*08 Ranks		A*02 Ranks		B*54 Ranks
1	17		1		1
2	18		2		17
3	26		14		28
4	55	AND	23	VERSUS	31
5	90		33		32
6	92		35		33
7	95		46		39
8	104		67		40

Table 5.2: Protective class-I alleles bind HBZ strongly: For each protein (in this example, Tax), the ranks of the top 8 binding peptides from the protein to the allele were compared between detrimental (B*54) and protective (A*02 and Cw*08) alleles.

5.2.8 HAM/TSP versus Asymptomatic Carriers

Method for Results Section 5.3.3.

The analysis was carried out on each HTLV-I protein in turn. For each individual in the Kagoshima cohort, the rank of the top binding peptide from the HTLV-I protein to each of the individuals A and B HLA class I alleles was found (see Section 5.2.4). These ranks were then split into two groups - those from HAM/TSP patients and those from asymptomatic carriers (AC). The two sets of ranks (HAM/TSP versus AC) were then compared for each protein using a Wilcoxon-Mann-Whitney test (null hypothesis: HAM/TSP patients and asymptomatic carriers bind the protein equally strongly). Table 5.3 shows an example of this calculation.

<table>
<tr><th colspan="5">HAM/TSP</th></tr>
<tr><th>Individual</th><th>Rank of strongest binding HBZ peptide to locus A1</th><th>Rank of strongest binding HBZ peptide to locus A2</th><th>Rank of strongest binding HBZ peptide to locus B1</th><th>Rank of strongest binding HBZ peptide to locus B2</th></tr>
<tr><td>HAM/TSP 1</td><td>{A2402, <b>208</b>}</td><td>{A2402, <b>208</b>}</td><td>{B4002, <b>3</b>}</td><td>{B4002, <b>3</b>}</td></tr>
<tr><td>HAM/TSP 2</td><td>{A2402, <b>208</b>}</td><td>{A3101, <b>42</b>}</td><td>{B5101, <b>42</b>}</td><td>{B0702, <b>84</b>}</td></tr>
<tr><td>HAM/TSP 3</td><td>{A2402, <b>208</b>}</td><td>{A2601, <b>2</b>}</td><td>{B5401, <b>125</b>}</td><td>{B3501, <b>93</b>}</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>HAM/TSP 230</td><td>{A2601, <b>2</b>}</td><td>{A3101, <b>42</b>}</td><td>{B3501, <b>93</b>}</td><td>{B3501, <b>93</b>}</td></tr>
<tr><th colspan="5">AC</th></tr>
<tr><td>AC 1</td><td>{A2402, <b>208</b>}</td><td>{A2601, <b>2</b>}</td><td>{B5401, <b>125</b>}</td><td>{B5601, <b>NA</b>}</td></tr>
<tr><td>AC 2</td><td>{A2402, <b>208</b>}</td><td>{A3301, <b>9</b>}</td><td>{B3501, <b>93</b>}</td><td>{B4402, <b>2</b>}</td></tr>
<tr><td>AC 3</td><td>{A2402, <b>208</b>}</td><td>{A2402, <b>208</b>}</td><td>{B3501, <b>93</b>}</td><td>{B4402, <b>2</b>}</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>AC 202</td><td>{A0201, <b>22</b>}</td><td>{A3101, <b>42</b>}</td><td>{B3501, <b>93</b>}</td><td>{B4001, <b>7</b>}</td></tr>
<tr><td colspan="5">All rank values (in bold) for the HAM/TSP group were compared against all rank values (in bold) for the AC group using a Wilcoxon-Mann-Whitney test.</td></tr>
</table>

Table 5.3: Asymptomatic carriers bind HBZ more strongly than HAM/TSP patients: For each protein (in this example, HBZ), the rank of the strongest binding peptide to each allele of the A and B loci was found for each individual. These were then compared between HAM/TSP and AC. Key = {allele, rank of strongest binding peptide from the protein of interest to that allele}.

5.2.9 Rank versus Proviral Load

Method for Results Section 5.3.4.

We considered each HTLV-I protein in turn. Firstly, we split the cohort by disease status (AC or HAM/TSP). Then, for each individual, we counted the number of alleles they possessed that were strong binders to the protein of interest and then tested for a correlation between the number of strong binders to the protein and proviral load using the Spearman rank correlation. A strong binding allele to a particular protein was defined as one that was in the top 40% of alleles. That is, the rank of the top binding peptide from the HTLV-I protein to each of the individuals A and B HLA class I alleles was found (see Section 5.2.4). This set of rank values (pooled HAM/TSP and AC) was then ordered from highest to lowest rank and the alleles that were represented in the top 40% of these ranks were defined as strong binding alleles to that protein (see Table 5.4). Importantly, for each protein, we looked at the relationship between strength of binding and proviral load separately in HAM/TSP patients and ACs and then combined the P values using Fishers combined test (rather than simply looking at the relationship in the whole cohort). Therefore we could be confident that any relationship between protein binding and proviral load that we found did not follow trivially from a relationship between protein binding and disease status and the fact that asymptomatic carriers have a significantly lower load than HAM/TSP patients.

Our alternative metric for this method used the Rank Measure to quantify the strength of binding of peptides from each HTLV-I protein to each individuals A and B alleles. We then tested for any correlation between these values and the individuals proviral load for HAM/TSP patients and asymptomatic carriers.

1	2	3	4	5	6	7	8	9	10	...	n = number of alleles in the cohort (1728)
{A2601, 2}	{A2601, 2}	{A2601, 2}	{B4402, 2}	{B4402, 2}	{B4002, 3}	{B4002, 3}	{B4001, 7}	{A3301, 9}	{A0201, 22}	...	{A2402, 208}

Table 5.4: HBZ peptide binding is a consistent predictor of proviral load: Again, for each protein (in this example, HBZ), the relationship between the number of strong binding alleles to peptides from that protein and proviral load was examined. For each protein, the definition of a strong binding allele to that protein was as follows: For HBZ, the {HLA, rank} data from both HAM/TSP and AC groups was ordered from strongest to weakest binding.

5.2.10 Prevented Fraction of Disease, $\mathbf{F_P}$

Method for Results Section 5.3.7.

To calculate the prevented fraction (F_P) of disease [1, 153], we used a 2 × 2 contingency table (Table 5.5):

	$\mathbf{G^+}$	$\mathbf{G^-}$
D	$a = 183$	$b = 47$
H	$c = 181$	$d = 21$

Table 5.5: The input matrix for the F_P test. The numbers indicated are those used for the calculation in Section 5.3.7.

D = disease (HAM/TSP), H = healthy, G^+ = positive for protective genotype and G^- = negative for protective genotype. The fraction (F_P) of potential cases of disease D in the population that is prevented by the genotype G^+ is given by Equation 5.2:

$$F_P = (1 - R) \times \left[1 - \left(\frac{dr_1}{br_2}\right)\right] \tag{5.2}$$

R is the prevalence rate of disease D in the population, $r_1 = a + b$ and $r_2 = c + d$. In the case of HAM/TSP, R is estimated as $\leq 1\%$ of the HTLV-I-infected population. F_P is approximately normally distributed: the standard deviation is given by Equation 5.3:

$$SD\left(F_P\right) = (1 - R - F_P) \times \sqrt{\left[\left(\frac{c}{d}r_2\right) + \left(\frac{a}{b}r_1\right)\right]} \tag{5.3}$$

5.2.11 Detection of HTLV-I specific CD8$^+$ T cells

All subjects attended the HTLV-I clinic at St Marys Hospital, London, gave written informed consent and the study was approved by the St Marys NHS Trust Local Research Ethics Committee. Peripheral blood mononuclear cells (PBMC) were isolated from whole blood from HTLV-I infected individuals by density gradient centrifugation. PBMC were depleted of CD4$^+$ T cells using MACS beads (Miltenyi Biotec). The resulting cells were cultured in duplicate at a density of 100,000 cells per well in the presence of a range of concentrations of pooled overlapping 20mer peptides (offset by 6 amino acids) spanning HBZ, Tax, or with medium alone. After 6 hours, IFN-γ producing cells were detected by ELISpot (Mabtech). The threshold for a positive response to peptide was defined as greater than the mean plus two standard deviations of the number of spots in the medium only control.

5.2.12 HTLV-I Proteome

The reference strain is from [154], with the exception of HBZ, which was identified more recently and described in [155] (see Appendix B, Table B.8).

5.3 Results

5.3.1 Verification of Epitope Prediction Software

Approximately 50 HLA class I-epitope pairs have been identified for HTLV-I [10, 83, 143, 156] (mainly from the immunodominant protein Tax [11] in the context of A*02); this represents a small and non-random fraction of the approximately 2200 nonamer epitopes that could be bound by the alleles of the Kagoshima cohort[1]. Therefore we used epitope prediction software to systematically predict HTLV-I epitopes. The epitope prediction software that we used has been extensively validated for other organisms, but because of the lack of experimental data, it has not previously been tested for HTLV-I. One of the most stringent requirements for a peptide to be an epitope is its ability to bind the HLA allele of interest, so to validate the epitope prediction software, we measured experimentally the binding affinity of 200 HTLV-I peptide-allele combinations (Appendix B, Table B.7). We found a strong positive correlation between experimental measurement and the theoretical prediction for each of the two epitope prediction methods used (Metaserver: all $P < 0.00001$, Spearmans rank correlation; Figure 5.1. Epipred: all $P < 0.001$, Spearmans rank correlation; Figure 5.2. Full P values in Table 5.6). We conclude that these epitope prediction software packages accurately predict relative (i.e. rank order) HTLV-I peptide binding affinities. Throughout this chapter, the tests based on Metaserver predicted affinities were repeated using Epipred predicted affintities. All conclusions were replicated by both methods and by the alternative metric of raw predicted binding affinities (Table 5.12). The SIR metric, an alternative method of quantifying protein specificity, was also tested in Appendix B, Section B.1.

	Metaserver		Epipred	
Allele	R_S	P	R_S	P
A0201 ($n = 50$)	0.76	1×10^{-10}	0.48	4×10^{-4}
B0702 ($n = 44$)	0.62	9×10^{-6}	0.65	2×10^{-6}
A2402 ($n = 49$)	0.65	5×10^{-7}	0.68	8×10^{-8}
B3501 ($n = 49$)	0.68	9×10^{-8}	0.47	6×10^{-4}
Combined		3×10^{-25}		1×10^{-16}

Table 5.6: The Spearman rank coefficients and P values for each of the comparisons between predicted and experimental binding data.

[1] This figure is 1% [103] of the 3,389 overlapping nonamers of the HTLV-I proteome multiplied by the number of unique alleles (65) in the cohort.

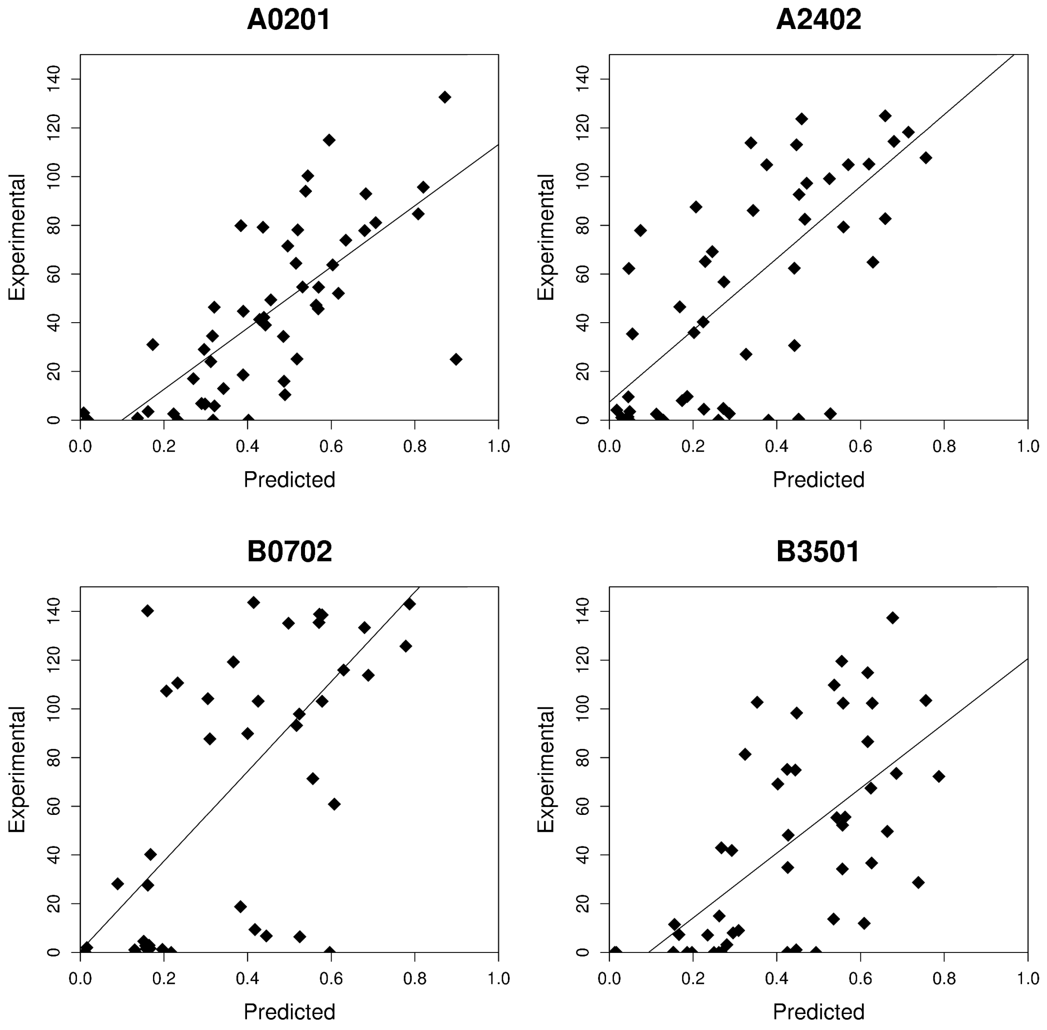


FIGURE 5.1: The correlation between the experimentally measured binding affinities (% binding compared to control peptide) and the predicted binding affinities ($1 - \log_{50000}(\text{affinity})$) of Metaserver for each of the 4 alleles analysed.

5.3.2 Protective class I alleles bind HBZ strongly

A number of associations between HLA class I alleles and protection or disease risk in HTLV-I infection have been identified in a population in southern Japan [1, 4]. We performed a rigorous reanalysis of these associations for verification and refinement in Chapter 3. From this data, *A*0201* and *Cw*0801* were classed as protective alleles and *B*5401* as a detrimental allele in the context of disease risk and proviral load. Hence, we started our analysis of HTLV-I epitopes with these alleles.

We compared the predicted HTLV-I peptide-binding affinities of *A*0201* and *Cw*0801*, with those of *B*5401* (see Methods, Section 5.2.7). Peptides from the HTLV-I protein HBZ bound to HLA A*0201 and Cw*0801 significantly more strongly compared to B*5401 ($P = 0.0002$, Wilcoxon-Mann-Whitney; Figure 5.3, Table 5.7). Repeating the

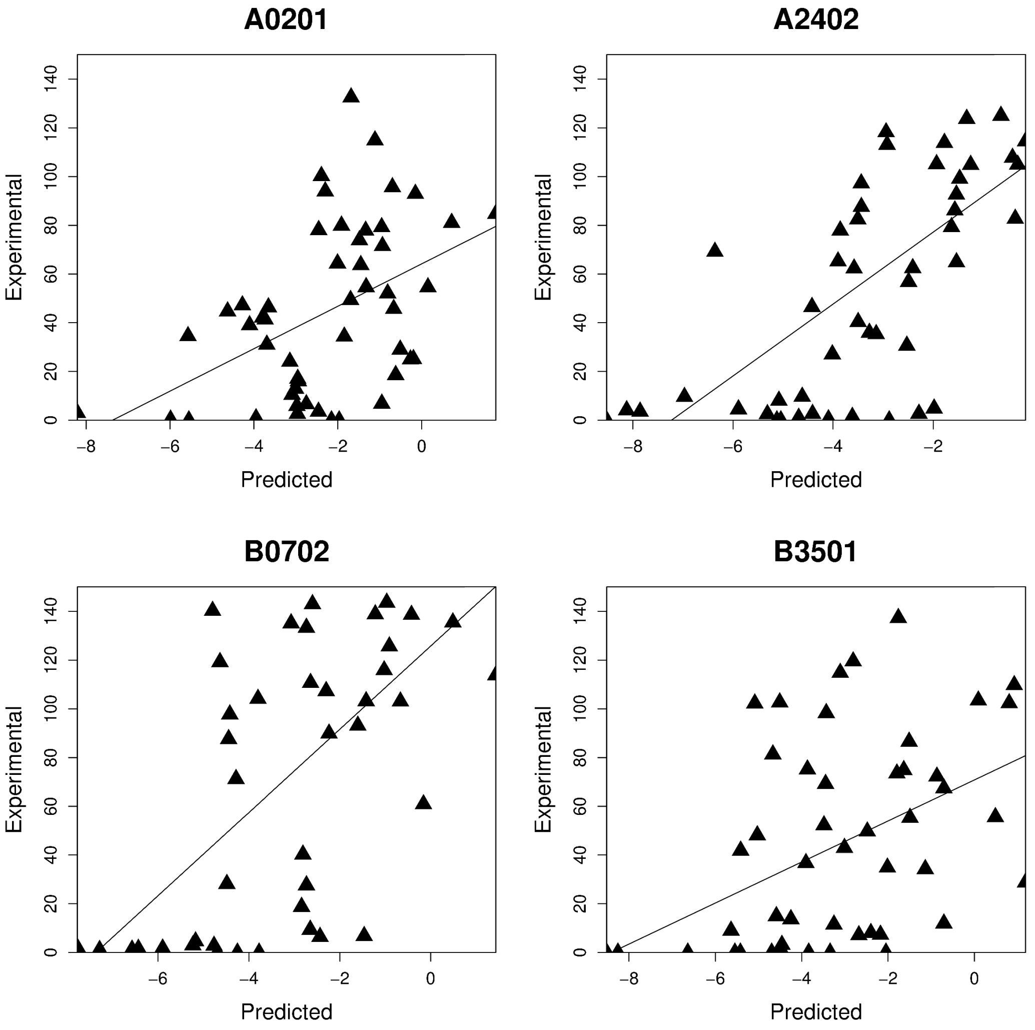


FIGURE 5.2: The correlation between the experimentally measured binding affinities (% binding compared to control peptide) and the predicted binding affinities ($1 - \log_{50000}(\text{affinity})$) of Epipred for each of the 4 alleles analysed.

analysis with another protective allele from the *A*02* family, *A*0206*, instead of *A*0201* yielded identical conclusions ($P = 0.0007$, Wilcoxon-Mann-Whitney; data not shown). These P values need to be treated with caution because the rank of the binding affinity of one HBZ peptide for A*0201 may not be independent of the rank of the binding affinity of a second peptide to A*0201 and similarly for Cw*0801 and B*5401. However, we also found that the difference in binding strength (i.e. the rank of the top A*0201 binding peptide minus the rank of the top B*5401 binding peptide) was significantly greater for HBZ than for other HTLV-I proteins ($P < 0.001$, binomial test). This statistic is based only on the top binding peptide so it does not assume different peptides have independent binding affinity ranks. Henceforth, we only considered the top binding peptide to avoid the potential problem of dependence (see Methods, Section 5.2.5).

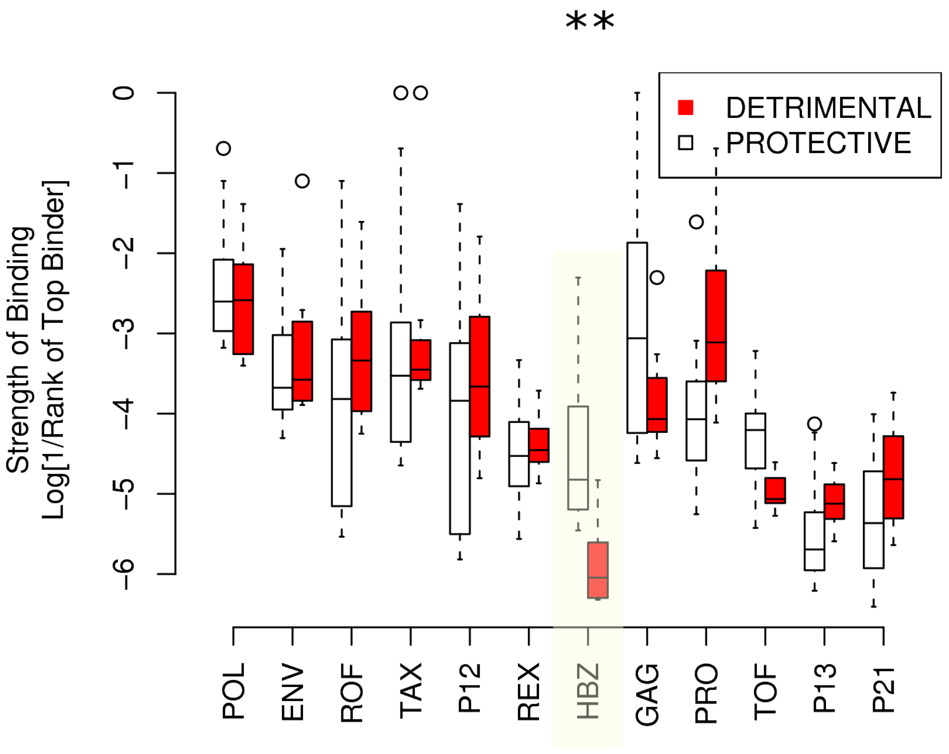


Figure 5.3: The strength of binding of protective alleles (A*02/Cw*08) and detrimental alleles (B*54) across the 12 HTLV-I proteins. The y-axis gives strength of binding of the top 8 binding peptides from each protein. The level of significance indicated is corrected for multiple comparisons. All *P* values are shown in Table 5.7.

Protein	Mann-Whitney 2-tailed *P* value
Pol	0.4999
Env	0.5003
Rof	0.2091
Tax	0.4257
P12	0.2978
Rex	0.5283
HBZ	0.0002
Gag	0.2572
Pro	0.0087
Tof	0.0131
P13	0.0523
P21	0.1200

Table 5.7: The *P* values associated with Figure 5.3.

5.3.3 Asymptomatic carriers bind HBZ more strongly than HAM/TSP patients

Having established that the known protective HLA class I alleles bind to peptides from HBZ more strongly than the known detrimental allele, we examined peptide binding by all alleles in the Kagoshima cohort. We compared the predicted epitopes for asymptomatic carriers ($n = 202$) and HAM/TSP patients ($n = 230$) from the Kagoshima cohort. We predicted the HTLV-I peptides bound most strongly by each individual, given their HLA class I types and then tested for differences between the two subject groups (see Methods, Section 5.2.8). The results are illustrated in Figure 5.4 and Table 5.8. One result remained highly statistically significant after correction for multiple comparisons and was consistent across both prediction methods: asymptomatic carriers have HLA class I alleles that bind more strongly to peptides from HBZ compared to HAM/TSP patients (Metaserver: $P = 0.0002$, Wilcoxon-Mann-Whitney. Epipred: $P < 0.0001$, Wilcoxon-Mann-Whitney; Figure 5.4). A bootstrap analysis was performed to confirm this result in both Metaserver and Epipred (Figure 5.5 and Figure 5.6).

	Metaserver			Epipred		
Protein	P value (2 tailed)	Group with strongest binding	Significance after correction	P value (2 tailed)	Group with strongest binding	Significance after correction
pol	0.0005	AC	**	0.0744	AC	-
env	0.0019	HAM	*	0.7203	HAM	-
rof	0.0023	HAM	*	0.0127	HAM	-
tax	0.3320	AC	-	0.8320	HAM	-
p12	0.0168	HAM	-	0.0940	HAM	-
rex	0.4706	AC	-	0.7639	AC	-
HBZ	0.0002	AC	**	0.000002	AC	***
gag	0.0011	AC	*	0.1265	HAM	-
pro	0.0970	HAM	-	0.0143	HAM	-
tof	0.4111	HAM	-	0.0256	HAM	-
p13	0.8524	AC	-	0.7308	AC	-
p21	0.0341	AC	-	0.0018	HAM	*

TABLE 5.8: The differences in the strength of binding of the alleles between AC and HAM/TSP patients to each of the 12 HTLV-I proteins.

To test whether this association was caused solely by the known protective and detrimental HLA alleles, the analysis for HBZ was repeated excluding A*02 and B*54. The results showed that, amongst the HLA-A alleles, A*02 was responsible for the protective effect, whereas in HLA-B more than one allele contributed significant effects. Overall, strong binding of HBZ peptides was associated with asymptomatic status, even

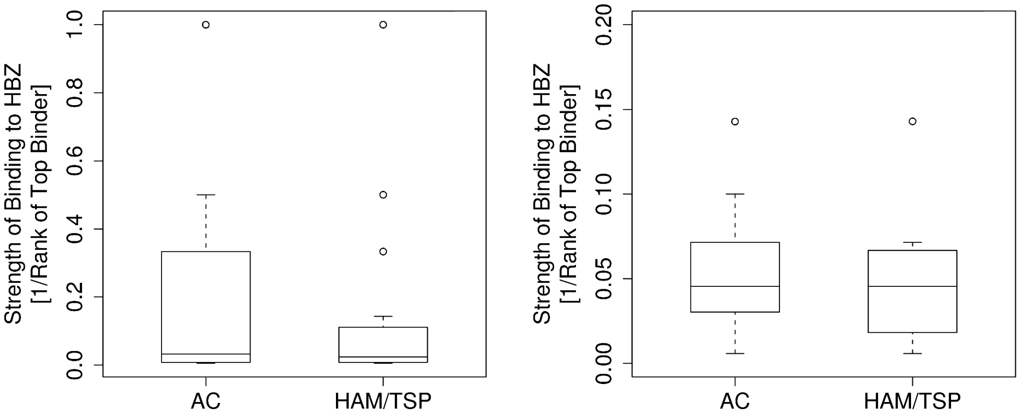


Figure 5.4: The strength of binding of the HLA class I alleles of asymptomatic carriers and HAM/TSP patients to HBZ. Asymptomatic carriers have HLA class I alleles that bind HBZ significantly more strongly than HAM/TSP patients (Metaserver, left panel: $P = 0.0002$. Epipred, right panel: $P = 2 \times 10^{-6}$).

when A*02, B*54 and Cw*08 were excluded from the analysis (Metaserver: $P = 0.04$, Wilcoxon-Mann-Whitney. Epipred: $P = 0.006$, Wilcoxon-Mann-Whitney; Table 5.9).

		Whole cohort ($N = 202, 230$)	Excluding A*02 & B*54 ($N = 84, 116$)
Metaserver	A alleles	0.006	0.81
	B alleles	0.001	0.01
	Combined	0.0005	0.04
Epipred	A alleles	0.0009	0.72
	B alleles	0.0002	0.001
	Combined	0.000001	0.006

Table 5.9: The difference in binding strength to HBZ between HAM/TSP patients and asymptomatic carriers. The first column gives the P values of the Wilcoxon-Mann-Whitney tests for the A and B loci. The second column repeats this analysis excluding individuals with either the A*02 or B*54 alleles.

5.3.4 Individuals whose HLA class I genotype predisposed them to bind HBZ peptides strongly had a significantly lower proviral load

Next we investigated why strong binding of HBZ peptides was associated with remaining asymptomatic. One of the best predictors of HAM/TSP is a high proviral load of HTLV-I [157]. We therefore tested the hypothesis that strong predicted binding of HBZ

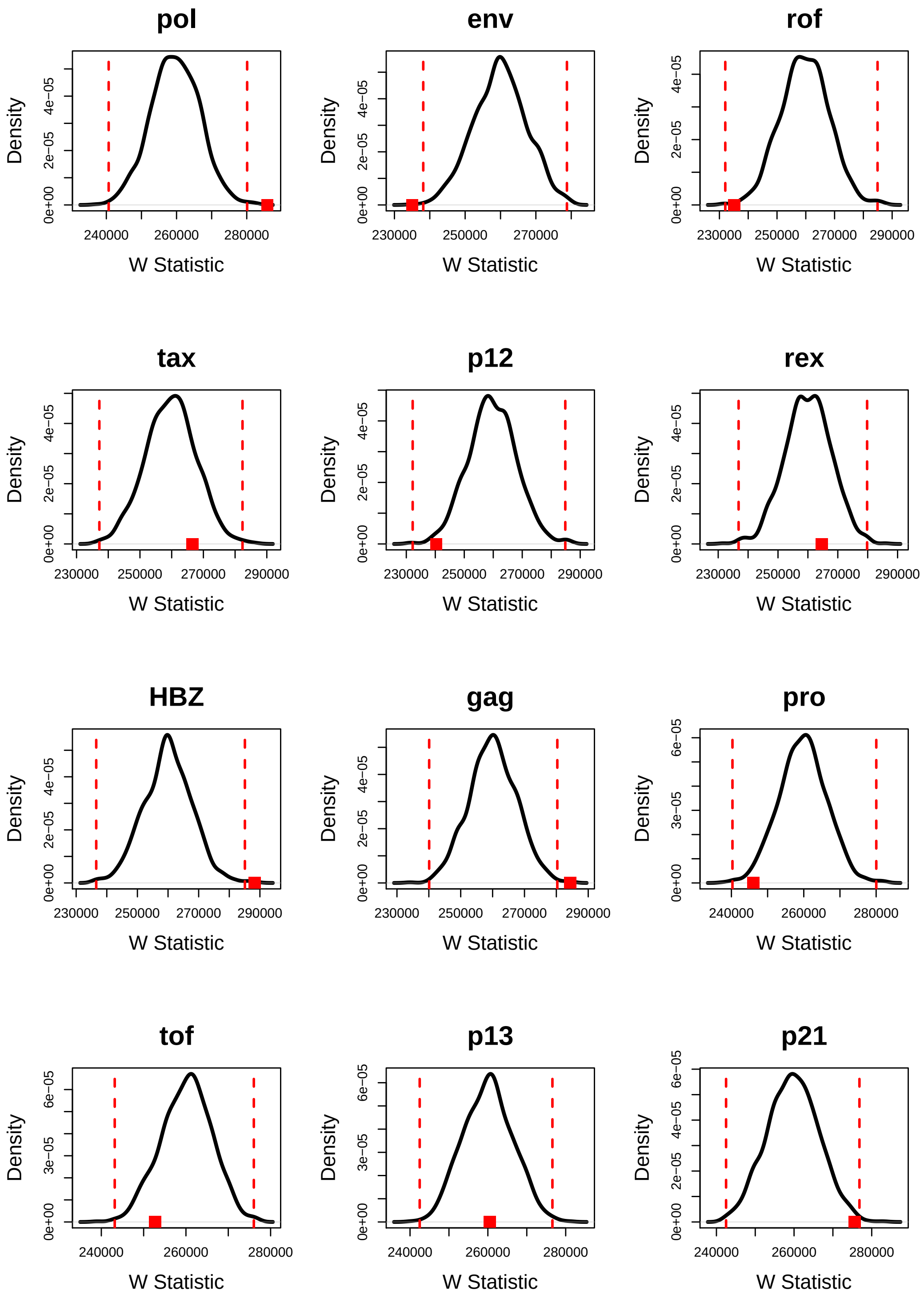


FIGURE 5.5: A bootstrapping method to validate the conclusion of Section 5.3.3 (ACs bind HBZ more strongly than HAM/TSP patients). The 432 individuals of the Kagoshima cohort were randomly assigned to an 'AC' and 'HAM/TSP' group. The Mann-Whitney test was then performed on these groups. This was repeated 1,000 times and the density plot of the resultant W statistics of each test was plotted, together with the W statistic from the 'true' test (red dot). The dotted lines represent the 2-tailed levels of significance after the Bonferroni adjustment for multiple comparisons. As can be seen from the HBZ graph, the W statistic value is still significantly different from the null distribution of bootstrapped W statistic values. This analysis for Metaserver was repeated for Epipred in Figure 5.6.

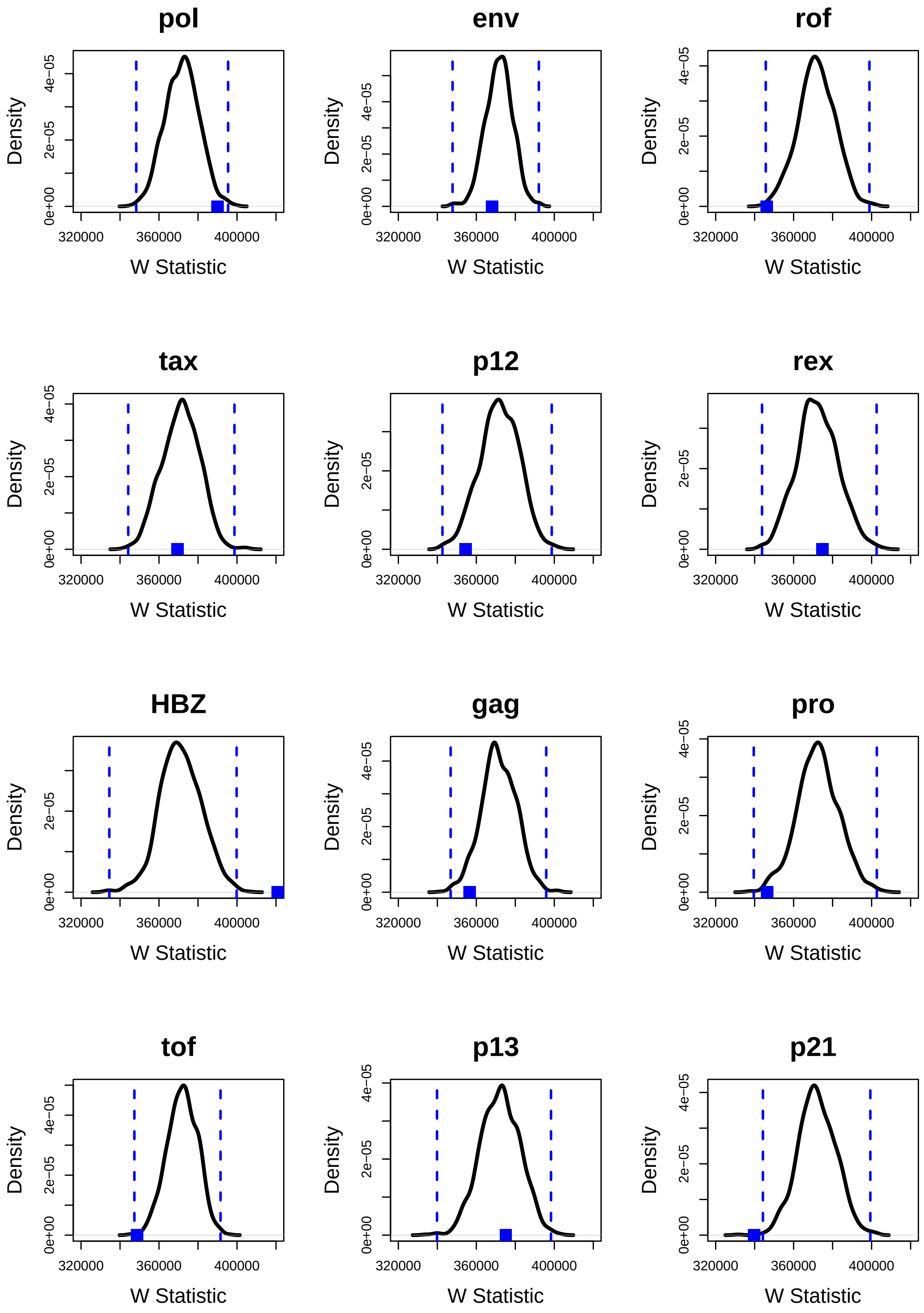


Figure 5.6: The bootstrap analysis for Epipred, described in Figure 5.5. The W statistic value for HBZ is significantly different from the null distribution of bootstrapped W statistic values.

peptides was associated with a lower proviral load. The number of alleles that each individual possessed that strongly bound peptides from HBZ was plotted against their proviral load (see Methods, Section 5.2.9). We found that the number of HLA Class I alleles that an individual had that strongly bound HBZ peptides was significantly negatively correlated with their proviral load (Metaserver: $P = 0.016$, Spearmans rank correlation. Epipred: $P = 0.1$, Spearmans rank correlation; Figure 5.7). We tested this correlation independently in HAM/TSP patients and asymptomatic carriers and then combined the P values (rather than simply testing the whole cohort), so this result does not follow trivially from our previous observation than asymptomatic carriers bind HBZ significantly more strongly than HAM/TSP patients. An alternative metric, the binding strength of the top HBZ-binding peptide to each allele instead of the number of strongly binding alleles, yielded an identical conclusion i.e. there was a significant negative correlation between the proviral load and the strength of binding to HBZ peptides (Metaserver: $P = 0.008$, Spearmans rank correlation. Epipred: $P = 0.003$, Spearmans rank correlation).

5.3.5 HBZ Peptide Binding is a Consistent Predictor of Proviral Load

Next we compared our peptide-binding analysis of HLA class I genotype with a traditional frequency-based “presence or absence of an allele” analysis. Previously a “traditional” analysis yielded inconsistent results [1, 4, 57]. For example, A*02 was a significant predictor of load in ACs but not in patients with HAM/TSP. We therefore directly compared the ability of the two methods to predict proviral load in ACs and HAM/TSP patients (Table 5.10). This analysis showed that whilst binding HBZ was a significant predictor of proviral load in both ACs and HAM/TSP patients ($P = 0.001$, $P = 0.017$), HLA-A*02 (presence/absence) was a significant predictor in ACs only ($P = 0.01$) and HLA-B*54 for HAM/TSP patients only ($P = 0.019$). The proportion of variance in proviral load explained was marginally higher for the peptide binding analysis. The observation that HBZ binding strength correlated with proviral load in both ACs and HAM/TSP patients suggests that peptide binding is the more fundamental predictor than HLA genotype. Further details of the predictive strength of peptide binding for proviral load and disease risk is described in Appendix B, Section B.2.

5.3.6 Proteins whose peptides are bound strongly by asymptomatic carriers are those associated with a lower proviral load

In the work above we established that the HTLV-I protein that is associated with the most significant reduction in HAM/TSP risk when bound by HLA class I molecules

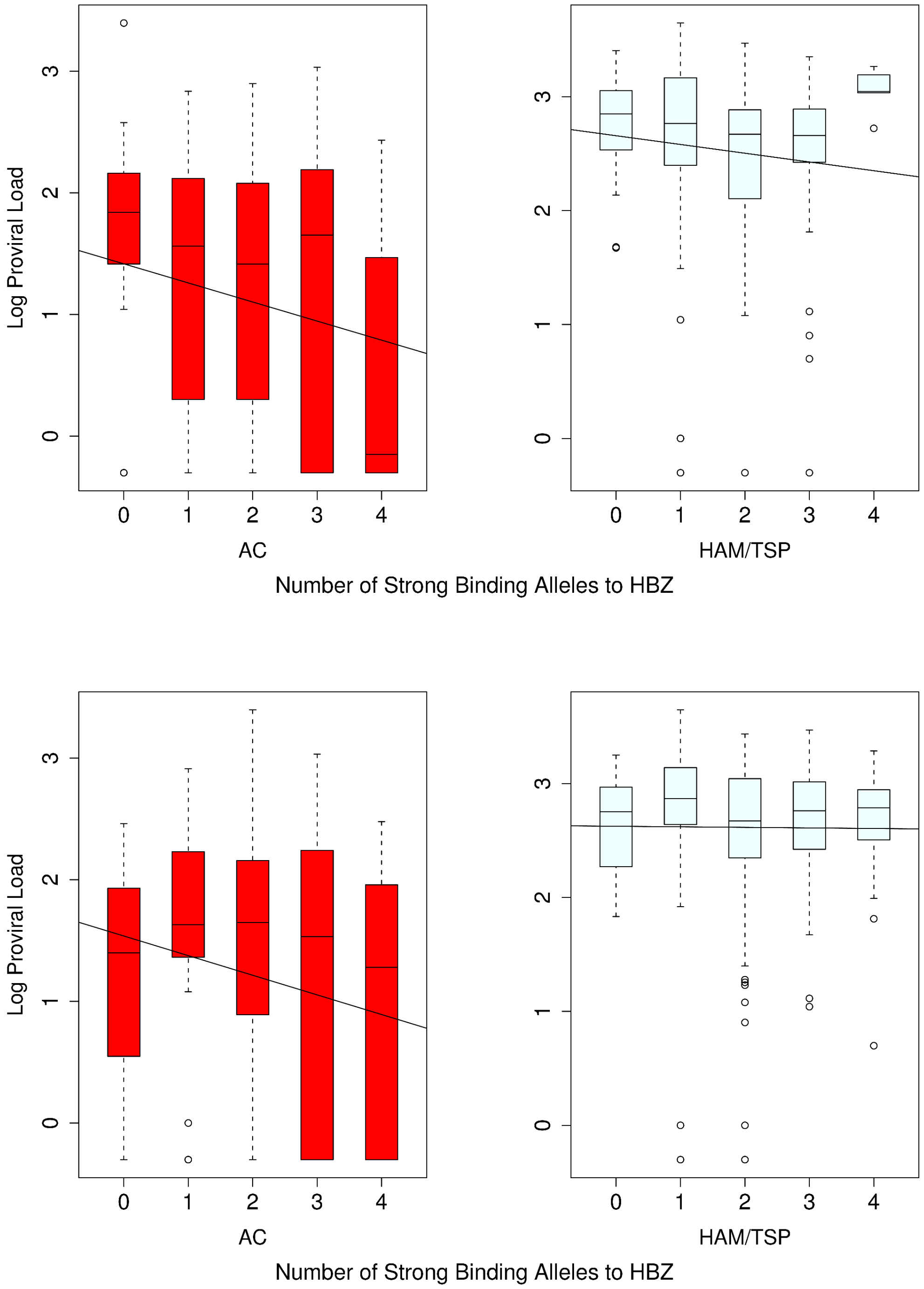


FIGURE 5.7: The count of strong binding alleles to HBZ per individual against their proviral load in AC and HAM/TSP groups. The number of strong binders to HBZ is significantly negatively correlated with proviral load (Metaserver, top panel: $P = 0.016$. Epipred, bottom panel: 0.1).

	Binding (A and B only)			Genotype (A and B only)		
AC Proviral Load	HBZ Pro	0.001 0.013	*** *	A*02	0.01	**
	$R^2 = 0.054$			$R^2 = 0.034$		
HAM/TSP Proviral Load	HBZ	0.017	*	B*54	0.019	*
	$R^2 = 0.026$			$R^2 = 0.025$		

Table 5.10: The significant predictors and their associated P values for each of the multiple regression models of proviral load.

(i.e. HBZ, Table 5.8) is also, independently, associated with a significant reduction in proviral load when bound (Figure 5.7). We wished to investigate whether this relationship held across all proteins. We ranked each HTLV-I protein by the following criteria:

1. Is strong binding of peptides from this protein associated with a lower HAM/TSP prevalence?
2. Is strong binding of peptides from this protein associated with a lower proviral load (tested independently in AC and HAM/TSP groups and then recombined, to avoid trivial associations)?

Table 5.11 illustrates this concept. The first column ranks the HTLV-I proteins according to whether they were bound more strongly by asymptomatic carriers or HAM/TSP patients (Figure 5.8 x-axis; at the extremes ACs were significantly more likely to bind peptides from HBZ, HAM/TSP patients were significantly more likely to bind peptides from Env). This list could be viewed as the "rank order of targets for a vaccine designed to reduce HAM/TSP risk". The second column ranks the proteins according to whether binding their peptides was associated with a lower proviral load (Figure 5.8, y-axis; at the extremes binding of HBZ was associated with a significantly lower proviral load, binding of Env was associated with a significantly higher proviral load). This list could be viewed as the "rank order of targets for a vaccine designed to reduce proviral load".

We then compared these two sets of ranks and found them to be strongly positively correlated (Metaserver: $R_S = 0.86$, $P = 0.0005$, Spearmans rank correlation; Figure 5.8. Epipred: $R_S = 0.66$, $P = 0.02$, Spearmans rank correlation; Figure 5.9). That is, proteins whose peptides are bound strongly by asymptomatic carriers are, independently, those associated with a lower load when bound. This observation has two important implications. Firstly, HLA class I binding of peptides from different proteins has a differential impact on both proviral load and HAM/TSP risk. Secondly, the fact that across all alleles and across all proteins, peptide binding associated with immune control (reduced proviral load) is strongly correlated with prevention of HAM/TSP is the strongest evidence yet that the CD8$^+$ T cell response can have a beneficial role in HTLV-I infection.

	Targeting this protein is associated with:	
	Reduced HAM/TSP Frequency	Reduced Proviral Load
Best	HBZ	Gag
	Pol	HBZ
	Gag	Pol
	P21	Rex
	Tax	Tof
↓	Rex	P21
	P13	Pro
	Tof	Tax
	Pro	P13
Worst	P12	P12
	Rof	Rof
	Env	Env

Table 5.11: The rank order of targeting HTLV-I proteins according to their potential to reduce the risk of HAM/TSP (first column) and reduce proviral load (second column).

5.3.7 The Prevented Fraction of Disease, $\mathbf{F_P}$

We calculated the prevented fraction of disease attributable to the possession of one or more strong binding alleles to HBZ [1] (see Methods, Section 5.2.10). This showed that the possession of strong HBZ-binding HLA alleles prevented (F_P) $\sim$ 48% (SD 12.3%) of potential cases of HAM/TSP in the study population. However, the strength of HBZ binding is not the only determinant of disease status: in a logistic regression model, the strength of HBZ binding alone could only predict 55% of cases of HAM/TSP.

5.3.8 HBZ Specific CD8$^+$ T Cells can be Detected *ex vivo*

This work strongly implies that HBZ-specific CD8$^+$ T cells play a protective role in HTLV-I infection. HBZ immunogenicity has been studied in ATL patients [155, 158] but it is unknown whether a HBZ-specific CD8$^+$ T cell response is generated or even whether HBZ protein is expressed in asymptomatic carriers and HAM/TSP patients. We therefore sought to identify HBZ-specific CD8$^+$ T cells in PBMCs from HTLV-I infected individuals. We assayed IFN-γ production by ELISpot following stimulation in vitro with a pool of overlapping peptides that spanned the entire HBZ protein. Of 45

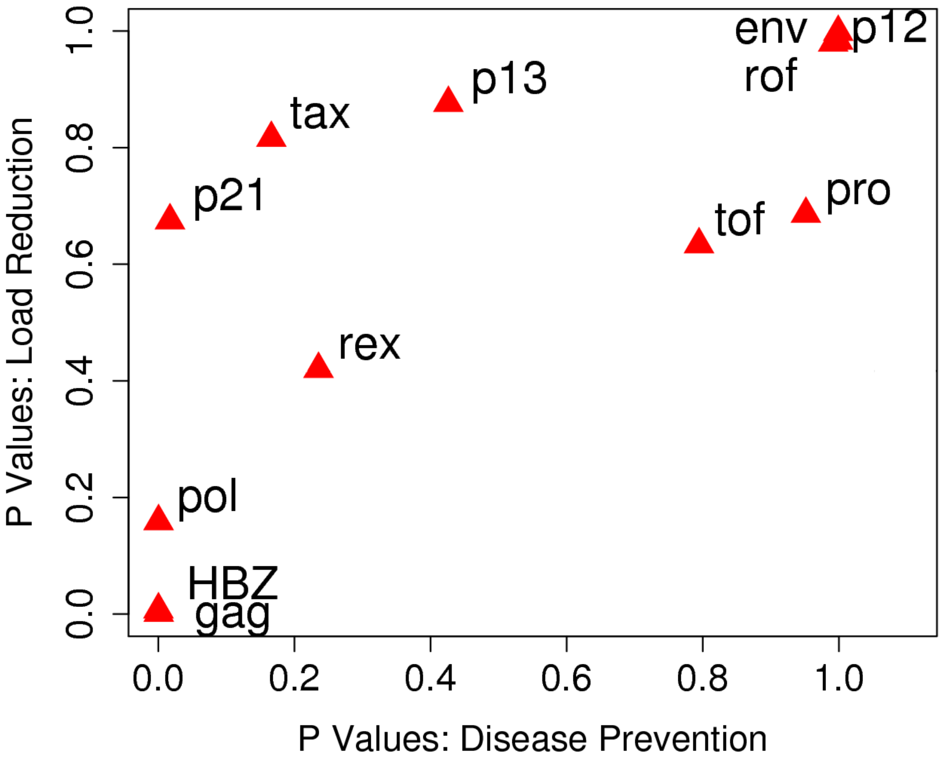


FIGURE 5.8: The correlation between the P values of the 1-tailed hypotheses: targeting this protein is associated with a lower proviral load and targeting this protein is associated with a lower HAM/TSP prevalence ($R_S = 0.86$, $P = 0.0005$). In addition to HBZ, Gag also produced significant results in this analysis (it was significantly associated with a lower HAM/TSP prevalence ($P = 0.0005$) and a lower proviral load ($P = 0.002$)). However, we did not focus on this result as it was not repeated independently with Epipred.

subjects tested, 31% had detectable HBZ-specific CD8$^+$ T cells. We conclude that HBZ protein is expressed in vivo and is immunogenic.

5.3.9 The Comparative Immunogenicity of HBZ and Tax

How does the immunogenicity of HBZ compare to Tax? We compared the predicted top binding peptide from HBZ and Tax to 43 alleles (the allele capacity of Metaserver). Peptides from Tax bind significantly more strongly than peptides from HBZ ($P = 0.00002$, paired Wilcoxon-Mann-Whitney; Figure 5.10, panel A). Consistent with this, the frequency of Tax-specific CD8$^+$ T cells by IFN-γ ELISpot was also greater than the frequency of HBZ-specific CD8$^+$ T cells in the 45 HTLV-I infected individuals ($P = 0.000006$, paired Wilcoxon-Mann-Whitney; Figure 5.10, panel B).

	Null hypothesis	Rank measure		Raw score		Conclusion
		Metaserver	Epipred	Metaserver	Epipred	
1	Protective and detrimental alleles target HBZ equally	0.0002		-	-	Protective alleles bind HBZ significantly more strongly than detrimental alleles
2	AC and HAM/TSP patients target HBZ equally	0.0002	0.000002	0.002	0.001	ACs have HLA alleles that bind HBZ significantly more strongly compared to HAM/TSP patients
3	AC and HAM/TSP patients target HBZ equally (excluding A02, B54 and Cw08)	0.04	0.006	0.14	0.03	ACs bind HBZ significantly more strongly compared to HAM/TSP patients even when known protective and detrimental alleles are excluded
4	There is no correlation between proviral load and the number of alleles that bind HBZ strongly	0.016	0.1	0.01	0.032	The higher the number of strong binding alleles to HBZ per individual, the lower their proviral load
5	There is no correlation between proviral load and the strength of HBZ binding	0.008	0.04	0.003	0.085	The greater the strength of HBZ binding (rank method), the lower the proviral load
6	There is no correlation between load reduction (count) and disease prevalence reduction	0.0005	0.02	0.004	0.03	Proteins that are strongly bound by asymptomatic carriers are, independently, those associated with a greater reduction in load when bound
7	There is no correlation between load reduction (rank) and disease prevalence reduction	$< 2.2 \times 10^{-16}$	0.003	0.002	0.2	As above, using the rank measure to quantify the effect of binding strength on proviral load

Table 5.12: The results of hypothesis testing repeated using different epitope prediction methods (Metaserver and Epipred) and different metrics (a rank measure which only compares within alleles (i.e. not between alleles) and a raw binding score measure which compares between as well as within alleles).

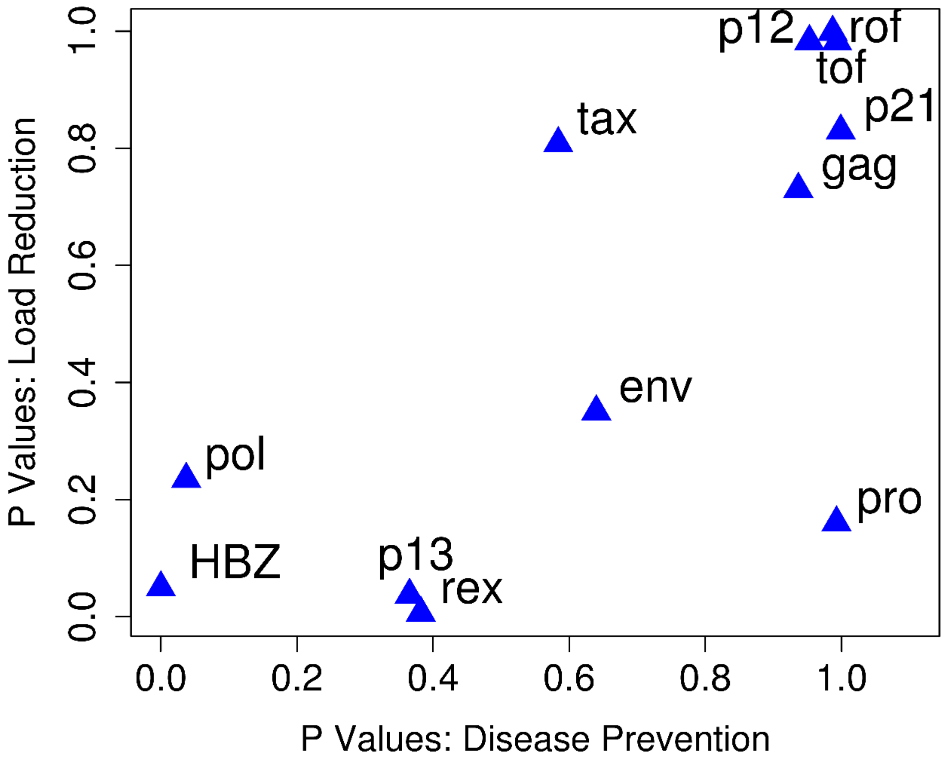


FIGURE 5.9: The correlation between the P values of the 2 hypotheses: targeting this protein is associated with a reduction in HAM/TSP prevalence and targeting this protein is associated with a reduction in proviral load ($R_S = 0.66$, $P = 0.02$). Epitope binding predictions made using Epipred.

5.4 Discussion

Using validated epitope prediction software, we show that strong binding of peptides from the HTLV-I basic leucine zipper factor (HBZ) protein is associated with a reduced risk of HAM/TSP and a reduced proviral load in a population with endemic HTLV-I infection in southern Japan. We demonstrated that protection is not limited to a small subset of HLA class I alleles previously associated with disease status and proviral load (HLA-A*02 and HLA-Cw*08), but is more generally associated with HLA class I alleles that bind strongly to HBZ.

Prior to this analysis CD8$^+$ T cells specific for the HTLV-I protein Tax were often considered as the best candidate for 'efficient' or 'protective' CD8$^+$ cells because of the immunodominance of Tax in the CD8$^+$ T cell response [8, 11]. Our finding that binding of HBZ peptides rather than Tax peptides is protective raises the question why HBZ?

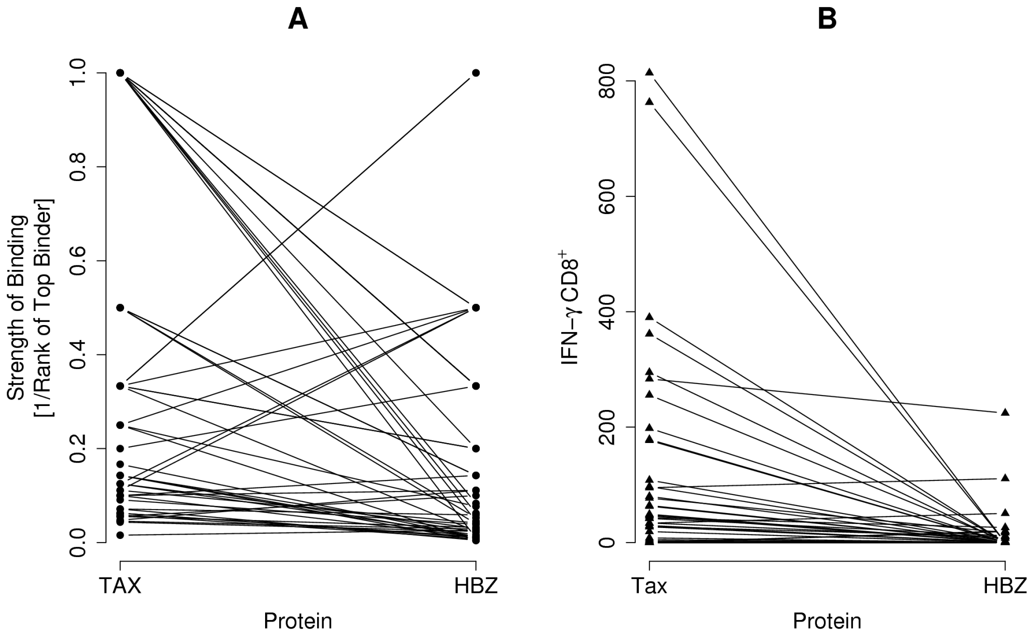


FIGURE 5.10: The comparative immunogenicity of HBZ and Tax. A, The predicted top binding peptide from Tax and HBZ to each of the 43 alleles for which Metaserver predicts binding affinities was found. Peptides from Tax are bound more strongly than peptides from HBZ ($P = 0.00002$, paired Wilcoxon-Mann-Whitney). B, Consistent with this, the frequency of Tax-specific CD8$^+$ T cells was also greater compared to HBZ CD8$^+$ T cells in the 45 HTLV-I infected individuals tested using by IFN-γ ELISpot ($P = 0.000006$, paired Wilcoxon-Mann-Whitney).

The HBZ gene was identified recently [159]. It is encoded by the complementary strand of the HTLV-I genome and its promoter lies in the 3 LTR rather than the 5 LTR. It functions by binding to the transcription factor CREB-2. There are two major splice variants of the HBZ transcript, SP1 and SP2; the variant SP1 is more abundant and is the variant used in this study [160]. Expression of HBZ suppresses Tax-mediated transactivation through the 5 LTR [159, 161] and thereby inhibits expression of other HTLV-I genes [159, 162]; HBZ can be expressed in the absence of transcription of other HTLV-I genes. Additionally, HBZ RNA promotes the proliferation of infected T-lymphocytes [155]. This dual action - reduction of HTLV-I expression and subsequent protection from immune surveillance, and enhancement of infected cell proliferation - probably confers a survival advantage on HBZ-expressing cells and is consistent with the observations that HBZ enhances persistence in HTLV-I inoculated rabbits [162] and that ATL cells often have a hypermethylated or deleted 5 LTR but an intact functional 3 LTR [155].

We hypothesise that if HBZ-specific CD8$^+$ T cells are weak or absent then infected cells that express HBZ but not other viral proteins will escape immune surveillance and proliferate rapidly, leading to a large increase in proviral load. HBZ-specific CD8$^+$ T cells would then play an important role in preventing this proliferation of provirus-positive cells and blocking this strategy of persistence. If this conclusion is correct that HLA class I recognition of HBZ plays a central role in the control of HTLV-I replication than one might expect that HBZ in HTLV-I would have evolved to minimize the effect of this class I recognition. Consistent with this hypothesis, we find that the predicted binding affinity to HBZ peptides is significantly weaker than that of Tax peptides and that the frequency of HBZ-specific CD8$^+$ T cells is significantly lower than the frequency of Tax-specific CD8$^+$ T cells. Although the low immunogenicity of HBZ is precisely what we predict given its central importance in maintaining HTLV-I persistence it is nevertheless striking that these low frequency responses are so important. This result challenges the prevailing assumption in HTLV-I research and in immunology in general that the immunodominant responses are the most interesting and important.

This approach to studying the association between HLA class I genotype and the outcome of infection has a number of strengths compared with a traditional frequency-based analysis. Firstly, it is more mechanistic: knowing that binding HBZ is associated with a reduced proviral load and disease risk compared with knowing that A*02 is associated with these outcomes is a simultaneously more fundamental and more applicable level of understanding. Secondly, identification of protective epitopes immediately suggests a practical approach to increase the efficiency of an individuals anti-viral response. Thirdly, because the same effect (e.g. HBZ binding) can be identified for many alleles it is less likely to be a spurious result of linkage disequilibrium or genetic stratification. Finally, effects due to multiple low-frequency alleles can be captured because analysis is made at the level of peptide binding rather than allelic frequency.

In summary, using a novel and generalizable approach, we have identified one of the constituents of an effective CD8$^+$ T cell response in HTLV-I infection.

Chapter 6

Antigen Expression as a Determinant of CTL Lysis in HTLV-I Infection

6.1 Introduction

So far, this study has concentrated on the epitope properties that are associated with disease risk and proviral load. For this chapter, we extended this work to examine the CD8$^+$ T cell response itself, specifically in terms of its dynamics as a function of antigen expression of infected CD4$^+$ T cells.

In human T-lymphotropic virus type 1 (HTLV-I) infection, a high frequency of HTLV-I-specific CTLs can co-exist stably with a high proviral load and the proviral load is strongly correlated with the risk of HTLV-I-associated inflammatory diseases. In Chapter 2, Section 2.3.3, it was discussed how these observations have led to the hypothesis that HTLV-I specific CTLs are ineffective in controlling HTLV-I replication but contribute to the pathogenesis of the inflammatory diseases. However, evidence from host and viral immunogenetics and gene expression microarrays suggests that a strong CTL response is associated with a low proviral load and a low risk of HAM/TSP. To further examine the role of CTLs in HTLV-I infection, a collaborative experiment was carried out to quantify the frequency, lytic activity and functional avidity of HTLV-I-specific CD8$^+$ cells in fresh, unstimulated PBMCs from individuals with natural HTLV-I infection. The ELISpot assays and flow cytometry experiments were performed by Aileen Rowan and Tarek Kattan. My role, using a system of ordinary differential equations, was to quantify the efficiency of CD8$^+$ lysis in their *ex vivo* experiments.

We have previously investigated methods of quantifying antiviral CTL efficiency. In HTLV-I infection, both effector CTLs and infected target cells are often present in fresh blood at frequencies sufficiently high to obviate the need for enrichment of specific subpopulations. We have exploited this feature to develop an assay of CD8$^+$ cell-mediated suppression of HTLV-I expression in fresh PBMCs [87]. As a marker of proviral expression we use the viral protein Tax, a regulatory protein expressed early in the life cycle of HTLV-I [87]. We previously showed that this suppression of HTLV-I depended on CD8$^+$ T cell frequency and required both perforin and a match in class 1 MHC genotype between effector and target cells [163], consistent with classical class 1 MHC-restricted CTL lysis. Mathematical modeling can be used to quantify the rate of killing of Tax-expressing CD4$^+$ cells per CD8$^+$ cell per day. We use the term lytic efficiency to denote this per-CD8$^+$-cell rate of lysis. This assay of lytic efficiency showed that the rate of CTL-mediated lysis of HTLV-I-infected cells in fresh PBMCs was inversely correlated with the proviral load, both in patients with HAM/TSP and in asymptomatic HTLV-I carriers (ACs) [87].

This measure of lytic efficiency has two chief limitations. First, the antiviral activity is expressed per CD8$^+$ cell, not per virus-specific CD8$^+$ cell, since there is no currently available method to measure in the same assay the lytic activity and the total frequency of CD8$^+$ T cells specific to all viral epitopes in each individual (an ELISpot assay that detects IFNγ production comes closest to this definition. However, this assay only detects the activation of CD8$^+$ cells and not their lytic efficiency). Second, rate of lysis is likely to be a composite parameter that is a function of both the frequency of the Ag specific CD8$^+$ cells and the "quality" of their effector functions at the single-cell level [164] (i.e. it is not clear what effector functions might affect the lysis rate). In an acute viral infection, efficient elimination of the virus is associated with a high frequency of Ag-specific T cells [165]. But in persistent infections, the complexity of the equilibrium dynamics makes it impossible to infer the efficiency of virus-specific CTLs directly from their steady-state frequency [166].

For this chapter, I modified this model of lytic efficiency to take account of the observation that Tax expression per cell is not constant in the HTLV-I-infected target CD4$^+$ cells, but increases over the 18 hour incubation time. The results of this analysis were estimates of the lytic efficiency of CTLs specific to high Tax expressing cells and, separately, low Tax expressing cells.

6.2 Methods

The cell preparation, culture and flow cytometry analysis was carried out by Tarek Kattan.

6.2.1 PBMC Separation

Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood by density gradient centrifugation using Histopaque-1077 (Sigma, Poole, United Kingdom) from EDTA-anticoagulated blood samples taken from HTLV-I infected individuals. All individuals attended the HTLV-I clinic at St Marys Hospital, London and gave written informed consent. Isolated PBMCs were washed twice in PBS and then cryopreserved in fetal calf serum (FCS, Sigma) with 10% dimethyl sulfoxide (DMSO, Sigma).

6.2.2 Cell Culture

Cells were thawed, washed twice in PBS and then cultured in complete medium consisting of RPMI 1640 medium (Sigma) supplemented with 10% FCS, 2 mM L-Glutamine, 100 U/ml Penicillin and 100μg/ml Streptomycin (Life Technologies). Cells were incubated for different times at 37 °C in 5% CO_2. When required, $CD8^+$ or $CD4^+$ cells were depleted by positive selection using Ab coated magnetic microbeads following the manufacturer's instructions (Miltenyi Biotec, Surrey, United Kingdom).

6.2.3 Flow Cytometric Detection of Tax Expression

After incubation, cells were surface-stained with mAbs specific to CD4 and CD8 at 15μg/ml in each case (Beckman Coulter, Marseille, France). Cells were fixed with 2% paraformaldehyde (PFA, Sigma) and then permeabilized using PBS/0.1% Triton X-100 (Sigma). Finally, cells were stained intracellularly with the FITC conjugated Ab anti-Tax protein Lt-4 [167], diluted 1/100. Cells were analyzed on a Coulter Epics XL flow cytometer. Thirty thousand events were routinely collected during acquisition of the data. The data were analyzed using Coulter Expo32 software (Beckman Coulter). Tax^+cells were divided in flow-cytometric analysis into two gates, corresponding respectively to Tax^{low} and Tax^{high} cells according to fluorescence intensity. The line dividing these gates was arbitrarily defined; the same definition was used in the analysis of all samples.

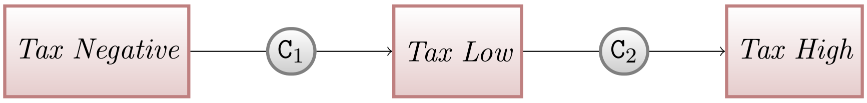
6.2.4 Rate of CD8$^+$ Cell-mediated Lysis

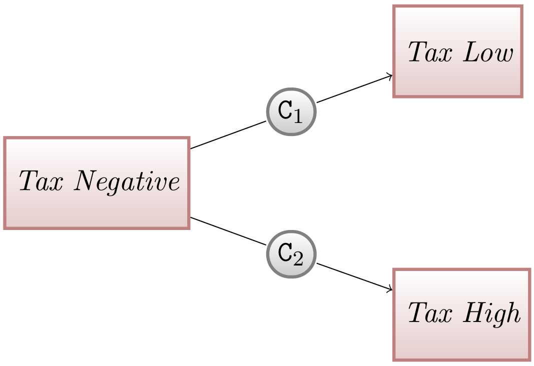
The rate ("efficiency") of CD8$^+$ cell-mediated lysis of HTLV-I-infected cells was estimated as previously described [87]. CD8$^+$ cell lytic efficiency (expressed as the proportion of Tax-expressing CD4$^+$ cells killed per CD8$^+$ cell per day) was calculated for each HTLV-I-infected individual tested using Equation 6.1:

$$\frac{dy}{dt} = c - \epsilon y z \tag{6.1}$$

where y is the proportion of CD4$^+$ cells expressing Tax, c is the rate of increase of Tax expression, which is assumed to be constant during the short-term culture, ϵ is the CD8$^+$ cell-mediated antiviral efficacy (expressed as the proportion of CD4$^+$ Tax$^+$ cells killed per CD8$^+$ cell per day) and z is the proportion of lymphocytes that are CD8$^+$. This model was solved analytically and fitted to the data using non-linear least-squares regression (SPSS v12), providing an estimate of the antiviral efficacy (ϵ) in each individual.

Equation 6.1 used a constant rate c to describe Tax expression in the absence of CD8$^+$ cells. However, since observations from experimental time course data (Figure 6.1) suggested a non-linear rate of increase of Tax expression, it was necessary to modify the existing model. 2 models were considered as possibilities:





The first model represents a progression from Tax$^{\text{negative}}$ to Tax$^{\text{low}}$ to Tax$^{\text{high}}$ expressing CD4$^+$ cells as a single population. The second models the appearance of Tax$^{\text{low}}$ and Tax$^{\text{high}}$ expressing CD4$^+$ cells as 2 distinct populations. Instead of a single parameter c that defines the rate of increase of Tax expression in Equation 6.1, each model introduces

2 rate parameters: c_1, the rate of increase of low Tax expression and c_2, the rate of increase of high Tax expression. Both models were fitted to the time course data and it was found that the 'single population' model was a more accurate fit of the data (see Section 6.3.1). The 'single population' model was designated 'Model 1'.

In Model 1, the Tax$^{\text{low}}$ population (as defined from the gated FACS) is produced at a constant rate c_1 and the Tax$^{\text{high}}$ population at a rate c_2. The following pair of linked ordinary differential equations, Equation 6.2 and Equation 6.3, describe the model:

$$\frac{dy}{dt} = c_1 - c_2 y \tag{6.2}$$

$$\frac{dw}{dt} = c_2 y \tag{6.3}$$

Here, y is the proportion of CD4$^+$ cells expressing low levels of Tax and w is the proportion of CD4$^+$ cells expressing high levels of Tax. Solving these equations, we have Equation 6.4 and Equation 6.5:

$$y = \left(\frac{c_1}{c_2}\right)\left(1 - e^{-c_2 t}\right) \tag{6.4}$$

$$w = c_1 \left(t + \frac{e^{-c_2 t}}{c_2} - \frac{1}{c_2}\right) \tag{6.5}$$

Equation 6.4 and Equation 6.5 were fitted to the data using non-linear least-squares regression (Table 6.2), providing an estimate for c_1 and c_2 in each individual. Equation 6.2 and Equation 6.3 were then modified to describe the rate of CD8$^+$ cell-mediated lysis of Tax$^{\text{low}}$ CD4$^+$ cells and Tax$^{\text{high}}$ CD4$^+$ cells separately:

$$\frac{dy}{dt} = c_1 - c_2 y - \epsilon^{\text{low}} yz \tag{6.6}$$

$$\frac{dw}{dt} = c_1 \left(1 - \frac{1}{e^{c_2 t}}\right) - \epsilon^{\text{high}} wz \tag{6.7}$$

Equation 6.6 and Equation 6.7 (Model 2) were solved analytically and fitted to the data using non-linear least-squares regression. From the resulting data, estimates for the rate of killing of CD4$^+$ cells expressing low levels of Tax (ϵ^{low}) and high levels of Tax (ϵ^{high}) were produced for each individual.

6.2.4.1 Bootstrap

A bootstapping method was used to test the robustness of the ϵ^{low} and ϵ^{high} parameters for each patient. 50 new datasets were generated per individual for both their expression levels of Tax$^{\text{low}}$ and Tax$^{\text{high}}$. Equation 6.6 and Equation 6.7 were then fitted to this data, which gave a standard deviation of the resulting values of ϵ^{low} and ϵ^{high} (see Table 6.2).

6.3 Results

6.3.1 Modeling Tax expression

Flow cytometric analysis was used to divide the Tax-expressing CD4$^+$ population into high Tax-expressing and low Tax-expressing cells. The 2 models of Tax expression, 'single' and 'dual' populations, were solved and fitted to the experimental data to give estimates of the parameters c_1 and c_2 in each case. Table 6.1 shows the R^2 values for the fits of the respective models for each of the patients. The accuracy of the 'single population' model was significantly better for low Tax expression and better for high Tax expression (Tax low: $P = 0.0001$, Tax high, $P = 0.095$; paired Wilcoxon-Mann-Whitney). From these results, we chose the 'single population' as our model of Tax expression over 18 hours.

	Names	Single Low	Single High	Dual Low	Dual High
1	HAP	0.4571	0.8297	0.1126	0.8575
2	HAY	0.8710	0.8408	0.8586	0.1214
3	HBE	0.2665	0.4938	-0.2782	0.4277
4	HBX	0.7376	-1.4986	0.6609	0.2523
6	HCH	0.1675	0.4303	-0.2767	0.7503
8	HFB	-1.0925	0.7753	-2.6407	0.7978
11	TAK	0.8165	0.4660	0.8179	0.0435
12	TAQ	0.9461	0.8541	0.9361	0.7629
13	TAW	-0.0586	-0.2140	-0.5068	0.4870
14	TAZ	0.8401	0.7504	0.8006	0.0256
15	TBC	0.6124	0.9268	-0.2076	-0.7341
16	TBJ	0.7501	0.6986	0.5346	-0.0517
17	TBP	0.7397	0.9457	0.3094	-0.1896
19	TCL	0.6367	0.5781	0.3577	0.1122
20	UV1	-0.1119	0.6705	-0.7201	-1.6188

Table 6.1: The R^2 values showing the accuracy of the single and dual population models fitting the experimental time course data of Tax$^{\text{low}}$ and Tax$^{\text{high}}$ expression.

Equation 6.2 and Equation 6.3 were then fitted to the Tax$^{\text{low}}$ and Tax$^{\text{high}}$ data respectively by non-linear least squares regression. Values for the parameters c_1 and c_2 were calculated from this analysis for each of the 15 patients (Table 6.2). Examples for 6 of the patients are shown in Figure 6.1. The data for the other 9 patients is in Appendix C, Figure C.1.

6.3.2 The CD8$^+$ Antiviral Efficacy Assay

Equation 6.6 and Equation 6.7 were fitted to the antiviral efficacy assay data for all 15 patients. The data shown for each patient is the proportion of CD4$^+$ lymphocytes that were Tax$^{\text{high}}$ and Tax$^{\text{low}}$ following 18 h co-culture with different proportions of CD8$^+$ lymphocytes. The parameters ϵ^{low} and ϵ^{high} were measured by non-linear least squares regression. The assay was repeated in each patient. Figure 6.2 shows this data for 3 patients. The data for the other 12 patients is in Appendix C, Figure C.2.

Figure 6.3 shows a statistically significantly higher rate of CD8$^+$ cell-mediated lysis of the Tax$^{\text{high}}$ cells than that of the Tax$^{\text{low}}$ cells ($P = 0.004$, Wilcoxon-Mann-Whitney, $n = 15$). Figure 6.4 shows the plot of ϵ^{low} against ϵ^{high} for each patient. The strong linear relationship ($R^2 = 0.855$, $P < 0.001$) suggests the ratio $\epsilon^{\text{high}}/\epsilon^{\text{low}}$ is maintained across patients.

There was strong agreememt between the 2 repeats per patients of the estimates for ϵ^{low} and ϵ^{high} (ϵ^{low}: $R^2 = 0.9492$, $P < 0.001$. ϵ^{high}: $R^2 = 0.890$, $P < 0.001$; Appendix C, Figure C.2). Appendix C also contains a comparison of the original ϵ and c parameters from Equation 6.1 against ϵ^{low} and ϵ^{high} (Figure C.4) and c_1 and c_2 (Figure C.5).

Finally, Appendix C, Figure C.6 shows there was no difference in the ratio of Tax expression between HAM/TSP patients and ACs ($P = 0.871$, Wilcoxon-Mann-Whitney, HAM/TSP: $n = 16$, AC: $n = 12$).

6.4 Discussion

CD8$^+$ T cells have been shown in vitro to require only 10 complexes of MHC/peptide to elicit a lytic effector response [168–170]. The detection of a significant difference in the rate of CTL-mediated lysis between cells with high Tax expression and those with low Tax expression was therefore surprising, since even the low Tax cells contain sufficient Tax protein to be readily detected by flow cytometry. Inefficient lysis might be caused by inefficient Ag processing, which would result in turn in few MHC/Tax peptide complexes being presented on the infected cell surface, despite the high level

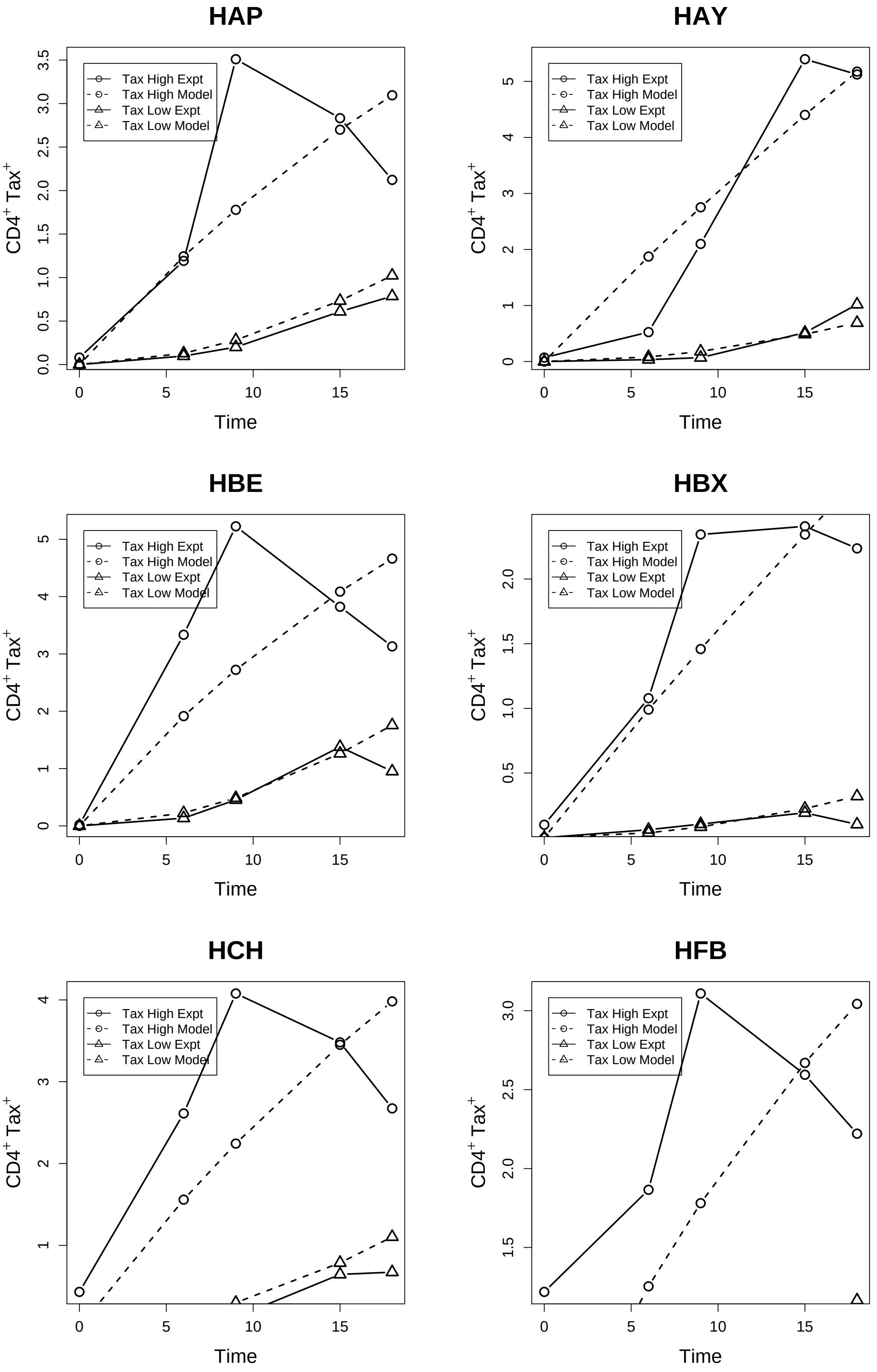


FIGURE 6.1: This figure shows examples of the time course of Tax expression as the proportion of CD4$^+$ lymphocytes that were Tax$^{\text{high}}$ or Tax$^{\text{low}}$ over 18 hours culture (see Section 6.2.2). Equation 6.2 and Equation 6.3 were fitted to the Tax$^{\text{low}}$ and Tax$^{\text{high}}$ data respectively.

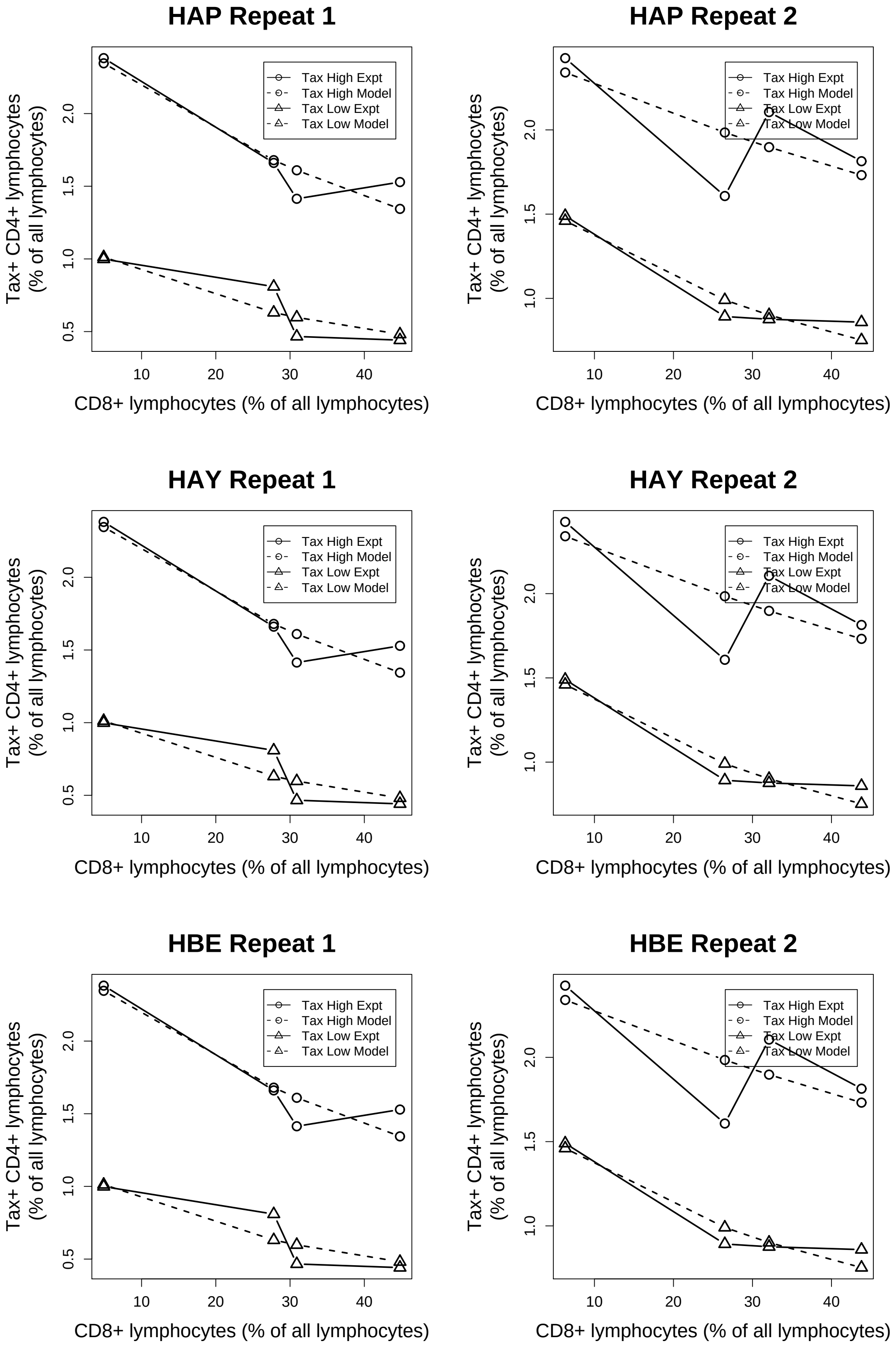


FIGURE 6.2: The figure shows examples of the antiviral efficacy assay for patients HAP, HAY and HBE. The proportion of CD4$^+$ lymphocytes that were Tax$^{\text{high}}$ and Tax$^{\text{low}}$ following 18 h co-culture with different proportions of CD8$^+$ lymphocytes was measured. The model (Equation 6.6 and Equation 6.7, Section 6.2.4) was fitted to this data and in this way the rate of clearance of Tax$^{\text{low}}$CD4$^+$ and Tax$^{\text{high}}$CD4$^+$ cells per day per CD8$^+$ cell (antiviral efficacy) was estimated. This was repeated in the same subject.

	Names	C_1	C_2	ϵ Low 1	ϵ Low 2	ϵ High 1	ϵ High 2
1	HAP	5.49168	0.80208	0.06325	0.03579	0.08192	0.04930
2	HAY	7.82112	0.34128	0.00998	0.01941	-0.00819	0.02481
3	HBE	8.55912	0.9036	0.10483	0.10640	0.41622	0.41790
4	HBX	4.1112	0.29808	0.06160	0.06127	0.13806	0.12907
6	HCH	6.78072	0.68136	0.07576	0.07223	0.24124	0.18569
8	HFB	5.61096	0.9168	-0.00280	-0.01707	-0.02105	0.02714
11	TAK	4.80936	0.07488	0.03289	0.02731	0.05000	0.07240
12	TAQ	3.87672	0.14112	0.03150	0.05664	0.04257	-0.13534
13	TAW	4.28592	0.86304	0.19241	0.18434	0.53716	0.59616
14	TAZ	12.6408	0.27312	0.03840	0.02830	0.07164	0.04612
15	TBC	53.83248	1.08096	0.00462	0.00097	0.03455	0.03117
16	TBJ	14.04216	0.54552	0.15732	0.13589	0.42991	0.31358
17	TBP	18.14064	0.87336	-0.00313	0.00972	0.03025	0.02982
19	TCL	2.9952	0.5952	0.05158	0.05430	0.11593	0.10564
20	UV1	2.5392	1.04928	-0.02919	-0.01079	0.04658	-0.00212

	Names	ϵ Low Avg	ϵ High Avg	ϵ Low 1 SD	ϵ Low 2 SD	ϵ High 1 SD	ϵ High 2 SD
1	HAP	0.04952	0.06561	0.02272	0.02713	0.06017	0.02021
2	HAY	0.01469	0.00831	0.0178	0.02471	0.01251	0.04833
3	HBE	0.10561	0.41706	0.0623	0.13736	0.30864	0.28325
4	HBX	0.06143	0.13356	0.01437	0.0671	0.06328	0.12195
6	HCH	0.07399	0.21347	0.0375	0.04678	0.18174	0.02042
8	HFB	-0.00993	0.00304	0.01811	0.02293	0.03427	0.04355
11	TAK	0.0301	0.0612	0.01943	0.01747	0.05339	0.03249
12	TAQ	0.04407	-0.04639	0.01176	0.13824	0.05061	0.10499
13	TAW	0.18837	0.56666	0.0659	0.07396	0.11129	0.32662
14	TAZ	0.03335	0.05888	0.03852	0.0228	0.07298	0.04592
15	TBC	0.0028	0.03286	0.0101	0.00806	0.02315	0.02376
16	TBJ	0.1466	0.37175	0.04261	0.02233	0.19724	0.08156
17	TBP	0.0033	0.03004	0.01462	0.00265	0.01026	0.0256
19	TCL	0.05294	0.11078	0.09165	0.11836	0.13564	0.17345
20	UV1	-0.01999	0.02223	0.0046	0.04018	0.11681	0.05295

TABLE 6.2: The summary statistics for the model of CTL-mediated lysis efficiency against $CD4^+$ cells expressing either high or low levels of Tax. Each row gives the statistics per patient. The columns from left to right: C_1 and C_2 refer to the estimated rates of change per 24 hours between $\text{Tax}^{\text{negative}}$ and Tax^{low} and Tax^{low} and Tax^{high} respectively. ϵ low 1 and 2 and ϵ high 1 and 2 give the values of ϵ derived from the model for the 2 repeats of low Tax and high Tax expressing cells respectively. ϵ low average and ϵ high average are the respective means of the 2 repeats. The next 4 columns show the standard deviation values derived from the estimates of ϵ for the 2 repeats of low and high. These were calculated using a bootstrap method (see Section 6.2.4.1).

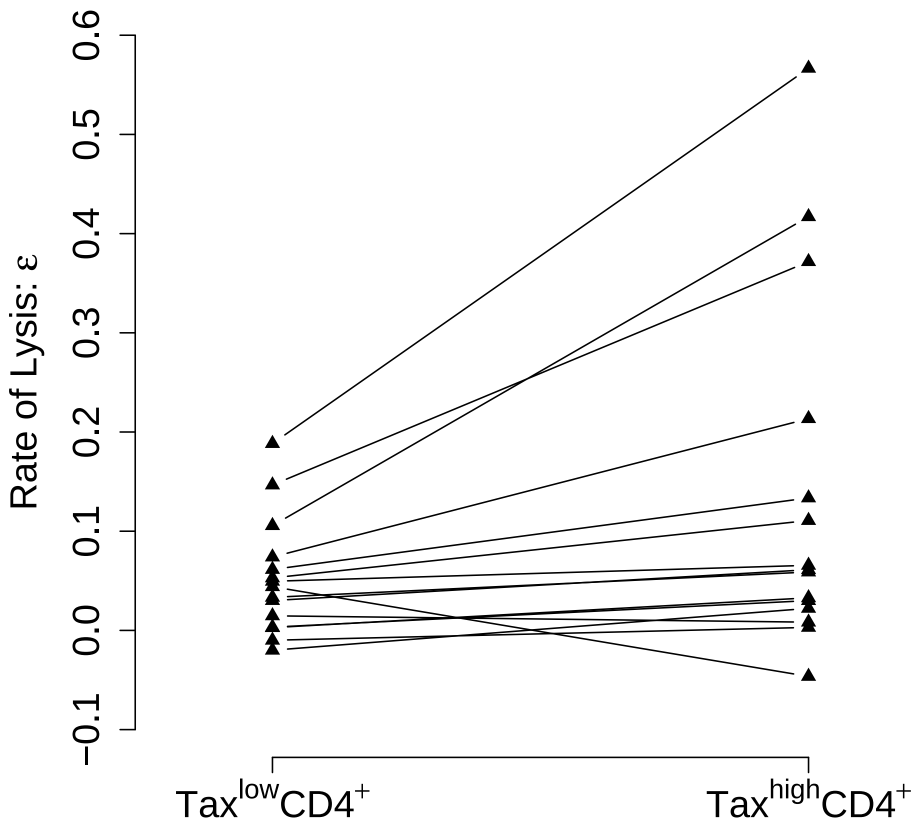


FIGURE 6.3: The rates of lysis ϵ^{low} and ϵ^{high} compared per patient. Tax$^{\text{high}}$ cells were killed faster than Tax$^{\text{low}}$ cells in the same individual ($P = 0.004$, Wilcoxon-Mann-Whitney, $n = 15$).

of intracellular Tax protein as indicated by intracellular staining. Alternatively, it is possible that there is not a uniformly high probability of CTL-mediated lysis when a low threshold of MHC/peptide density on the cell surface is exceeded, but rather that the probability of CTL-mediated lysis increases progressively with increasing density of MHC peptide complexes. Finally, it is possible that despite the immunogenicity of Tax protein in the CD8^{+} T cell response [11], recognition of another HTLV-I Ag by CTLs might be the factor that limits the rate of HTLV-I replication in vivo [7, 171].

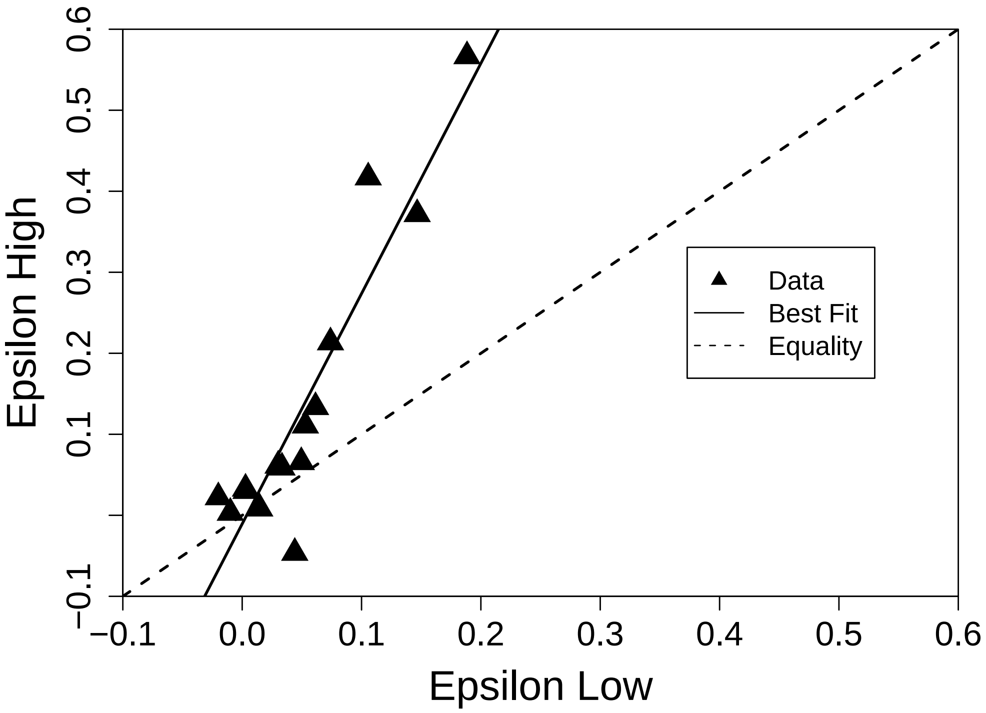


FIGURE 6.4: A comparison of the rates of lysis ϵ^{low} and ϵ^{high}. The data represents the average ϵ^{low} and ϵ^{high} from the 2 repeats per patient. The was a strong linear correlation across all patients for this ratio ($R^2 = 0.855$, $P < 0.001$).

	Names	R^2 TC Low	R^2 TC High	R^2 High 1	R^2 High 2	R^2 Low 1	R^2 Low 2	C Rep 1	C Rep 2	ϵ Rep 1	ϵ Rep 2
1	HAP	0.45711	0.82968	0.76929	0.92048	0.86973	0.47586	4.817	5.454	0.047	0.034
2	HAY	0.87095	0.84082	0.3946	0.29511	0.43637	0.46641	8.705	9.021	0.008	0.021
3	HBE	0.26648	0.49379	0.80528	0.93747	0.82434	0.83602	6.585	13.191	0.081	0.2
4	HBX	0.7376	-1.49857	0.96126	0.78407	0.95216	0.66433	4.552	4.294	0.077	0.067
6	HCH	0.16752	0.43028	0.55425	0.97523	0.78886	0.55311	4.052	4.09	0.037	0.032
8	HFB	-1.09248	0.77528	0.12536	0.39027	0.00464	0.1863	3.106	3.617	-0.015	-0.002
11	TAK	0.81652	0.46595	0.78497	0.95456	0.82724	0.85024	5.798	5.664	0.04	0.035
12	TAQ	0.94614	0.85412	0.8894	0.14617	0.913	0.51971	5.63	5.78	0.045	0.044
13	TAW	-0.05858	-0.21404	0.98371	0.89219	0.9684	0.92851	10.931	3.999	0.452	0.161
14	TAZ	0.84013	0.75042	0.64513	0.51961	0.46895	0.63307	14.711	12.33	0.044	0.027
15	TBC	0.61237	0.92676	0.87283	0.76828	0.089	0.00652	51.142	52.394	0.013	0.01
16	TBJ	0.75008	0.69859	0.86004	0.88222	0.72689	0.83239	4.94	5.435	0.041	0.036
17	TBP	0.73966	0.94565	0.94029	0.95061	0.06192	0.87595	24.185	23.942	0.014	0.022
19	TCL	0.63667	0.57808	0.89715	0.73001	0.64048	0.75114	3.351	3.81	0.059	0.066
20	UV1	-0.11191	0.67047	0.43589	0.00057	0.88353	0.02592	2.216	1.988	0.008	-0.004

Table 6.2: Continued: The R^2 values for fits of the model against the experimental data are shown for the time course model (R^2 TC low and R^2 TC high) and the 2 repeats in the lysis model (R^2 high and R^2 low). Finally the original values of the C constant and ϵ are shown for 2 repeats per patient (C Rep 1, C Rep 2, ϵ Rep 1 and ϵ Rep 2).

Chapter 7

The KIR Gene Cluster and HTLV-I

7.1 Introduction

Natural killer (NK) cells are critical components of the innate immune system that have direct involvement in the anti-viral immune response [172]. In addition to direct cytotoxic and antiviral functions, they have the potential to interact with components of the adaptive immune system, including T cells and dendritic cells [173]. This interaction implies a broad role in immunity and potential involvement in a wide range of diseases, including infections, cancers, and autoimmune disorders. NK cells are controlled by many activating and inhibitory receptors [174, 175]. In humans, one of the key receptor families contributing to the NK cell receptor repertoire is the killer cell immunoglobulin (Ig)-like receptor (KIR) family. They were first identified by their ability to impart some specificity on natural killer (NK) cytolysis [176, 177]. Similar to many NK cell receptors, KIRs are expressed on T cells in addition to NK cells, affirming their role in adaptive immunity.

There is extensive diversity at the KIR gene locus, which stems from both its polygenic and multi-allelic polymorphism [178]. KIR gene expression patterns can vary clonally [179], adding yet another layer of complexity to the system. Diversity at the locus may be the result of selection pressures, in a manner analogous to that proposed for the HLA class I loci. Overall, however, KIR genes can be classified as activating or inhibitory. Combinations of these genes occur to generate haplotypes with widely differing balances between activating and inhibitory types.

In this chapter, we tested the hypothesis that certain KIR-HLA associations are beneficial or detrimental regarding disease status and proviral load in HTLV-I infection. This was based on the model that KIRs synergize with HLAs to generate compound genotypes that provide different levels of activation and inhibition, which impacts viral control.

7.2 Methods

7.2.1 Database

The Kagoshima dataset (see Section 3.2.1) provided information on the presence or absence of expression of the KIR genes of Table 7.1.

7.2.2 Tested Associations

For each HAM/TSP patient ($n = 230$) and AC ($n = 202$) in the dataset, we tested for the presence of each KIR and its associated HLA ligand (Table 7.1). The total number of inhibitory and activating pairs were counted per individual and tested against proviral load separately for both HAM/TSP and AC groups.

Inhibitory					Activating			
2DL1	2DL2	2DL3	3DL1	3DL2	2DS1[1]	2DS2	2DS4	3DS1
C02 C04 C05 C06	C01 C03 C07 C08	C01 C03 C07 C08	B08 B13 B27 B44 B51 B52 B53 B57 B58	A03 A11	C02 C04 C05 C06	C01 C03 C07 C08	C04	B08 B13 B27 B44 B51 B52 B53 B57 B58

Table 7.1: A summary of KIR ligand specificity. Each individual in the Kagoshima dataset was labelled yes/no for the expression of the KIR alleles. For both the inhibitory and activating KIRs, their respective ligands [180] are listed. The ligands are grouped according to sequence similarites. The B alleles are those that contain the Bw4 serological motif (HLA-B$^{\text{Bw4}}$) and the C alleles are grouped according to their amino acid residue at position 80: the group C01 . . . C08 contain asparagine at position 80 (HLA-C$^{\text{Asn80}}$) and the C02 . . . C06 group contain lysine (HLA-C$^{\text{Lys80}}$).

[1] The Kagoshima dataset contained no information for 2DS1.

We also tested whether the presence of known protective KIR-HLA associations from other pathogenic infections (see Section 7.4) were protective in HTLV-I infection.

7.3 Results

We found no significant relationship between the count of inhibitory HLA-KIR interactions and proviral load, for either HAM/TSP or AC groups (Figure 7.1; AC: $R^2 = 0.01$, $P = 0.154$. HAM/TSP: $R^2 < 0.001$, $P = 0.96$).

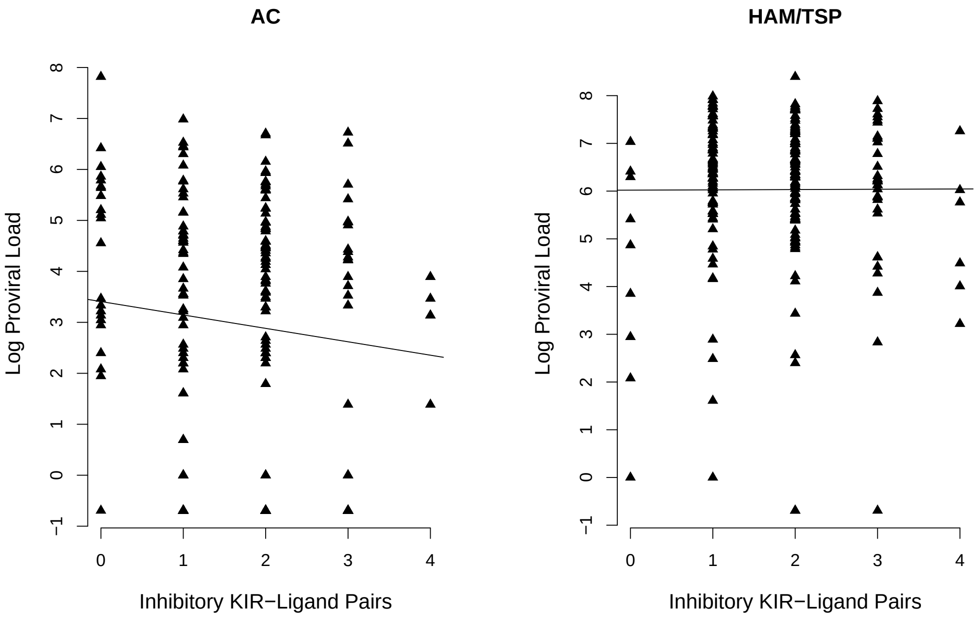


FIGURE 7.1: The count of inhibitory HLA-KIR interactions per individual plotted against their proviral load. No significant monotonic univariate relationship was found for AC or HAM/TSP groups (AC: $R^2 = 0.01$, $P = 0.154$. HAM/TSP: $R^2 < 0.001$, $P = 0.96$).

Figure 7.2 shows no relationship between the count of activating KIR-HLA interactions per individual and proviral load, again for both HAM/TSP and AC groups (AC: $R^2 < 0.001$, $P = 0.867$. HAM/TSP: $R^2 = 0.003$, $P = 0.397$). We also found no relationship between the difference in activating and inhibitory KIR-HLA interactions and proviral load for the 2 groups (AC: $R^2 = 0.012$, $P = 0.117$. HAM/TSP: $R^2 = 0.002$, $P = 0.498$).

Table 7.2 and Table 7.3 also demonstrate that KIR genes associated with disease outcome in other pathogens (see Section 7.4) show no protective effect in terms of HTLV-I disease status (2DL3: $\chi^2 = 1.243$, $P = 0.265$. 3DS1: $\chi^2 = 0.006$, $P = 0.938$). There was no difference in proviral load between individuals expressing 2DL3 and those that did not

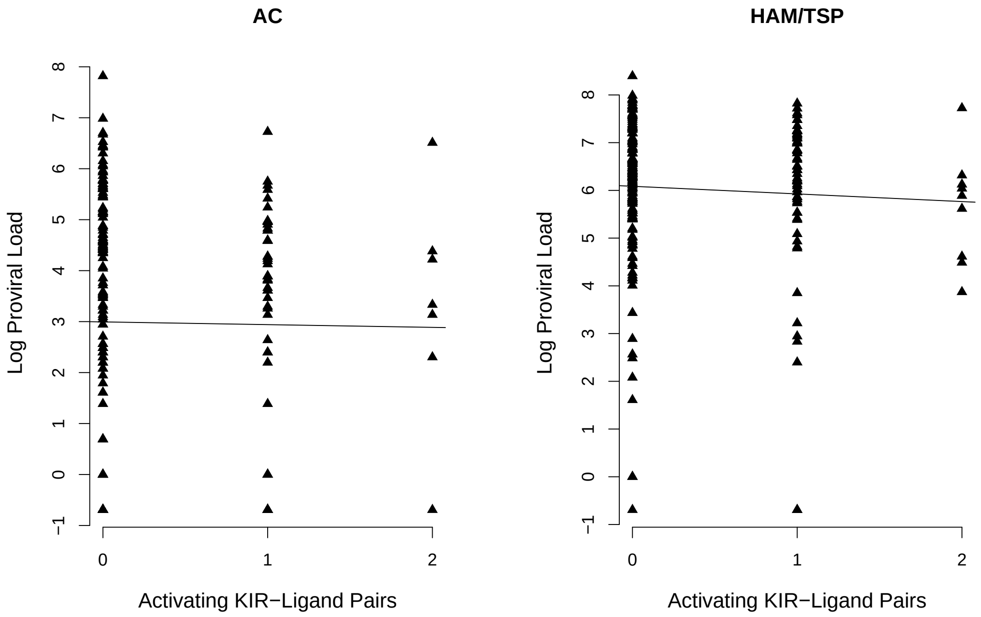


Figure 7.2: The count of activating HLA-KIR interactions per individual plotted against their proviral load. No significant monotonic univariate relationship was found for AC or HAM/TSP groups (AC: $R^2 < 0.001$, $P = 0.867$. HAM/TSP: $R^2 = 0.003$, $P = 0.397$).

(HAM/TSP: $P = 0.874$, AC: $P = 0.207$, Wilcoxon-Mann-Whitney). This was also the case for 3DS1 (HAM/TSP: $P = 0.393$, AC: $P = 0.289$, Wilcoxon-Mann-Whitney).

	HAM/TSP	AC
2DL3$^+$	$n = 198$	$n = 165$
2DL3$^-$	$n = 32$	$n = 37$

Table 7.2: The number of HAM/TSP and AC individuals that express KIR2DL3. There was no significant difference in the frequency of expression between the 2 groups ($\chi^2 = 1.243$, $P = 0.265$).

	HAM/TSP	AC
3DS1$^+$	$n = 38$	$n = 33$
3DS1$^-$	$n = 192$	$n = 169$

Table 7.3: The number of HAM/TSP and AC individuals that express KIR2DS1. There was no significant difference in the frequency of expression between the 2 groups ($\chi^2 = 0.006$, $P = 0.938$).

7.4 Discussion

Previous data regarding the effect of NK cells on HTLV-I infection has been sparse (Section 2.3.5) and, to my knowledge, this is the first time KIR-HLA associations have been examined in HTLV-I. Using the expression data of several KIR alleles and the presence of their associated MHC class I ligands in 230 HAM/TSP patients and 202 AC individuals, no significant associations were found with proviral load or disease status.

A number of assumptions have been made with this analysis regarding the interaction between KIRs and MHC class I. Combinations of KIR genes combine to generate haplotypes with widely differing balances between activating and inhibitory types. Summing across both types and analysing separately may be an over-simplification of how the KIR genes interact. This is based on previous KIR-HLA association studies [181], as well as evidence of a quantitative model of KIR protection against disease in HCV [182]. However, it should be noted that, although binding of KIR to MHC class I is determined by simple motifs (Table 7.1), this binding may not be strictly observed. The receptor-ligand interaction can be modulated by the peptide bound to HLA. For instance, KIR3DL2 binds HLA-A3 and -A11, but in binding studies using an HLA-A11 tetramer, this was only the case when specific viral peptides were refolded with the HLA molecule [181].

Bearing in mind these assumptions, the results of summing both the inhibitory and activating KIR-HLA interactions for each individual and comparing the count against proviral load yielded no significant relationships (Figure 7.1 and Figure 7.2).

The combination of KIR3DS1 and HLA-B alleles that contain the Bw4 serological motif (HLA-B^{Bw4}) has been found to be protective in both HIV [183, 184] and HCV [182] infection. In HCV, it was also found that there was an increased frequency of the inhibitory receptor KIR2DL3 in combination with HLA-C alleles with asparagine at position 80 (HLA-C^{Asn80}) [182]. This raises the hypothesis that certain KIR-HLA combinations confer a level of non-specific protection against multiple viral infections. However, we found no such association for either KIR-HLA interaction with disease status in HTLV-I infection (Table 7.2 and Table 7.3).

Chapter 8

Conclusion

8.1 Introduction

The immune response to HTLV-I infection has remained an intriguing and difficult area of research for the 30 years since its discovery and classification. Why approximately 95% of individuals who are infected with the virus remain asymptomatic and the remainder develop either debilitating inflammatory conditions (e.g. HAM/TSP) or an aggressive T-cell lymphoma (ATL) is still not fully understood. The immune response to HTLV-I infection includes cytotoxic T lymphocytes (CTLs), antibodies, T_{reg} cells and natural killer (NK) cells (Section 2.3). However among these adaptive immune responses, it is the CTL response that is the most important determinant of disease and proviral load.

The conflicting results of the role of CTLs (or CD8$^+$ T cells) in controlling HTLV-I infection has painted a confused picture of this arm of the immune response. On the one hand, the protective effect of a strong CD8$^+$ T cell response has been demonstrated by selection pressure on the dominant target antigen, Tax [80, 83], host genetics [1, 4, 57] and gene expression microarrays [100]. Contrary to these results, the effect of HTLV-I-specific CD8$^+$ T cell frequency has been ambiguous, with different groups reporting little effect of frequency on proviral load [9, 10] and some reporting a positive correlation with proviral load [86, 185].

These data has led to the hypothesis that it is the quality of the virus-specific CTLs, and not their frequency, that determines their effectiveness in controlling viral load and the risk of associated disease [186]. This quality has been defined by a mathematical model of CTL-mediated lysis [87] and has yielded the observation that approximately 35% of the between-individual variation in proviral load can be explained by variation in their CD8$^+$ T cell lytic efficiency [87, 89].

So what is the functional basis of this variation in CD8$^+$ T cell lysis efficiency? The are a number of possibilites that have been investigated: variation in the sensitivity of CD8$^+$ T cells to antigen concentration ("functional avidity" [89]), the ability of HTLV-I-specific CD8$^+$ T cells to respond in multiple ways to antigen ("polyfunctionality" [149]) and the role of regulatory T cells in the CTL response [187]. However, the focus of my PhD was the observation of the protective effect of the HLA class I alleles *A*02* and *Cw*08* and the detrimental effect of *B*54* [1], as well as the overall protective effect of HLA class I heterozygosity [4].

This variation in HLA class I effect across different alleles in terms of proviral load and the risk of associated disease suggested two observations :

(i) The association between HLA class I and proviral load and disease status in HTLV-I infection was - given the unequivocal function of HLA class I to display viral antigen to CTLs - the strongest evidence that CTL efficiency controls HTLV-I infection.

(ii) The functional difference between protective and detrimental HLA class I alleles could be understood by studying the viral epitopes that bind to these alleles.

From this information, we formulated the following structure to try and understand the underlying mechanism of HLA class I protection:

(a) Refine and, if possible, improve upon existing knowledge of the associations between HLA class I alleles and markers of HTLV-I infection.

(b) Rigorously test existing epitope prediction software to predict the HTLV-I epitopes that bind to the HLA class I alleles of interest.

(c) Use this information to test hypotheses about the epitope properties of protective and detrimental alleles.

(d) Model the CD8$^+$ T cell response in terms of its rate of lysis of infected CD4$^+$ T cells.

(e) Further understand the role of CTLs and NK cells in HTLV-I infection.

8.2 A Summary of Results by Chapter

8.2.1 Refining HLA class I Allele Associations

Chapter 3 built upon the conclusions of the series of papers by Jefferys *et al.* [1, 4] detailing the HLA class I associations with proviral load and the risk of HAM/TSP in HTLV-I infection. The conclusions of these papers can be summarised as follows:

(i) The presence of *HLA-A*02* is associated with a lower risk of HAM/TSP and a reduced proviral load in aymptomatic carriers of HTLV-I (but not in HAM/TSP individuals).

(ii) Independent of the *HLA-A*02* effect, the presence of *HLA-Cw*08* is associated with a lower risk of HAM/TSP and a reduced proviral load in aymptomatic carriers of HTLV-I (but not in HAM/TSP individuals).

(iii) *HLA-B*054* is associated with an increased risk of HAM/TSP and a higher proviral load in HAM/TSP patients (but not in asymptomatic carriers).

(iv) Individals who are heterozygous at all three HLA class I loci have a lower proviral load than those who are homozygous at one or more loci.

Our goal was to reanalyse this data and test for any further associations. We wanted to increase the pool of statistically significant protective and detrimental alleles, which would provide us with greater power in the analysis of their epitopes. We used a combination of Mann-Whitney U tests, χ^2 analysis and a novel ranking test (Section 3.2). This analysis suggested other HLA class I alleles could be associated with HTLV-I infection outcome (e.g. Figure 3.3). However, these associations did not prove significant using the the data of the Kagoshima Cohort (Section 3.2.1).

8.2.2 Rescaling in Epitope Prediction

The initial goal of Chapter 4 was to verify epitope prediction software for our use in predicting HTLV-I epitopes. Based on our preliminary analysis, we decided to focus on the subject of rescaling predicted binding affinities - a normalization procedure built on the assumption that different HLA class I molecules will bind to the same number of viral peptides. We used a combination of ROC curve and ranking analysis (Section 4.3) to show that, when comparing predicted binding affinitities between alleles, non-rescaled affinities are more accurate for epitope discovery. We incorporated these results into a web-based epitope prediction server (Metaserver, Section 4.4) and our analysis of rescaling has been verified and acknowledged by other members of the epitope prediction community [188].

8.2.3 Using Predicted Binding Affinities

In Chapter 5, we used epitope prediction software to define the predicted epitopes of HTLV-I. We defined our approach to understanding the basis of HLA class I protection in terms of specificity i.e. is it advantageous (in terms of proviral load and risk of HAM/TSP) to possess alleles that bind strongly to specific HTLV-I proteins? In order to test this, we built upon a novel ranking method by Borghans *et al.* [117] for the study of HIV epitopes. Our method defined the strength of binding of 9-mer peptides from a specific HTLV-I protein in terms of their rank binding strength compared to every other 9-mer peptide in the HTLV-I proteome. Using this method, along with raw predicted binding affinities and the SIR metric (Section B.1), and also 2 independent methods of epitope prediction (Metaserver and Epipred, Section 5.2.1), we reached the following conclusions:

(1) HLA Class I alleles previously associated with reduced proviral load and HAM/TSP prevalence (HLA-A*02 and -Cw*08) were predicted to bind epitopes from the viral protein HBZ significantly more strongly than an allele associated with increased proviral load and HAM/TSP prevalence (HLA-B*54).

(2) Asymptomatic carriers had a HLA class I genotype that predisposed them to bind epitopes from HBZ significantly more strongly than HAM/TSP patients. This result remained significant even when all individuals who possessed *A*02*, *B*54* and/or *Cw*08* were excluded from the cohort.

(3) Individuals whose HLA class I genotype predisposed them to strongly bind HBZ epitopes had a significantly reduced proviral load. This result was independent of the disease association reported above.

(4) Across all HTLV-I proteins, those proteins that were preferentially targeted by asymptomatic carriers were those associated with a greater reduction in proviral load when bound.

(5) HBZ-specific CD8$^+$ cells were detectable by IFNγ ELISpot in fresh PBMC from HTLV-I infected individuals.

The importance of HBZ in inhibiting expression of other HTLV-I genes [159, 162] and in promoting the proliferation of infected T-lymphocytes [155] certainly complemented our findings that the protein is an important target for the host immune system. The recent observation, however, that HBZ-specific CTLs could not lyse HTLV-I infected cells in ATL patients had cast some doubt on this theory [189]. In response to this data, our laboratory (Aileen Rowan) has shown that HBZ-specific CD8$^+$ cells are detectable by a CD107a degranulation assay (as an alternative to IFNγ ELISpot, mentioned above). We also showed that naturally infected cells from asymptomatic carriers and HAM/TSP patients are susceptible to lysis by HBZ-specific CTLs. The difference in susceptibility between non-leukaemic and ATL patients to HBZ-specific CD8$^+$ cells may be because ATL cells are inherently harder to kill and/or express lower levels of HBZ (B. Asquith, pers. comm.).

It was also interesting to compare the ranking of proteins in terms of proviral load and disease risk (Section 5.3.6) and the immunodominance hierarchy established by Goon *et al.* [11]. The immunodominance hierarchy ranked the HTLV-I proteins in terms of the frequency of protein-specific CD8$^+$ T cell responses. This hierarchy of proteins (highest to lowest frequency: Tax, Pol, Env, Gag, Rof, Tof, Pro, Rex) beared no relationship to our rank order of protective immune responses Table 5.11. This suggested again that immunodominance, as we found with Tax, was not necessarily related to protection in terms of targeting specific proteins in HTLV-I.

The relationship between Tax, as the immunodominant target, and HBZ is becoming more apparent through results such as those above. The data suggests that HBZ is important during the chronic stage of infection, where Tax expression is suppressed and is thus not presented to the immune system. Tax expression then increases to drive short phases of rapid expansion of infected T lymphocytes (C. Bangham, pers. comm.). Work is ongoing in our laboratory to test hypotheses generated from this relationship.

8.2.4 Understanding CTL Lysis and Antigen Expression

Chapter 6 built upon the model describing the CD8$^+$ T cell-mediated lysis of HTLV-I-infected cells [87]. This was modified by fitting a new model to time course expression data of the HTLV-I protein Tax, which better described the non-linear rate of increase of Tax expression. Using this model yielded the following results:

(1) A new model of CD8$^+$ T cell-mediated lysis of Tax$^{\text{low}}$ and Tax$^{\text{high}}$ expressing CD4$^+$ T cells.

(2) Target CD4$^+$ T cells expressing high levels of antigen (Tax) were killed significantly quicker than CD4$^+$ T cells expressing low levels of antigen.

(3) The ratio of killing rates of high and low Tax expressing CD4$^+$ T cells was maintained across patients.

Why the ratio of the high and low rates of lysis (ϵ) should remain constant across individuals is unknown. A possible explanation is that the probability of CTL-mediated lysis increases progressively with increasing density of MHC class I - peptide complexes and that the change in probability remains constant across individuals. Unfortunately, we did not have time to explore this question further.

8.2.5 The KIR:HLA Relationship in HTLV-I Infection

Chapter 7 examined the effect of KIR-HLA genotype on proviral load and HAM/TSP prevalence in HTLV-I infection:

(1) We found no significant relationship between the count of inhibitory or activiating HLA-KIR interactions and proviral load, for both HAM/TSP and AC groups.

(2) KIR genes associated with disease outcome in HIV and HCV showed no protective effect in terms of HTLV-I disease status or proviral load.

8.3 Final Remarks

The question of how HTLV-I persists as a chronic infection despite the activation of a strong T lymphocyte and antibody response remains unanswered. This is perhaps not surprising considering the multiple factors that have been shown to influence HTLV-I expression, proviral load and the risk of developing associated inflammatory diseases and ATL (Table 8.1). However, despite these multiple factors, there is strong evidence that it is the CTL response to HTLV-I infection that is the most important determinant of disease progression [1, 4, 81, 83, 87, 100]. One of these strands of evidence was the association between specific HLA class I alleles and protection from disease in HTLV-I infection. The goal of this research was to understand the underlying mechanism of this protection.

Our use of epitope prediction software resulted in a method of theoretically defining the presentation of viral peptides by HLA class I to cytotoxic T lymphocytes. The resulting identification of the HTLV-I protein HBZ as an important target of the immune system has aided the understanding of how HTLV-I persists *in vivo* and has focused experimental work on this protein. The rank method we developed can also be applied to other disease causing pathogens, especially where HLA class I associations with protection have been found (e.g. HCV).

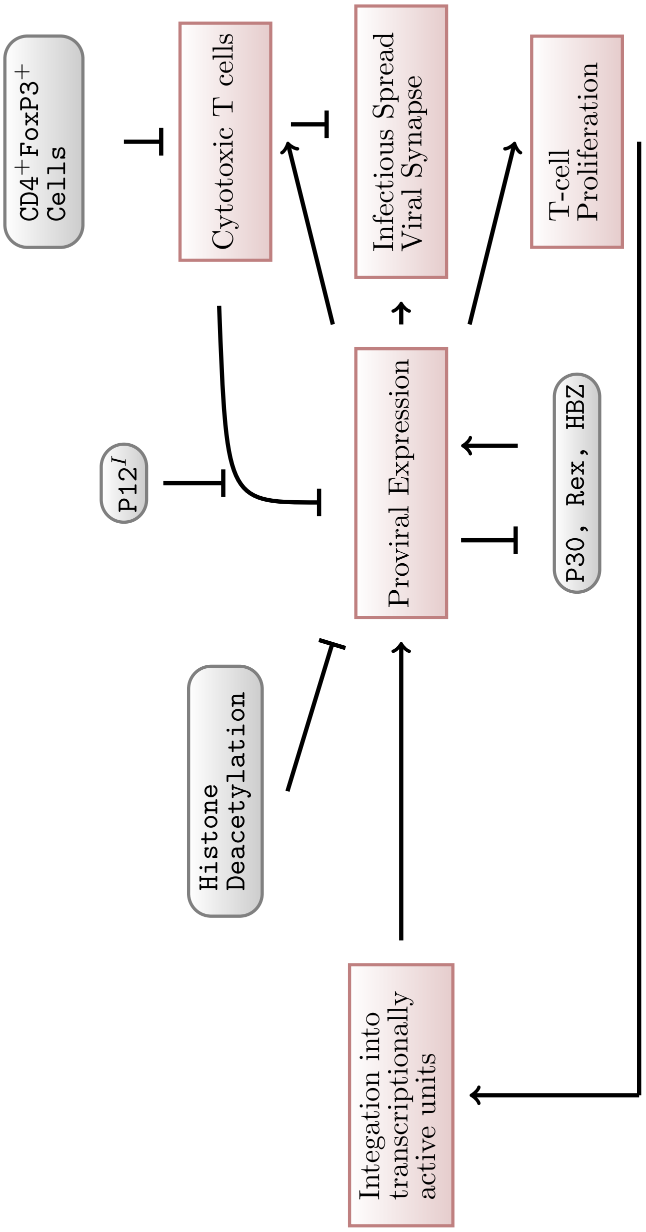


Table 8.1: The persistence of HTLV-I *in vivo*. The stimulatory and inhibitory factors controlling HTLV-I proliferation (C. Bangham, pers. comm.).

Appendix A

Supplementary Data for Chapter 4

A.1 Datasets

The epitope datasets used in Chapter 4 are detailed below. Each table gives the epitope, MHC class I allele and source protein of the epitope.

Epitope	Allele	Protein	Start	Epitope	Allele	Protein	Start	Epitope	Allele	Protein	Start
AYSSWMYSY	A1	EBN3_EBV	44	SVASTITGV	A2	ADFP_HUMAN	129	NAYEYFTKI	B7	BAK2_HUMAN	106
AIVDKVPSV	A2	COPG_HUMAN	147	SVFAGVVGV	A2	CYG3_HUMAN	581	NAYVNINRI	B7	HAPP_HUMAN	363
ALADGVQKV	A2	APL1_HUMAN	160	TIHDIILEC	A2	VE6_HPV16	29	NPVPVGNIY	B7	GAG_HV2G1	257
ALANGIEEV	A2	APL3_HUMAN	172	TILLGIFFL	A2	MSHR_HUMAN	244	NSSKVSQNY	B7	GAG_HV1J3	123
ALASHLIEA	A2	EHD2_HUMAN	507	VLEETSVML	A2	VIE1_HCMVA	316	PPIPVGDIY	B7	GAG_HV1MA	259
ALFGALFLA	A2	PLTP_HUMAN	2	VMAPRTLVL	A2	1A23_HUMAN	3	PPSGKGGNY	B7	GAG_HV2CA	127
ALLNIKVKL	A2	K1CR_HUMAN	364	WLNEVEFKL	A2	DMD_HUMAN	1281	SPKLPVSSL	B7	DRI1_HUMAN	372
ALLVLYSFA	A2	LMP1_EBV	51	YLDNGVVFV	A2	DDB1_HUMAN	316	SPYQNIKIL	B7	SPSY_HUMAN	145
ALPHAILRL	A2	ACTB_HUMAN	170	YLVTRHADV	A2	POLG_HCVH	1131	TSEHSHFSL	B7	DCE2_HUMAN	277
FLALIICNA	A2	MSHR_HUMAN	283	YVDPVITSI	A2	MET_HUMAN	654	TVLDVGDAY	B7	POL_HV1BR	274
FLDGNELTL	A2	CLI1_HUMAN	167	DADKYAVTV	B7	OM1E_CHLTR	367	VFPTKDVAL	B7	PP65_HCMVA	187
FLDGNEMTL	A2	CLI4_HUMAN	178	DAEMTTRMV	B7	PSBA_HUMAN	90	VPSEPGGVL	B7	PTN6_HUMAN	422
FLDPRPLTV	A2	CP1B_HUMAN	190	DAENAMRYI	B7	CB20_HUMAN	93	WASRELERF	B7	GAG_HV1BR	35
FLLDKKIGV	A2	TCPB_HUMAN	218	DALLIIPKV	B7	TCPW_HUMAN	441	YAFNMKATV	B7	HS7C_HUMAN	545
GLIEKNIEL	A2	DNM1_HUMAN	425	DALLKFSHI	B7	BI1_HUMAN	11	ELRRKMMYM	B8	VIE1_HCMVA	199
GLLGTLVQL	A2	CTNB_HUMAN	400	DALLQMITI	B7	EF2_HUMAN	347	ELRSRYWAI	B8	VNUC_IAPUE	380
GLYPGLIWL	A2	IRF6_HUMAN	21	DALRSILTI	B7	SYM_HUMAN	703	GEIYKRWII	B8	GAG_HV1BR	258
KASEKIFYV	A2	SSX2_HUMAN	41	DAYVLPKLY	B7	RS26_HUMAN	60	VMLRWGVLA	B8	NCAP_HRSVA	256
LLFDRPMHV	A2	ROM_HUMAN	268	DGYEQAARV	B7	TCPE_HUMAN	135	WVKEKVVAL	B8	NK4_HUMAN	175
LLMGTLGIV	A2	VE7_HPV16	82	DPYEVSYRI	B7	BTG1_HUMAN	107	ARFGLIQSM	B27	Y174_HUMAN	81
LTAGFLIFL	A2	LMP2_EBV	453	DPYKVYRIV	B7	IRF4_HUMAN	120	GRFSGLLGR	B27	IL16_HUMAN	180
NLLPKLHIV	A2	CLI1_HUMAN	179	FAYVQIKTI	B7	CP51_HUMAN	454	GRNVVLDKS	B27	CH60_YEREN	35
NLTISDVSV	A2	MUC1_HUMAN	1133	HPDIVIYQY	B7	POL_HV1U4	329	IRAAPPPLF	B27	PRTP_HUMAN	2
QLIDKVWQL	A2	SC14_HUMAN	593	IPQQHTQVL	B7	CEA5_HUMAN	632	IRAASAITA	B27	CH60_YEREN	420
RLVDDFLLV	A2	TERT_HUMAN	865	KPAFFAEKL	B7	ANX1_HUMAN	273	IRGAIILAK	B27	RS2_HUMAN	151
SLFPGKLEV	A2	FLIH_HUMAN	1010	KPSLPFTSL	B7	CYRG_HUMAN	3	IRLRPGGKK	B27	GAG_HV1BR	18
SLIGHLQTL	A2	DUS5_HUMAN	337	LPRSTVINI	B7	IFM1_HUMAN	19	IRNDEELNK	B27	H2AC_HUMAN	87
SLSEKTVLL	A2	CD59_HUMAN	106	MPMNVADLI	B7	IF42_HUMAN	399	IRRGVMLAV	B27	CH60_HUMAN	140
SLWGQPAEA	A2	CA54_HUMAN	18	MPWFKGWKV	B7	EF11_HUMAN	208	KRGIDKAVI	B27	CH60_YEREN	117
STAPPVHNV	A2	MUC1_HUMAN	950	NAACMALNI	B7	OM1E_CHLTR	121	KRIQEIIEQ	B27	CH60_HUMAN	369

Table A.1: The SYF[1] dataset.

Epitope	Allele	Protein	Start	Epitope	Allele	Protein	Start
KRTLKIPAM	B27	CH60_HUMAN	469	ALNFPGSQK	A3	PM17_HUMAN	87
MRMATPLLM	B27	HG2A_HUMAN	107	RLGVRATRK	A3	POLG_HCV1	43
NRIVYLYTK	B27	RL34_HUMAN	27	GPISGHVLK	A3	PP65_HCMVA	16
QRNLYIAGF	B27	CDM_HUMAN	100	YTPTISRER	A3	PSB2_HUMAN	147
QRNVNIFKF	B27	LDHA_HUMAN	110	QAIKGMHIR	A3	RL17_HUMAN	34
QRVNVQPEL	B27	PGTB_HUMAN	321	SLADIMAKR	A3	RL24_HUMAN	86
TRYQGVNLY	B27	PAB1_HUMAN	289	DTIEIITDR	A3	ROA2_HUMAN	139
AENLWVTVY	B44	ENV_HV1S3	30	ETIGEILKK	A3	ROK_HUMAN	95
AETPDIKLF	B44	RS5_HUMAN	12	FCVGFTKKR	A3	RS3A_HUMAN	137
NEGLGWAGW	B44	POLG_HCVJ6	88	EVVVSGKLR	A3	RS3_HUMAN	135
IALYLQQNW	B58	LMP1_EBV	156	ETFSGVYKK	A3	RS7_HUMAN	171
ITTKAISRW	B58	TCPG_HUMAN	159	STIEYVIQR	A3	S23B_HUMAN	115
KTKEVIQEW	B58	TALI_HUMAN	343	TPAGGGFPR	A3	TISB_HUMAN	43
VSFIEFVGW	B58	EBN4_EBV	279	AIYKQSQHM	A24	P53_HUMAN	161
AFHHVAREL	B62	NEF_HV1BR	190	AYSQQTRGL	A24	POLG_HCVBK	1031
GFYPGSIEV	B62	HB2I_HUMAN	150	EYLQLVFGI	A24	MAG2_HUMAN	156
IKADHVSTY	B62	HA2Q_HUMAN	32	EYLVSFGVW	A24	CORA_HPBVJ	117
WQYFFPVIF	B62	MAG3_HUMAN	143	HYTNASDGL	A24	LCK_HUMAN	207
DTAAQITQR	A3	1B35_HUMAN	161	KYTSFPWLL	A24	DPOL_HPBVJ	756
TIIDILTKR	A3	ANX1_HUMAN	63	QFQSIYAKF	A24	RECO_HUMAN	47
TIVNILTNR	A3	ANX2_HUMAN	54	QYDPVAALF	A24	PP65_HCMVA	341
TIIDIITHR	A3	ANX6_HUMAN	384	RWPSCQKKF	A24	WT1_HUMAN	417
ATIGTAMYK	A3	BRL1_EBV	134	TFDYLRSVL	A24	LCK_HUMAN	485
GSPATWTTR	A3	CA34_HUMAN	1436	TYGEIFEKF	A24	N4BM_HUMAN	107
YVNVNMGLK	A3	CORA_HPBV4	88	VYAETKHFL	A24	TERT_HUMAN	324
RFKMFPEVK	A3	DCE2_HUMAN	255	VYALPLKML	A24	PP65_HCMVA	113
TLYCVHQRI	A3	GAG_HV1BR	83	VYGFVRACL	A24	TERT_HUMAN	461
IVGLNKIVR	A3	GAG_HV1EL	266	YYMIGEQKF	A24	NNMT_HUMAN	203
DVFVVGTER	A3	GTFI_HUMAN	53				
SIMKWNRER	A3	NB6M_HUMAN	48				

Table A.1: Continued

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
RRGWEALKY	A1	gp160	787	TLNAWVKVV	A2	p24-p2p7p1	19	RLRPGGKKK	A3	p17	20
WIYHTQGYF	A1	Nef	113	AMQMLKETI	A2	p24-p2p7p1	65	TVRLIKLLY	A3	Rev	15
YFPDWQNYT	A1	Nef	120	AEWDRVHPV	A2	p24-p2p7p1	78	KLLYQSNPP	A3	Rev	20
GSEELRSLY	A1	p17	71	TLQEQIGWM	A2	p24-p2p7p1	110	RILGTYLGR	A3	Rev	58
QRPLVTIKI	A1	Protease	7	MTNNPPIPV	A2	p24-p2p7p1	118	ALVEICTEM	A3	RT	33
ISERILGTY	A1	Rev	55	RMYSPTSIL	A2	p24-p2p7p1	143	NTPVFAIKK	A3	RT	57
LWVTVYYGV	A2	gp160	34	YVDRFYKTL	A2	p24-p2p7p1	164	GIPHPAGLK	A3	RT	93
VTVYYGVPV	A2	gp160	36	VLAEAMSQV	A2	p24-p2p7p1	230	AIFQSSMTK	A3	RT	158
NVWATHACV	A2	gp160	67	LVGPTPVNI	A2	Protease	76	QIYPGIKVR	A3	RT	269
QMHEDIISL	A2	gp160	103	ALVEICTEM	A2	RT	33	QIIEQLIKK	A3	RT	520
KLTPLCVSL	A2	gp160	121	YTAFTIPSI	A2	RT	127	KVYLAWVPA	A3	RT	530
KLTSCNTSV	A2	gp160	192	VIYQYMDDL	A2	RT	179	TACTNCYCK	A3	Tat	20
QRGPGRAFV	A2	gp160	310	YQYMDDLYV	A2	RT	181	HMYVSGKAR	A3	Vif	28
TLKQIASKL	A2	gp160	341	KIEELRQHL	A2	RT	201	KLTEDRWNK	A3	Vif	168
TMGAASMTL	A2	gp160	529	ILKEPVHGV	A2	RT	309	IQRGPGRAF	A24	gp160	309
AVLSIVNRV	A2	gp160	700	PLVKLWYQL	A2	RT	421	FYCNSTQLF	A24	gp160	383
RLVNGSLAL	A2	gp160	747	KLGKAGYVT	A2	RT	451	RYLKDQQLL	A24	gp160	585
RLRDLLLIV	A2	gp160	770	ALQDSGLEV	A2	RT	485	WYIKLFIMI	A24	gp160	680
LLNATAIAV	A2	gp160	814	AIIRILQQL	A2	Vpr	59	SYHRLRDLL	A24	gp160	767
RVIEVVQGA	A2	gp160	828	RILQQLLFI	A2	Vpr	62	HSQRRQDIL	A24	Nef	102
RIRQGLERI	A2	gp160	846	VVAIIIAIV	A2	Vpu	13	RQDILDLWI	A24	Nef	106
QVRDQAEHL	A2	Integrase	164	SLWDQSLKP	A3	gp160	110	GYFPDWQNY	A24	Nef	119
LLWKGEGAV	A2	Integrase	241	VSFEPIPIH	A3	gp160	208	DSRLAFHHV	A24	Nef	186
ATNAACAWL	A2	Nef	50	HSFNCGGEF	A3	gp160	374	AFHHVAREL	A24	Nef	190
AAVDLSHFL	A2	Nef	83	AVDLSHFLK	A3	Nef	84	KYKLKHIVW	A24	p17	28
YPLTFGWCY	A2	Nef	135	DLSHFLKEK	A3	Nef	86	EIYKRWIIL	A24	p24-p2p7p1	128
LTFGWCYKL	A2	Nef	137	ILDLWIYHT	A3	Nef	109	DYVDRFYKT	A24	p24-p2p7p1	163
LEWRFDSRL	A2	Nef	181	PLTFGWCYK	A3	Nef	136	VYYDPSKDL	A24	RT	317
AFHHVAREL	A2	Nef	190	AFHHVAREL	A3	Nef	190	IYQEPFKNL	A24	RT	341
SLYNTVATL	A2	p17	77	KIRLRPGGK	A3	p17	18	EVIPMFSAL	A26	p24-p2p7p1	35

TABLE A.2: The Lanl[179] dataset.

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
YVDRFYKTL	A26	p24-p2p7p1	164	GRAFVTIGK	B27	gp160	314	IVLPEKDSW	B58	RT	244
ETKLGKAGY	A26	RT	449	GRRGWEALK	B27	gp160	786	ITTESIVIW	B58	RT	375
IPRRIRQGL	B7	gp160	843	KIRLRPGGK	B27	p17	18	VSGKARGWF	B58	Vif	31
LPPVVAKEI	B7	Integrase	28	IRLRPGGKK	B27	p17	19	AVRHFPRIW	B58	Vpr	30
FPVTPQVPL	B7	Nef	68	KRWIILGLN	B27	p24-p2p7p1	131	SFNCGGEFF	B62	gp160	375
TPQVPLRPM	B7	Nef	71	RWIILGLNK	B27	p24-p2p7p1	132	RAIEAQQHL	B62	gp160	557
RPMTYKAAV	B7	Nef	77	VRHFPRIWL	B27	Vpr	31	THLEGKVIL	B62	Integrase	66
TPGPGVRYP	B7	Nef	128	TPQDLNTML	B39	p24-p2p7p1	48	IKQEFGIPY	B62	Integrase	135
YPLTFGWCY	B7	Nef	135	HPVHAGPIA	B39	p24-p2p7p1	84	RKAKIIRDY	B62	Integrase	263
KIRLRPGGK	B7	p17	18	LEKHGAITS	B44	Nef	37	RMRRAEPAA	B62	Nef	19
SPRTLNAWV	B7	p24-p2p7p1	16	AAVDLSHFL	B44	Nef	83	MTYKAAVDL	B62	Nef	79
ATPQDLNTM	B7	p24-p2p7p1	47	KEKGGLEGL	B44	Nef	92	AAVDLSHFL	B62	Nef	83
TPQDLNTML	B7	p24-p2p7p1	48	GELDRWEKI	B44	p17	11	YFPDWQNYT	B62	Nef	120
HPVHAGPIA	B7	p24-p2p7p1	84	SEGATPQDL	B44	p24-p2p7p1	44	LTFGWCYKL	B62	Nef	137
ANPDCKTIL	B7	p24-p2p7p1	194	KETINEEAA	B44	p24-p2p7p1	70	WRFDSRLAF	B62	Nef	183
GPGHKARVL	B7	p24-p2p7p1	223	EEAAEWDRV	B44	p24-p2p7p1	75	RLRPGGKKK	B62	p17	20
YPLTSLRSL	B7	p24-p2p7p1	352	AEWDRVHPV	B44	p24-p2p7p1	78	RFAVNPGLL	B62	p17	43
SPAIFQSSM	B7	RT	156	CTERQANFL	B44	p24-p2p7p1	294	VKVVEEKAF	B62	p24-p2p7p1	24
IPLTEEAEL	B7	RT	293	KELYPLTSL	B44	p24-p2p7p1	349	FSPEVIPMF	B62	p24-p2p7p1	32
YLAWVPAHK	B7	RT	532	IEELRQHLL	B44	RT	202	GHQAAMQML	B62	p24-p2p7p1	61
FPRIWLHGL	B7	Vpr	34	REPHNEWTL	B44	Vpr	12	GLNKIVRMY	B62	p24-p2p7p1	137
RVKEKYQHL	B8	gp160	2	RAIEAQQHL	B58	gp160	557	YVDRFYKTL	B62	p24-p2p7p1	164
FNCGGEFFY	B8	gp160	376	KTAVQMAVF	B58	Integrase	173	GHKAIGTVL	B62	Protease	68
GGKKKYKLK	B8	p17	24	KAAVDLSHF	B58	Nef	82	IHSISERIL	B62	Rev	52
ELRSLYNTV	B8	p17	74	HTQGYFPDW	B58	Nef	116	IPLTEEAEL	B62	RT	293
EIKDTKEAL	B8	p17	93	YTPGPGVRY	B58	Nef	127	DVKQLTEAV	B62	RT	364
GEIYKRWII	B8	p24-p2p7p1	127	ISPRTLNAW	B58	p24-p2p7p1	15	ITKALGISY	B62	Tat	39
NANPDCKTI	B8	p24-p2p7p1	193	FSPEVIPMF	B58	p24-p2p7p1	32	WHLGQGVSI	B62	Vif	79
DCKTILKAL	B8	p24-p2p7p1	197	STLQEQIGW	B58	p24-p2p7p1	109	AVRHFPRIW	B62	Vpr	30
GPKVKQWPL	B8	RT	18	QASQEVKNW	B58	p24-p2p7p1	176				

Table A.2: Continued

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
RRGWEALKY	A0101	gp160	787	KLTSCNTSV	A0202	gp160	192	LLWKGEGAV	A0203	Integrase	241
WIYHTQGYF	A0101	Nef	113	QRGPGRAFV	A0202	gp160	310	ATNAACAWL	A0203	Nef	50
YFPDWQNYT	A0101	Nef	120	TLKQIASKL	A0202	gp160	341	AAVDLSHFL	A0203	Nef	83
GSEELRSLY	A0101	p17	71	TMGAASMTL	A0202	gp160	529	YPLTFGWCY	A0203	Nef	135
QRPLVTIKI	A0101	Protease	7	AVLSIVNRV	A0202	gp160	700	LTFGWCYKL	A0203	Nef	137
ISERILGTY	A0101	Rev	55	RLVNGSLAL	A0202	gp160	747	LEWRFDSRL	A0203	Nef	181
LWVTVYYGV	A0201	gp160	34	RLRDLLLIV	A0202	gp160	770	AFHHVAREL	A0203	Nef	190
VTVYYGVPV	A0201	gp160	36	LLNATAIAV	A0202	gp160	814	SLYNTVATL	A0203	p17	77
NVWATHACV	A0201	gp160	67	RVIEVVQGA	A0202	gp160	828	TLNAWVKVV	A0203	p24-p2p7p1	19
QMHEDIISL	A0201	gp160	103	RIRQGLERI	A0202	gp160	846	AMQMLKETI	A0203	p24-p2p7p1	65
KLTPLCVSL	A0201	gp160	121	QVRDQAEHL	A0202	Integrase	164	AEWDRVHPV	A0203	p24-p2p7p1	78
KLTSCNTSV	A0201	gp160	192	LLWKGEGAV	A0202	Integrase	241	TLQEQIGWM	A0203	p24-p2p7p1	110
QRGPGRAFV	A0201	gp160	310	ATNAACAWL	A0202	Nef	50	MTNNPPIPV	A0203	p24-p2p7p1	118
TLKQIASKL	A0201	gp160	341	AAVDLSHFL	A0202	Nef	83	RMYSPTSIL	A0203	p24-p2p7p1	143
TMGAASMTL	A0201	gp160	529	YPLTFGWCY	A0202	Nef	135	YVDRFYKTL	A0203	p24-p2p7p1	164
AVLSIVNRV	A0201	gp160	700	LTFGWCYKL	A0202	Nef	137	VLAEAMSQV	A0203	p24-p2p7p1	230
RLVNGSLAL	A0201	gp160	747	LEWRFDSRL	A0202	Nef	181	LVGPTPVNI	A0203	Protease	76
RLRDLLLIV	A0201	gp160	770	AFHHVAREL	A0202	Nef	190	ALVEICTEM	A0203	RT	33
LLNATAIAV	A0201	gp160	814	SLYNTVATL	A0202	p17	77	YTAFTIPSI	A0203	RT	127
RVIEVVQGA	A0201	gp160	828	TLNAWVKVV	A0202	p24-p2p7p1	19	VIYQYMDDL	A0203	RT	179
RIRQGLERI	A0201	gp160	846	AMQMLKETI	A0202	p24-p2p7p1	65	YQYMDDLYV	A0203	RT	181
QVRDQAEHL	A0201	Integrase	164	AEWDRVHPV	A0202	p24-p2p7p1	78	KIEELRQHL	A0203	RT	201
LLWKGEGAV	A0201	Integrase	241	TLQEQIGWM	A0202	p24-p2p7p1	110	ILKEPVHGV	A0203	RT	309
ATNAACAWL	A0201	Nef	50	MTNNPPIPV	A0202	p24-p2p7p1	118	PLVKLWYQL	A0203	RT	421
AAVDLSHFL	A0201	Nef	83	RMYSPTSIL	A0202	p24-p2p7p1	143	KLGKAGYVT	A0203	RT	451
YPLTFGWCY	A0201	Nef	135	YVDRFYKTL	A0202	p24-p2p7p1	164	ALQDSGLEV	A0203	RT	485
LTFGWCYKL	A0201	Nef	137	VLAEAMSQV	A0202	p24-p2p7p1	230	AIIRILQQL	A0203	Vpr	59
LEWRFDSRL	A0201	Nef	181	LVGPTPVNI	A0202	Protease	76	RILQQLLFI	A0203	Vpr	62
AFHHVAREL	A0201	Nef	190	ALVEICTEM	A0202	RT	33	VVAIIIAIV	A0203	Vpu	13
SLYNTVATL	A0201	p17	77	YTAFTIPSI	A0202	RT	127	LWVTVYYGV	A0206	gp160	34
TLNAWVKVV	A0201	p24-p2p7p1	19	VIYQYMDDL	A0202	RT	179	VTVYYGVPV	A0206	gp160	36
AMQMLKETI	A0201	p24-p2p7p1	65	YQYMDDLYV	A0202	RT	181	NVWATHACV	A0206	gp160	67
AEWDRVHPV	A0201	p24-p2p7p1	78	KIEELRQHL	A0202	RT	201	QMHEDIISL	A0206	gp160	103
TLQEQIGWM	A0201	p24-p2p7p1	110	ILKEPVHGV	A0202	RT	309	KLTPLCVSL	A0206	gp160	121
MTNNPPIPV	A0201	p24-p2p7p1	118	PLVKLWYQL	A0202	RT	421	KLTSCNTSV	A0206	gp160	192
RMYSPTSIL	A0201	p24-p2p7p1	143	KLGKAGYVT	A0202	RT	451	QRGPGRAFV	A0206	gp160	310
YVDRFYKTL	A0201	p24-p2p7p1	164	ALQDSGLEV	A0202	RT	485	TLKQIASKL	A0206	gp160	341
VLAEAMSQV	A0201	p24-p2p7p1	230	AIIRILQQL	A0202	Vpr	59	TMGAASMTL	A0206	gp160	529
LVGPTPVNI	A0201	Protease	76	RILQQLLFI	A0202	Vpr	62	AVLSIVNRV	A0206	gp160	700
ALVEICTEM	A0201	RT	33	VVAIIIAIV	A0202	Vpu	13	RLVNGSLAL	A0206	gp160	747
YTAFTIPSI	A0201	RT	127	LWVTVYYGV	A0203	gp160	34	RLRDLLLIV	A0206	gp160	770
VIYQYMDDL	A0201	RT	179	VTVYYGVPV	A0203	gp160	36	LLNATAIAV	A0206	gp160	814
YQYMDDLYV	A0201	RT	181	NVWATHACV	A0203	gp160	67	RVIEVVQGA	A0206	gp160	828
KIEELRQHL	A0201	RT	201	QMHEDIISL	A0203	gp160	103	RIRQGLERI	A0206	gp160	846
ILKEPVHGV	A0201	RT	309	KLTPLCVSL	A0203	gp160	121	QVRDQAEHL	A0206	Integrase	164
PLVKLWYQL	A0201	RT	421	KLTSCNTSV	A0203	gp160	192	LLWKGEGAV	A0206	Integrase	241
KLGKAGYVT	A0201	RT	451	QRGPGRAFV	A0203	gp160	310	ATNAACAWL	A0206	Nef	50
ALQDSGLEV	A0201	RT	485	TLKQIASKL	A0203	gp160	341	AAVDLSHFL	A0206	Nef	83
AIIRILQQL	A0201	Vpr	59	TMGAASMTL	A0203	gp160	529	YPLTFGWCY	A0206	Nef	135
RILQQLLFI	A0201	Vpr	62	AVLSIVNRV	A0203	gp160	700	LTFGWCYKL	A0206	Nef	137
VVAIIIAIV	A0201	Vpu	13	RLVNGSLAL	A0203	gp160	747	LEWRFDSRL	A0206	Nef	181
LWVTVYYGV	A0202	gp160	34	RLRDLLLIV	A0203	gp160	770	AFHHVAREL	A0206	Nef	190
VTVYYGVPV	A0202	gp160	36	LLNATAIAV	A0203	gp160	814	SLYNTVATL	A0206	p17	77
NVWATHACV	A0202	gp160	67	RVIEVVQGA	A0203	gp160	828	TLNAWVKVV	A0206	p24-p2p7p1	19
QMHEDIISL	A0202	gp160	103	RIRQGLERI	A0203	gp160	846	AMQMLKETI	A0206	p24-p2p7p1	65
KLTPLCVSL	A0202	gp160	121	QVRDQAEHL	A0203	Integrase	164	AEWDRVHPV	A0206	p24-p2p7p1	78

Table A.3: The Lanl^{661} dataset.

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
TLQEQIGWM	A0206	p24-p2p7p1	110	ILKEPVHGV	A0211	RT	309	KLTPLCVSL	A0216	gp160	121
MTNNPPIPV	A0206	p24-p2p7p1	118	PLVKLWYQL	A0211	RT	421	KLTSCNTSV	A0216	gp160	192
RMYSPTSIL	A0206	p24-p2p7p1	143	KLGKAGYVT	A0211	RT	451	QRGPGRAFV	A0216	gp160	310
YVDRFYKTL	A0206	p24-p2p7p1	164	ALQDSGLEV	A0211	RT	485	TLKQIASKL	A0216	gp160	341
VLAEAMSQV	A0206	p24-p2p7p1	230	AIIRILQQL	A0211	Vpr	59	TMGAASMTL	A0216	gp160	529
LVGPTPVNI	A0206	Protease	76	RILQQLLFI	A0211	Vpr	62	AVLSIVNRV	A0216	gp160	700
ALVEICTEM	A0206	RT	33	VVAIIIAIV	A0211	Vpu	13	RLVNGSLAL	A0216	gp160	747
YTAFTIPSI	A0206	RT	127	LWVTVYYGV	A0212	gp160	34	RLRDLLLIV	A0216	gp160	770
VIYQYMDDL	A0206	RT	179	VTVYYGVPV	A0212	gp160	36	LLNATAIAV	A0216	gp160	814
YQYMDDLYV	A0206	RT	181	NVWATHACV	A0212	gp160	67	RVIEVVQGA	A0216	gp160	828
KIEELRQHL	A0206	RT	201	QMHEDIISL	A0212	gp160	103	RIRQGLERI	A0216	gp160	846
ILKEPVHGV	A0206	RT	309	KLTPLCVSL	A0212	gp160	121	QVRDQAEHL	A0216	Integrase	164
PLVKLWYQL	A0206	RT	421	KLTSCNTSV	A0212	gp160	192	LLWKGEGAV	A0216	Integrase	241
KLGKAGYVT	A0206	RT	451	QRGPGRAFV	A0212	gp160	310	ATNAACAWL	A0216	Nef	50
ALQDSGLEV	A0206	RT	485	TLKQIASKL	A0212	gp160	341	AAVDLSHFL	A0216	Nef	83
AIIRILQQL	A0206	Vpr	59	TMGAASMTL	A0212	gp160	529	YPLTFGWCY	A0216	Nef	135
RILQQLLFI	A0206	Vpr	62	AVLSIVNRV	A0212	gp160	700	LTFGWCYKL	A0216	Nef	137
VVAIIIAIV	A0206	Vpu	13	RLVNGSLAL	A0212	gp160	747	LEWRFDSRL	A0216	Nef	181
LWVTVYYGV	A0211	gp160	34	RLRDLLLIV	A0212	gp160	770	AFHHVAREL	A0216	Nef	190
VTVYYGVPV	A0211	gp160	36	LLNATAIAV	A0212	gp160	814	SLYNTVATL	A0216	p17	77
NVWATHACV	A0211	gp160	67	RVIEVVQGA	A0212	gp160	828	TLNAWVKVV	A0216	p24-p2p7p1	19
QMHEDIISL	A0211	gp160	103	RIRQGLERI	A0212	gp160	846	AMQMLKETI	A0216	p24-p2p7p1	65
KLTPLCVSL	A0211	gp160	121	QVRDQAEHL	A0212	Integrase	164	AEWDRVHPV	A0216	p24-p2p7p1	78
KLTSCNTSV	A0211	gp160	192	LLWKGEGAV	A0212	Integrase	241	TLQEQIGWM	A0216	p24-p2p7p1	110
QRGPGRAFV	A0211	gp160	310	ATNAACAWL	A0212	Nef	50	MTNNPPIPV	A0216	p24-p2p7p1	118
TLKQIASKL	A0211	gp160	341	AAVDLSHFL	A0212	Nef	83	RMYSPTSIL	A0216	p24-p2p7p1	143
TMGAASMTL	A0211	gp160	529	YPLTFGWCY	A0212	Nef	135	YVDRFYKTL	A0216	p24-p2p7p1	164
AVLSIVNRV	A0211	gp160	700	LTFGWCYKL	A0212	Nef	137	VLAEAMSQV	A0216	p24-p2p7p1	230
RLVNGSLAL	A0211	gp160	747	LEWRFDSRL	A0212	Nef	181	LVGPTPVNI	A0216	Protease	76
RLRDLLLIV	A0211	gp160	770	AFHHVAREL	A0212	Nef	190	ALVEICTEM	A0216	RT	33
LLNATAIAV	A0211	gp160	814	SLYNTVATL	A0212	p17	77	YTAFTIPSI	A0216	RT	127
RVIEVVQGA	A0211	gp160	828	TLNAWVKVV	A0212	p24-p2p7p1	19	VIYQYMDDL	A0216	RT	179
RIRQGLERI	A0211	gp160	846	AMQMLKETI	A0212	p24-p2p7p1	65	YQYMDDLYV	A0216	RT	181
QVRDQAEHL	A0211	Integrase	164	AEWDRVHPV	A0212	p24-p2p7p1	78	KIEELRQHL	A0216	RT	201
LLWKGEGAV	A0211	Integrase	241	TLQEQIGWM	A0212	p24-p2p7p1	110	ILKEPVHGV	A0216	RT	309
ATNAACAWL	A0211	Nef	50	MTNNPPIPV	A0212	p24-p2p7p1	118	PLVKLWYQL	A0216	RT	421
AAVDLSHFL	A0211	Nef	83	RMYSPTSIL	A0212	p24-p2p7p1	143	KLGKAGYVT	A0216	RT	451
YPLTFGWCY	A0211	Nef	135	YVDRFYKTL	A0212	p24-p2p7p1	164	ALQDSGLEV	A0216	RT	485
LTFGWCYKL	A0211	Nef	137	VLAEAMSQV	A0212	p24-p2p7p1	230	AIIRILQQL	A0216	Vpr	59
LEWRFDSRL	A0211	Nef	181	LVGPTPVNI	A0212	Protease	76	RILQQLLFI	A0216	Vpr	62
AFHHVAREL	A0211	Nef	190	ALVEICTEM	A0212	RT	33	VVAIIIAIV	A0216	Vpu	13
SLYNTVATL	A0211	p17	77	YTAFTIPSI	A0212	RT	127	LWVTVYYGV	A0219	gp160	34
TLNAWVKVV	A0211	p24-p2p7p1	19	VIYQYMDDL	A0212	RT	179	VTVYYGVPV	A0219	gp160	36
AMQMLKETI	A0211	p24-p2p7p1	65	YQYMDDLYV	A0212	RT	181	NVWATHACV	A0219	gp160	67
AEWDRVHPV	A0211	p24-p2p7p1	78	KIEELRQHL	A0212	RT	201	QMHEDIISL	A0219	gp160	103
TLQEQIGWM	A0211	p24-p2p7p1	110	ILKEPVHGV	A0212	RT	309	KLTPLCVSL	A0219	gp160	121
MTNNPPIPV	A0211	p24-p2p7p1	118	PLVKLWYQL	A0212	RT	421	KLTSCNTSV	A0219	gp160	192
RMYSPTSIL	A0211	p24-p2p7p1	143	KLGKAGYVT	A0212	RT	451	QRGPGRAFV	A0219	gp160	310
YVDRFYKTL	A0211	p24-p2p7p1	164	ALQDSGLEV	A0212	RT	485	TLKQIASKL	A0219	gp160	341
VLAEAMSQV	A0211	p24-p2p7p1	230	AIIRILQQL	A0212	Vpr	59	TMGAASMTL	A0219	gp160	529
LVGPTPVNI	A0211	Protease	76	RILQQLLFI	A0212	Vpr	62	AVLSIVNRV	A0219	gp160	700
ALVEICTEM	A0211	RT	33	VVAIIIAIV	A0212	Vpu	13	RLVNGSLAL	A0219	gp160	747
YTAFTIPSI	A0211	RT	127	LWVTVYYGV	A0216	gp160	34	RLRDLLLIV	A0219	gp160	770
VIYQYMDDL	A0211	RT	179	VTVYYGVPV	A0216	gp160	36	LLNATAIAV	A0219	gp160	814
YQYMDDLYV	A0211	RT	181	NVWATHACV	A0216	gp160	67	RVIEVVQGA	A0219	gp160	828
KIEELRQHL	A0211	RT	201	QMHEDIISL	A0216	gp160	103	RIRQGLERI	A0219	gp160	846

Table A.3: Continued

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
QVRDQAEHL	A0219	Integrase	164	DLSHFLKEK	A1101	Nef	86	RLRPGGKKK	A3001	p17	20
LLWKGEGAV	A0219	Integrase	241	TLYCVHQRI	A1101	p17	84	KQNPDIVIY	A3001	RT	173
ATNAACAWL	A0219	Nef	50	AIFQSSMTK	A1101	RT	158	KLNWASQIY	A3001	RT	263
AAVDLSHFL	A0219	Nef	83	QIYPGIKVR	A1101	RT	269	QRGPGRAFV	A3002	gp160	310
YPLTFGWCY	A0219	Nef	135	FVNTPPLVK	A1101	RT	416	IVNRVRQGY	A3002	gp160	704
LTFGWCYKL	A0219	Nef	137	QIIEQLIKK	A1101	RT	520	KYWWNLLQY	A3002	gp160	794
LEWRFDSRL	A0219	Nef	181	RYLKDQQLL	A2301	gp160	585	KIQNFRVYY	A3002	Integrase	219
AFHHVAREL	A0219	Nef	190	KYKLKHIVW	A2301	p17	28	RLRPGGKKK	A3002	p17	20
SLYNTVATL	A0219	p17	77	IYQEPFKNL	A2301	RT	341	KQNPDIVIY	A3002	RT	173
TLNAWVKVV	A0219	p24-p2p7p1	19	IQRGPGRAF	A2402	gp160	309	KLNWASQIY	A3002	RT	263
AMQMLKETI	A0219	p24-p2p7p1	65	FYCNSTQLF	A2402	gp160	383	VRYPLTFGW	A3301	Nef	133
AEWDRVHPV	A0219	p24-p2p7p1	78	RYLKDQQLL	A2402	gp160	585	RLAFHHVAR	A3301	Nef	188
TLQEQIGWM	A0219	p24-p2p7p1	110	WYIKLFIMI	A2402	gp160	680	MVHQAISPR	A3301	p24-p2p7p1	10
MTNNPPIPV	A0219	p24-p2p7p1	118	SYHRLRDLL	A2402	gp160	767	AIFQSSMTK	A3301	RT	158
RMYSPTSIL	A0219	p24-p2p7p1	143	HSQRRQDIL	A2402	Nef	102	FYVDGAANR	A3301	RT	440
YVDRFYKTL	A0219	p24-p2p7p1	164	RQDILDLWI	A2402	Nef	106	EYRKILRQR	A3301	Vpu	29
VLAEAMSQV	A0219	p24-p2p7p1	230	GYFPDWQNY	A2402	Nef	119	ETAYFLLKL	A6801	Integrase	96
LVGPTPVNI	A0219	Protease	76	DSRLAFHHV	A2402	Nef	186	VTLWQRPLV	A6801	Protease	3
ALVEICTEM	A0219	RT	33	AFHHVAREL	A2402	Nef	190	DTVLEEMSL	A6801	Protease	30
YTAFTIPSI	A0219	RT	127	KYKLKHIVW	A2402	p17	28	NTPVFAIKK	A6801	RT	57
VIYQYMDDL	A0219	RT	179	EIYKRWIIL	A2402	p24-p2p7p1	128	AIFQSSMTK	A6801	RT	158
YQYMDDLYV	A0219	RT	181	DYVDRFYKT	A2402	p24-p2p7p1	163	ETAYFLLKL	A6802	Integrase	96
KIEELRQHL	A0219	RT	201	VYYDPSKDL	A2402	RT	317	VTLWQRPLV	A6802	Protease	3
ILKEPVHGV	A0219	RT	309	IYQEPFKNL	A2402	RT	341	DTVLEEMSL	A6802	Protease	30
PLVKLWYQL	A0219	RT	421	IQRGPGRAF	A2403	gp160	309	NTPVFAIKK	A6802	RT	57
KLGKAGYVT	A0219	RT	451	FYCNSTQLF	A2403	gp160	383	AIFQSSMTK	A6802	RT	158
ALQDSGLEV	A0219	RT	485	RYLKDQQLL	A2403	gp160	585	RAEPAADRV	A6901	Nef	22
AIIRILQQL	A0219	Vpr	59	WYIKLFIMI	A2403	gp160	680	RVGAASRDL	A6901	Nef	29
RILQQLLFI	A0219	Vpr	62	SYHRLRDLL	A2403	gp160	767	IPRRIRQGL	B0702	gp160	843
VVAIIIAIV	A0219	Vpu	13	HSQRRQDIL	A2403	Nef	102	LPPVVAKEI	B0702	Integrase	28
SLWDQSLKP	A0301	gp160	110	RQDILDLWI	A2403	Nef	106	FPVTPQVPL	B0702	Nef	68
VSFEPIPIH	A0301	gp160	208	GYFPDWQNY	A2403	Nef	119	TPQVPLRPM	B0702	Nef	71
HSFNCGGEF	A0301	gp160	374	DSRLAFHHV	A2403	Nef	186	RPMTYKAAV	B0702	Nef	77
AVDLSHFLK	A0301	Nef	84	AFHHVAREL	A2403	Nef	190	TPGPGVRYP	B0702	Nef	128
DLSHFLKEK	A0301	Nef	86	KYKLKHIVW	A2403	p17	28	YPLTFGWCY	B0702	Nef	135
ILDLWIYHT	A0301	Nef	109	EIYKRWIIL	A2403	p24-p2p7p1	128	KIRLRPGGK	B0702	p17	18
PLTFGWCYK	A0301	Nef	136	DYVDRFYKT	A2403	p24-p2p7p1	163	SPRTLNAWV	B0702	p24-p2p7p1	16
AFHHVAREL	A0301	Nef	190	VYYDPSKDL	A2403	RT	317	ATPQDLNTM	B0702	p24-p2p7p1	47
KIRLRPGGK	A0301	p17	18	IYQEPFKNL	A2403	RT	341	TPQDLNTML	B0702	p24-p2p7p1	48
RLRPGGKKK	A0301	p17	20	EVIPMFSAL	A2601	p24-p2p7p1	35	HPVHAGPIA	B0702	p24-p2p7p1	84
TVRLIKLLY	A0301	Rev	15	YVDRFYKTL	A2601	p24-p2p7p1	164	ANPDCKTIL	B0702	p24-p2p7p1	194
KLLYQSNPP	A0301	Rev	20	ETKLGKAGY	A2601	RT	449	GPGHKARVL	B0702	p24-p2p7p1	223
RILGTYLGR	A0301	Rev	58	EVIPMFSAL	A2602	p24-p2p7p1	35	YPLTSLRSL	B0702	p24-p2p7p1	352
ALVEICTEM	A0301	RT	33	YVDRFYKTL	A2602	p24-p2p7p1	164	SPAIFQSSM	B0702	RT	156
NTPVFAIKK	A0301	RT	57	ETKLGKAGY	A2602	RT	449	IPLTEEAEL	B0702	RT	293
GIPHPAGLK	A0301	RT	93	SFEPIPIHY	A2902	gp160	209	YLAWVPAHK	B0702	RT	532
AIFQSSMTK	A0301	RT	158	FNCGGEFFY	A2902	gp160	376	FPRIWLHGL	B0702	Vpr	34
QIYPGIKVR	A0301	RT	269	RIKQIINMW	A2902	gp160	419	RVKEKYQHL	B0801	gp160	2
QIIEQLIKK	A0301	RT	520	YFPDWQNYT	A2902	Nef	120	FNCGGEFFY	B0801	gp160	376
KVYLAWVPA	A0301	RT	530	LYNTVATLY	A2902	p17	78	GGKKKYKLK	B0801	p17	24
TACTNCYCK	A0301	Tat	20	NCYCKKCCF	A2902	Tat	24	ELRSLYNTV	B0801	p17	74
HMYVSGKAR	A0301	Vif	28	HIVSPRCEY	A2902	Vif	127	EIKDTKEAL	B0801	p17	93
KLTEDRWNK	A0301	Vif	168	QRGPGRAFV	A3001	gp160	310	GEIYKRWII	B0801	p24-p2p7p1	127
SVITQACPK	A1101	gp160	199	IVNRVRQGY	A3001	gp160	704	NANPDCKTI	B0801	p24-p2p7p1	193
NTLKQIASK	A1101	gp160	340	KYWWNLLQY	A3001	gp160	794	DCKTILKAL	B0801	p24-p2p7p1	197
AVDLSHFLK	A1101	Nef	84	KIQNFRVYY	A3001	Integrase	219	GPKVKQWPL	B0801	RT	18

Table A.3: Continued

Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2	Epitope	Allele	Protein	HXB2
RVKEKYQHL	B0802	gp160	2	TVLDVGDAY	B3501	RT	107	DSRLAFHHV	B5101	Nef	186
FNCGGEFFY	B0802	gp160	376	FSVPLDEDF	B3501	RT	116	AFHHVAREL	B5101	Nef	190
GGKKKYKLK	B0802	p17	24	SPAIFQSSM	B3501	RT	156	NANPDCKTI	B5101	p24-p2p7p1	193
ELRSLYNTV	B0802	p17	74	NPDIVIYQY	B3501	RT	175	EKEGKISKI	B5101	RT	42
EIKDTKEAL	B0802	p17	93	IPLTEEAEL	B3501	RT	293	QGWKGSPAI	B5101	RT	151
GEIYKRWII	B0802	p24-p2p7p1	127	EPIVGAETF	B3501	RT	432	IPLTEEAEL	B5101	RT	293
NANPDCKTI	B0802	p24-p2p7p1	193	DARLVITTY	B3501	Vif	61	EPIVGAETF	B5101	RT	432
DCKTILKAL	B0802	p24-p2p7p1	197	HTGERDWHL	B3501	Vif	73	EAVRHFPRI	B5101	Vpr	29
GPKVKQWPL	B0802	RT	18	TPQDLNTML	B3901	p24-p2p7p1	48	YPLTFGWCY	B5301	Nef	135
SFNCGGEFF	B1501	gp160	375	HPVHAGPIA	B3901	p24-p2p7p1	84	TPQDLNTML	B5301	p24-p2p7p1	48
RAIEAQQHL	B1501	gp160	557	LEKHGAITS	B4001	Nef	37	PPIPVGEIY	B5301	p24-p2p7p1	122
THLEGKVIL	B1501	Integrase	66	AAVDLSHFL	B4001	Nef	83	QASQEVKNW	B5301	p24-p2p7p1	176
IKQEFGIPY	B1501	Integrase	135	KEKGGLEGL	B4001	Nef	92	ASQEVKNWM	B5301	p24-p2p7p1	177
RKAKIIRDY	B1501	Integrase	263	GELDRWEKI	B4001	p17	11	TPPQKQEPI	B5301	p24-p2p7p1	339
RMRRAEPAA	B1501	Nef	19	SEGATPQDL	B4001	p24-p2p7p1	44	YPLTSLRSL	B5301	p24-p2p7p1	352
MTYKAAVDL	B1501	Nef	79	KETINEEAA	B4001	p24-p2p7p1	70	IPLTEEAEL	B5301	RT	293
AAVDLSHFL	B1501	Nef	83	EEAAEWDRV	B4001	p24-p2p7p1	75	RAIEAQQHL	B5701	gp160	557
YFPDWQNYT	B1501	Nef	120	AEWDRVHPV	B4001	p24-p2p7p1	78	KTAVQMAVF	B5701	Integrase	173
LTFGWCYKL	B1501	Nef	137	CTERQANFL	B4001	p24-p2p7p1	294	KAAVDLSHF	B5701	Nef	82
WRFDSRLAF	B1501	Nef	183	KELYPLTSL	B4001	p24-p2p7p1	349	HTQGYFPDW	B5701	Nef	116
RLRPGGKKK	B1501	p17	20	IEELRQHLL	B4001	RT	202	YFPDWQNYT	B5701	Nef	120
RFAVNPGLL	B1501	p17	43	REPHNEWTL	B4001	Vpr	12	YTPGPGVRY	B5701	Nef	127
VKVVEEKAF	B1501	p24-p2p7p1	24	LEKHGAITS	B4002	Nef	37	LTFGWCYKL	B5701	Nef	137
FSPEVIPMF	B1501	p24-p2p7p1	32	AAVDLSHFL	B4002	Nef	83	ISPRTLNAW	B5701	p24-p2p7p1	15
GHQAAMQML	B1501	p24-p2p7p1	61	KEKGGLEGL	B4002	Nef	92	FSPEVIPMF	B5701	p24-p2p7p1	32
GLNKIVRMY	B1501	p24-p2p7p1	137	GELDRWEKI	B4002	p17	11	STLQEQIGW	B5701	p24-p2p7p1	109
YVDRFYKTL	B1501	p24-p2p7p1	164	SEGATPQDL	B4002	p24-p2p7p1	44	QASQEVKNW	B5701	p24-p2p7p1	176
GHKAIGTVL	B1501	Protease	68	KETINEEAA	B4002	p24-p2p7p1	70	FSVPLDEDF	B5701	RT	116
IHSISERIL	B1501	Rev	52	EEAAEWDRV	B4002	p24-p2p7p1	75	IVLPEKDSW	B5701	RT	244
IPLTEEAEL	B1501	RT	293	AEWDRVHPV	B4002	p24-p2p7p1	78	ITTESIVIW	B5701	RT	375
DVKQLTEAV	B1501	RT	364	CTERQANFL	B4002	p24-p2p7p1	294	VSGKARGWF	B5701	Vif	31
ITKALGISY	B1501	Tat	39	KELYPLTSL	B4002	p24-p2p7p1	349	AVRHFPRIW	B5701	Vpr	30
WHLGQGVSI	B1501	Vif	79	IEELRQHLL	B4002	RT	202	RAIEAQQHL	B5801	gp160	557
AVRHFPRIW	B1501	Vpr	30	REPHNEWTL	B4002	Vpr	12	KTAVQMAVF	B5801	Integrase	173
TEKLWVTVY	B1801	gp160	31	TEKLWVTVY	B4402	gp160	31	KAAVDLSHF	B5801	Nef	82
YDTEVHNVW	B1801	gp160	61	LYNTVATLY	B4402	p17	78	HTQGYFPDW	B5801	Nef	116
QDILDLWIY	B1801	Nef	107	EEKAFSPEV	B4402	p24-p2p7p1	28	YTPGPGVRY	B5801	Nef	127
YPLTFGWCY	B1801	Nef	135	SEGATPQDL	B4402	p24-p2p7p1	44	ISPRTLNAW	B5801	p24-p2p7p1	15
NNETPGIRY	B1801	RT	136	QEPIDKELY	B4402	p24-p2p7p1	344	FSPEVIPMF	B5801	p24-p2p7p1	32
NPDIVIYQY	B1801	RT	175	EEMSLPGRW	B4402	Protease	34	STLQEQIGW	B5801	p24-p2p7p1	109
GRAFVTIGK	B2705	gp160	314	EELRQHLLR	B4402	RT	203	QASQEVKNW	B5801	p24-p2p7p1	176
GRRGWEALK	B2705	gp160	786	TEKLWVTVY	B4403	gp160	31	IVLPEKDSW	B5801	RT	244
KIRLRPGGK	B2705	p17	18	LYNTVATLY	B4403	p17	78	ITTESIVIW	B5801	RT	375
IRLRPGGKK	B2705	p17	19	EEKAFSPEV	B4403	p24-p2p7p1	28	VSGKARGWF	B5801	Vif	31
RWIILGLNK	B2705	p24-p2p7p1	132	SEGATPQDL	B4403	p24-p2p7p1	44	AVRHFPRIW	B5801	Vpr	30
VRHFPRIWL	B2705	Vpr	31	QEPIDKELY	B4403	p24-p2p7p1	344				
DPNPQEVVL	B3501	gp160	78	EEMSLPGRW	B4403	Protease	34				
RPVVSTQLL	B3501	gp160	252	EELRQHLLR	B4403	RT	203				
TAVPWNASW	B3501	gp160	606	EEKAFSPEV	B4501	p24-p2p7p1	28				
FPVTPQVPL	B3501	Nef	68	AETFYVDGA	B4501	RT	437				
WIYHTQGYF	B3501	Nef	113	DPNPQEVVL	B5101	gp160	78				
YPLTFGWCY	B3501	Nef	135	LPCRIKQII	B5101	gp160	416				
RPGGKKKYK	B3501	p17	22	RAIEAQQHL	B5101	gp160	557				
WASRELERF	B3501	p17	36	GACRAIRHI	B5101	gp160	835				
HSNQVSQNY	B3501	p17	124	LPPVVAKEI	B5101	Integrase	28				
PPIPVGEIY	B3501	p24-p2p7p1	122	YFPDWQNYT	B5101	Nef	120				

Table A.3: Continued

A.2 The HIV HXB2 Proteome

The viral strain HXB2 (Table A.4) (GenBank Accession Number K03455) is used as a reference strain for the HIV epitope datasets in Chapter 4. The position of the defined epitope location relative to the sequence of the HXB2 protein is indicated in these datasets. HXB2 was selected as the reference strain because so many studies use HXB2, and because crystal structures for HXB2-related proteins are available.

A.3 The Effect of Set Size on ROC Curves

From Section 4.3.6, we examined the effect of a reduced negative dataset on rescaling and we found that rescaling still significantly reduced epitope detection accuracy. Figure A.1 and Figure A.2 show a more general examination of the effect on AUC values when the positive and negative set sizes are altered.

We used epitope data from the SYF[1] dataset (Section 4.2.2) to test two effects: the change in AUC values when the positive (epitope) set size is changed with a constant negative (random peptide) set size, and vice versa. Figure A.1 shows that changing the positive set size has no significant effect on the AUC values. However, increasing the negative set size significantly increases the AUC value (Figure A.2, $P < 0.001$). The reason for this is that, assuming the positive set size is drawn from the same distribution, the proportion of epitopes discovered as the threshold changes would not vary even as the positive set size increases. However, increasing the negative set size increases the ratio of true positives to true negatives at any threshold, which would result in a larger AUC value. Although this may be viewed as a trivial result, it is worth bearing in mind when comparing AUC values across different datasets.

gp160 MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYYGVPVWKEATT TLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLVNVTENFNMWKNDM VEQMHEDIISLWDQSLKPCVKLTPLCVSLKCTDLKNDTNTNSSSGRMIME KGEIKNCSFNISTSIRGKVQKEYAFFYKLDIIPIDNDTTSYKLTSCNTSV ITQACPKVSFEPIPIHYCAPAGFAILKCNNKTFNGTGPCTNVSTVQCTHG IRPVVSTQLLLNGSLAEEEVVIRSVNFTDNAKTIIVQLNTSVEINCTRPN NNTRKRIRIQRGPGRAFVTIGKIGNMRQAHCNISRAKWNNTLKQIASKLR EQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLFNSTWFNSTW STEGSNNTEGSDTITLPCRIKQIINMWQKVGKAMYAPPISGQIRCSSNIT GLLLTRDGGNSNNESEIFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTK AKRRVVQREKRAVGIGALFLGFLGAAGSTMGAASMTLTVQARQLLSGIVQ QQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSG KLICTTAVPWNASWSNKSLEQIWNHTTWMEWDREINNYTSLIHSLIEESQ NQQEKNEQELLELDKWASLWNWFNITNWLWYIKLFIMIVGGLVGLRIVFA VLSIVNRVRQGYSPLSFQTHLPTPRGPDRPEGIEEEGGERDRDRSIRLVN GSLALIWDDLRSLCLFSYHRLRDLLLIVTRIVELLGRRGWEALKYWWNLL QYWSQELKNSAVSLLNATAIAVAEGTDRVIEVVQGACRAIRHIPRRIRQG LERILL
Integrase FLDGIDKAQDEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKGEAM HGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYF LLKLAGRWPVKTIHTDNGSNFTGATVRAACWWAGIKQEFGIPYNPQSQGV VESMNKELKKIIGQVRDQAEHLKTAVQMAVFIHNFKRKGGIGGYSAGERI VDIIATDIQTKELQKQITKIQNFRVYYRDSRNPLWKGPAKLLWKGEGAVV IQDNSDIKVVPRRKAKIIRDYGKQMAGDDCVASRQDED
Nef MGGKWSKSSVIGWPTVRERMRRAEPAADRVGAASRDLEKHGAITSSNTAA TNAACAWLEAQEEEEVGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGL IHSQRRQDILDLWIYHTQGYFPDWQNYTPGPGVRYPLTFGWCYKLVPVEP DKIEEANKGENTSLLHPVSLHGMDDPEREVLEWRFDSRLAFHHVARELHP EYFKNC
p17 MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGL LETSEGCRQILGQLQPSLQTGSEELRSLYNTVATLYCVHQRIEIKDTKEA LDKIEEEQNKSKKKAQQAAADTGHSNQVSQNY
p24-p2p7p1p6 PIVQNIQGQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQ DLNTMLNTVGGHQAAMQMLKETINEEAAEWDRVHPVHAGPIAPGQMREPR GSDIAGTTSTLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPTSI LDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKT ILKALGPAATLEEMMTACQGVGGPGHKARVLAEAMSQVTNSATIMMQRGN FRNQRKIVKCFNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDCTERQAN FLGKIWPSYKGRPGNFLQSRPEPTAPPEESFRSGVETTTPPQKQEPIDKE LYPLTSLRSLFGNDPSSQ

Table A.4: The HIV HXB2 proteome

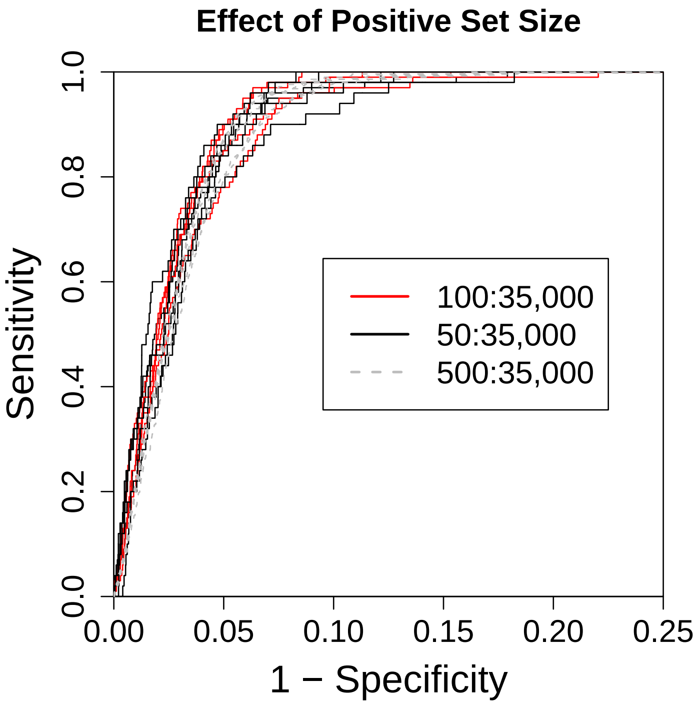


Figure A.1: The effect of positive set size on AUC values. The data for each ROC curve was randomly assigned from the SYF[1] dataset. For each size classification, 5 random samples of positive and negative data was used.

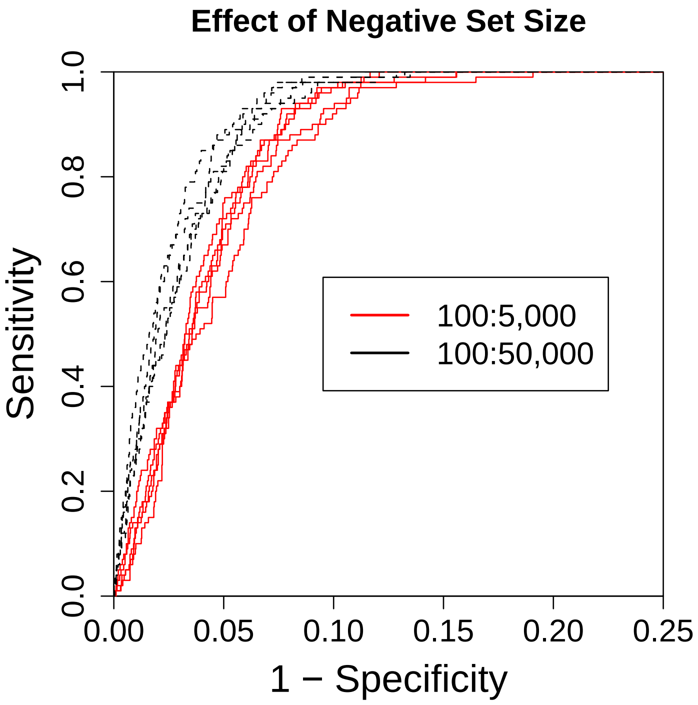


Figure A.2: The effect of neagative set size on AUC values. The positive and negative data was selected as in Figure A.1.

Protease PQVTLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
Rev MAGRSGDSDEELIRTVRLIKLLYQSNPPPNPEGTRQARRNRRRRWRERQRQIHSISERILGTYLGRSAEPVPLQLPPLERLTLDCNEDCGTSGTQGVGSPQILVESPTVLESGTKE
RT PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGL

Table A.4: Continued

Appendix B

Supplementary Data for Chapter 5

B.1 The SIR Score

The SIR (Size of the Immune Repertoire) score [190] is a permutation measure where the number of predicted binders per allele is compared for each HTLV-I protein against a random protein of the same size. Briefly, for each protein-allele combination, the number of predicted strong binders ($< 50\ nM$) and weak binders ($< 500\ nM$) was found for both the consensus HTLV-I sequence and a randomized counterpart (using the amino acid frequencies present in HTLV-I). This ratio could then be used in place of the rank measure (Section 5.2.4). There are a number of reasons though why this measure is not as effective as the rank measure for detecting the strength of binding to a given protein. Firstly, the SIR measure fails to take into account competition with other peptides for the same allele and secondly, it relates more to evolution i.e. does this protein sequence contain more epitopes than would be expected? Therefore, one sequence could have a higher SIR score than another while containing fewer epitopes. However, it is useful to use as many independent methods as possible. Most of the tests outlined in Section 5.3 were repeated using the SIR measure (Table B.1).The only finding that was not replicated was that the protective alleles A*0201 and Cw*08 bind HBZ more strongly than B*5401. We believe this is probably a question of power as we tested this hypothesis using a chi-squared table of the form:

	A*0201 & Cw*0801	**B5401**
# Predicted Epitopes	A	B
# Random Predicted Epitopes	C	D

But the numbers A-D were small (HBZ being poorly bound by most alleles, see Section 5.4).

B.2 Multiple and logistic regression

Chapter 5, Section 5.3.5 described the power of peptide binding to predict proviral load, compared to associations with HLA genotype. For the analysis in Section 5.3.5, the binding strength of an individual's HLA class I repertoire to a specific HTLV-I protein was defined as the median rank value (see Section 5.2.4) of the individual's A and B alleles. Table B.2 compares this measure against other metrics defining the strength of binding of an individual's HLA class I alleles to a specific protein. Each model was tested on the HAM/TSP and AC groups separately. The HTLV-I proteins in the table are whose peptide - MHC class I binding ranks significantly explain a proportion (R^2) of the variance in proviral load. 5 different metrics were used. 'Count' refers to the number of strong binding alleles that each individual posseses for a specific protein (Section 5.2.4). 'Median Rank' and 'Median Affinity' refer to the median rank value or median raw predicted binding affinity of the individual's 4 A and B alleles (Section 5.2.9). Finally, 'Max Rank' and 'Max Affinity' refer to the highest rank value or median raw predicted binding affinity of the individual's 4 A and B alleles.

The targeting of HBZ is a significant predictor of proviral load in 5/10 of these models. In each case, the HBZ parameter coefficient is the 'correct' sign for a beneficial effect of targeting HBZ: positive for the rank models where a low rank corresponds to strong binding and negative for the raw binding affinity scores where a high score corresponds to strong binding.

We also tested some logistic regression models that predicted the disease status (asymptomatic or HAM/TSP) using predicted binding specificity to HTLV-I proteins as predictors. As in Table B.2, Table B.3 shows the metric used for each model. As mentioned in Section 5.3.7, these models are weakly predictive of disease status because of the many other determinants involved. However, they do illustrate again the effect of targeting HBZ, with 3/4 of the models showing HBZ as a significant beneficial predictor of HAM/TSP risk.

B.3 Methods of epitope definition

One of the advantages of the rank statistic (Section 5.2.4) that we developed is that it negates the necessity of quantifying a predicted binding affinity threshold to define an

	Null hypothesis	SIR < 50	SIR < 500	Conclusion
1	Protective and detrimental alleles target HBZ equally	-	-	Not enough power for this test using the SIR measure
2	AC and HAM/TSP patients target HBZ equally	0.07	0.004	ACs have HLA alleles that bind HBZ significantly more strongly compared to HAM/TSP patients
3	There is no correlation between proviral load and the number of alleles that bind HBZ strongly	0.014	0.01	The higher the number of strong binding alleles to HBZ per individual, the lower their proviral load
4	There is no correlation between load reduction (count) and disease prevalence reduction	0.02	0.013	Proteins that are strongly bound by asymptomatic carriers are, independently, those associated with a greater reduction in load when bound

Table B.1: The results of hypothesis testing using the SIR measure to define the strength of binding of HLA class I alleles to HBZ.

Group	Method	R^2	Protein	Effect	P Value
AC	Count	0.044029	Env	0.552145	0.002725
HAM	Count	0.059642	Pol Gag	0.287735 -0.466769	0.044292 0.000276
AC	Median Rank	0.053639	HBZ Pro	0.014960 -0.023219	0.000981 0.012906
HAM	Median Rank	0.025510	HBZ	0.003767	0.016983
AC	Median Affinity	0.025705	HBZ	-1.216011	0.022650
HAM	Median Affinity	0.032798	HBZ	-0.856337	0.006694
AC	Max Rank	0.048615	Rof P21	-0.041019 0.019808	0.025676 0.016166
HAM	Max Rank	0.082557	P12 Gag P13	0.013849 0.127677 -0.028241	0.033824 0.000383 0.048216
HAM	Max Affinity	0.060414	Pol Gag	3.081140 -2.730876	0.003063 0.000299
AC	Max Affinity	0.088883	Rof HBZ	3.057874 -1.983844	0.004153 0.000140

Table B.2: The results of multiple regression analysis to predict proviral load using different metrics defining specificity.

Method	Protein	Effect	P Value
Median Rank Median Rank Median Rank	Pol Tax P13	-0.343359 -0.100128 -0.012620	0.000244 0.031052 0.045861
Max Rank Max Rank	HBZ Pro	-0.011648 0.017276	0.0006 0.0115
Median Score Median Score Median Score	Pol HBZ Pro	2.3074 1.2674 -1.5082	0.029 0.00045 0.0015
Max Score Max Score	Env HBZ	-3.0374 1.2742	0.004 0.00005

Table B.3: The results of logistic regression analysis to predict disease status using different metrics defining specificity.

epitope. However, for reference purposes, it is useful to define these thresholds for the epitope prediction methods that we used. Generally, it is assumed that strong binding epitopes have an affinity $< 50\ nM$ and weak binding epitopes an affinity of $< 500\ nM$ [124]. An alternative and less stringent definition of an epitope threshold can be found from a ROC curve. This is the binding affinity score that defines the most 'top-left' point of the ROC curve i.e. the point at which the difference between the true positive fraction (on the y-axis) and the false negative fraction (on the x-axis) is greatest.

Table B.4 gives the method-specific scores for the 3 epitope prediction methods that we used - Metaserver rescaled, Metaserver non-rescaled and Epipred. The scores are defined for each of the epitope definitions: $< 50\ nM$, $< 500\ nM$ and the ROC curve value.

	Epipred	Metaserver Rescaled	Metaserver Non-Rescaled
ROC curve	-4.0706	0.3324	0.3175
$< 50\ nM$	0.0183	1.1705	1.4348
$< 500\ nM$	-1.3978	0.7860	0.9458

Table B.4: The method-speciific raw binding scores that define the epitope thresholds described in Section B.3. The dataset Lanl^{661} (Section 4.2.2.2) was used to obtain the ROC curve values.

B.4 The relationship between the CD8$^+$ T cell functional response and binding specificity

In Chapter 6, we showed that the lytic efficiency of HTLV-I Tax specific CD8$^+$ T cells is related to the levels of Tax expression in the target cells. This work was performed in collaboration with experimentalists within our laboratory and further work from them included an examination of frequency and functional avidity as predictors of CD8$^+$ lytic efficiency [89]. They showed that the functional avidity of HTLV-I specific CD8$^+$ cells was strongly correlated with their lytic efficiency.

Our use of epitope prediction software demonstrated that strong MHC class I binding of HBZ reduces disease risk and lowers proviral load (Chapter 5). Given that the function of MHC class I is to present viral epitopes to CD8$^+$ T cells, it followed that there may be some relationship between lytic efficiency (ϵ) or functional avidity and our measure of how strongly an individual's MHC class I repertoire binds to peptides from the HBZ protein. To answer this question, we used the dataset from Kattan *et al.* [89] shown in Table B.5. Together with lytic efficiency (ϵ), CD8$^+$ T cell frequency and avidity, the genotype of each individual's HLA class I was typed to a resolution of 4 digits.

Metaserver did not provide sufficient allele coverage for this dataset. Instead, we used another method of epitope prediction: NetMHCpan [191].

As in Metaserver, which encompasses NetCTL and NetMHC (Section 4.4), NetMHCpan uses artificial neural networks to predict binding of peptides to MHC class I molecules. However, it also predicts binding to uncharacterized MHC class I molecules by using amino acid sequence information from their peptide binding sites. Hence, this method provided full coverage of the HLA class I genotypes in Table B.5.

No significant relationship was found between the strength of HBZ binding and either ϵ ($R^2 = 0.024$, $P = 0.52$), frequency ($R^2 = 0.00005$, $P = 0.99$) or avidity ($R^2 = 0.006$, $P = 0.73$). This is not a surprising result as these CD8$^+$ T cell variables were all measured in terms of tax expression (i.e. ϵ is the rate of killing of Tax$^+$CD4$^+$ cells per CD8$^+$ cell, frequency and avidity are based the quantity of IFNγ HTLV-I Tax-specific CD8$^+$ T cells). Figure B.1 shows these relationships for Tax instead of HBZ. In Figure B.1 A, the strength of binding to HTLV-I Tax protein is significantly positively correlated with ϵ ($R^2 = 0.34$, $P = 0.008$). Figure B.1 B shows a positive correlation between the strength of binding to Tax and avidity, although the relationship is not significant ($R^2 = 0.14$, $P = 0.09$).

The correlation between strength of binding to Tax and ϵ is an interesting result that shows a relationship between 2 measures of the CD8$^+$ T cell response to HTLV-I infection. In order to explore this relationship, however, it would be necessary to verify the epitope prediction software NetMHCpan and produce the corresponding Tax data (ϵ, frequency and avidity) for HBZ.

Patient	ϵ	Frequency	Avidity	A1	A2	B1	B2	C1	C2
HAO	NA	0.404	0.967	A2402	A3001	B0702	B5701	C0602	C0701
HAP	0.041	0.197	0.945	A0201	A0301	B0702	B5301	C0701	C0401
HAY	0.015	0.123	0.927	A1101	A0301	B1501	B2702	C0202	C0401
HBE	0.141	0.224	8.643	A0205	A3001	B0705	B5301	C0401	C0702
HBX	0.072	1.442	2.069	A0101	A6801	B0702	B5101	C0202	C1504
HBZ	0.090	0.593	0.385	A2901	A7401	B3501	B3501	C0401	C0401
HCH	0.035	0.099	1.081	A2901	A3601	B4901	B5301	C0401	C0701
HCL	NA	1.719	8.230	A0101	A2501	B0801	B1801	C0701	C1203
HDS	0.292	0.837	12.410	A0201	A0201	B0702	B3503	C0401	C0702
HFB	-0.009	0.528	0.824	A2601	A2601	B3801	B3801	C1202	C1202
HFG	0.147	2.984	9.560	A0201	A6901	B4001	B5501	C0303	C0303
N10	-0.038	0.197	0.945	A0201	A6801	B4402	B5802	C0501	C0602
N11	0.143	0.430	NA	NA	NA	NA	NA	NA	NA
N12	0.166	1.260	4.792	A0201	A3002	B3910	B5703	C0701	C1203
TAK	0.038	0.458	1.560	A0201	A0301	B5601	NA	C0102	NA
TAW	0.307	1.685	33.681	A0101	A2601	B4101	B1801	C1701	C1203
TAZ	0.036	1.354	1.370	A2402	A2601	B5101	B5201	C1202	C1402
TBR	0.072	0.814	3.781	A2402	A2402	B1515	B3906	C0102	C0702
TCF	0.046	0.178	0.825	A0301	A3301	B4501	B5301	C0401	C1601
TCI	0.167	0.552	1.627	A3601	A6802	B1510	B5301	C0401	C0802
TCJ	0.051	0.239	NA	A0301	A2301	B4403	B4501	C0401	C1601
TCL	0.063	0.172	2.621	A0201	A2301	B0705	B4501	C0702	C1601
TCP	NA	0.198	10.983	A0205	A2301	B1516	B5301	C0401	C1402
TCR	NA	1.676	34.223	A3301	A3601	B1516	B5703	C0701	C1402

Table B.5: The details of the HTLV-I infected individuals used in Section B.4. ϵ is the proportion of Tax-expressing CD4$^+$ cells killed per CD8$^+$ cell per day. The frequency is IFNγ^+ Tax-specific CTLs [89]. Avidity (10^6 M^{-1}) is calculated as in [89].

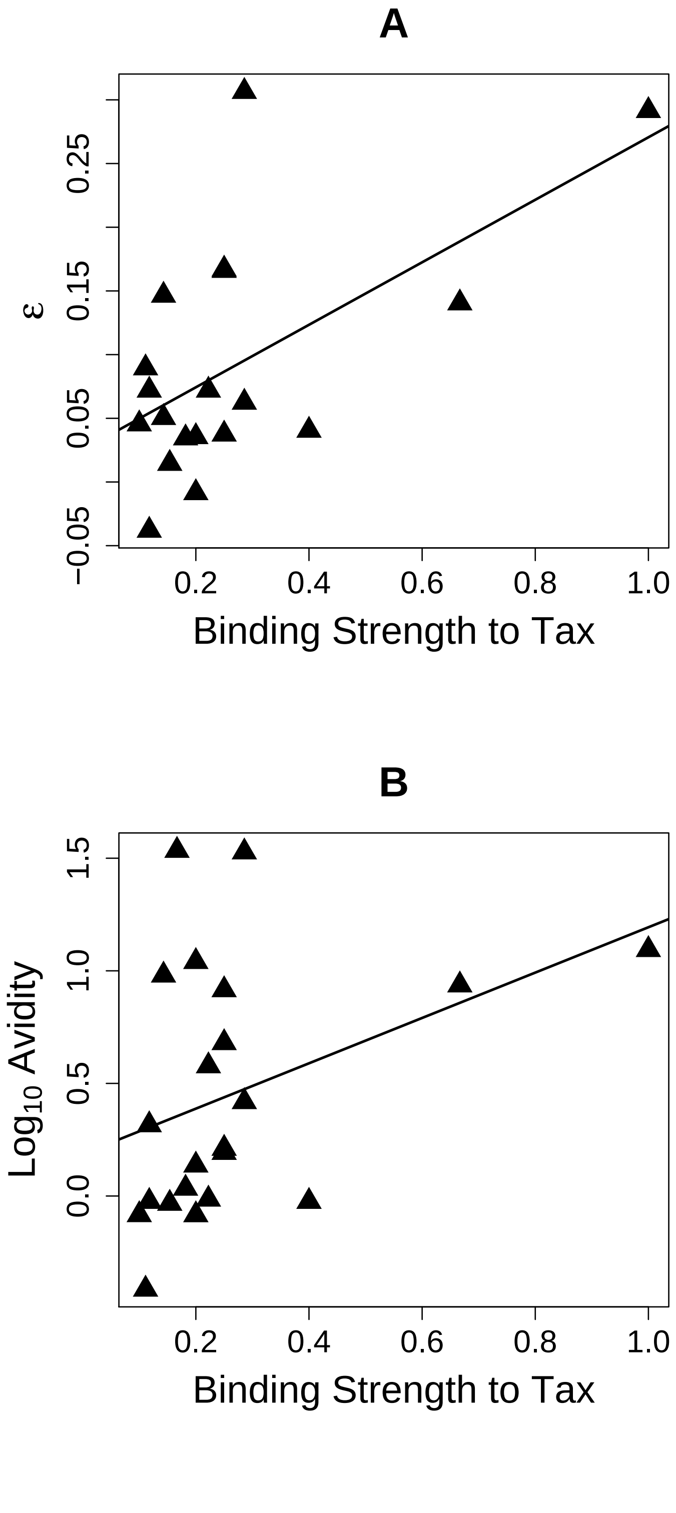


FIGURE B.1: A: The relationship between ϵ and the binding strength to Tax ($R^2 = 0.34$, $P = 0.008$). The binding strength to Tax is the reciprocal of the median rank value of the top binding Tax peptide to the individual's HLA class I repertoire (A, B and C alleles). B: The relationship between avidity (10^6 M^{-1}) and the binding strength to Tax ($R^2 = 0.14$, $P = 0.09$).

	Pol		Env		Rof		Tax		P12		Rex	
alleles	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity
A0101	HTDPRDQIY	10	SSSSTPLLY	14	PSDVSGLLL	1153	GSVVCMYLY	703	PSDVSGLLL	1153	QSTCLETVY	917
A02	ALLGEIQWV	7.86207	ALQTGITLV	14.1921	ILSGLLFLL	16.1626	YLYQLSPPI	5.06404	ILSGLLFLL	16.1626	IVTPYWPPV	153.291
A0201	YLYHYLRTL	10	VLYSPNVSV	11	ILSGLLFLL	13	LLFGYPVYV	3	ILSGLLFLL	13	SMDALSAQL	72
A0203	YLYHYLRTL	3	FLNTEPSQL	2	MLFRLLSPL	2	YLYQLSPPI	3	MLFRLLSPL	2	SLGDYVRPA	4
A0206	FQPYFAFTV	2	AQPVCSWTL	8	MLFRLLSPL	15	FLIPRLPSF	4	MLFRLLSPL	15	AQLYSSLSL	7
A0301	LLYKYFTDK	7	YLFPHWTKK	9	TMRFPARWR	90	FIFHKFQTK	12	TMRFPARWR	90	STCLETVYK	665
A1101	TTTVVFQSK	7	SSSSTPLLY	24	STMLFRLLS	227	QSSSFIFHK	7	PITMRFPAR	2025	STCLETVYK	6
A2402	KYKNTLYRL	40	KFLATLILF	17	LFLLFLPLF	53	SFHSLHLLF	14	LFLLFLPLF	53	PYWPPVQSI	224
A26	WTINHLNVL	87.3461	YTGAVSSPY	73.1346	PTPWQLPPF	3523.25	TTPGLIWTF	173.75	QILSGLLFL	4815.65	TGAPSLGDY	7105.98
A2601	WTINHLNVL	105	YTGAVSSPY	83	PTPWQLPPF	4259	TTPGLIWTF	203	QILSGLLFL	5815	TGAPSLGDY	8542
A2602	FTVPQQCNY	2	EVDKDISQL	5	AINPQLLHF	8	FLIPRLPSF	6	QILSGLLFL	41	TGAPSLGDY	245
A3001	RPWARTPPK	8	YLFPHWTKK	11	RRRPRRSQR	102	AMRKYSPFR	4	RFPARWRFL	1347	STCLETVYK	58
A3101	KVVYLHHVR	7	RQLRHLPSR	6	MPKTRRRPR	17	AMRKYSPFR	3	TMRFPARWR	39	MPKTRRRPR	17
A3301	ILRSCHACR	23	DLGLSQWAR	7	MPKTRRRPR	50	AMRKYSPFR	119	PITMRFPAR	180	MPKTRRRPR	50
B0702	APRNQPVPF	8	VPSSSSTPL	5	APSQPAAAF	8	SARLHRHAL	7	APSQPAAAF	8	RPAYIVTPY	60
B1501	KQFQPYFAF	46	YTGAVSSPY	81	SLQGLHLAF	110	GQHLPTLSF	56	FQILSGLLF	61	LSACTSTSF	221
B3501	FPQCTILQY	14	LPFNWTHCF	6	APSQPAAAF	14	IPPSFLQAM	19	APSQPAAAF	14	RPAYIVTPY	8
B3901	YHYLRTLAL	6	YHATYSLYL	10	LHFFFPSTM	2596	IQYSSFHSL	147	MRFPARWRF	12998	YKATGAPSL	107
B40	FEMQLAHIL	44.9888	WKFQHDVNF	3789.06	FQILSGLLF	4594.8	EELLYKISL	39.8315	FQILSGLLF	4594.8	AQLYSSLSL	8671.56
B4001	FEMQLAHIL	11	AQPVCSWTL	1872	FQILSGLLF	3488	EELLYKISL	25	FQILSGLLF	3488	AQLYSSLSL	636
B4002	FEMQLAHIL	66	REALQTGIT	654	FFFPSTMLF	8666	EELLYKISL	49	FQILSGLLF	5279	GEAPLSACT	2863
B4402	LEAGHIEPY	593	LENRVLTGW	1253	SQPAAAFLF	8074	EELLYKISL	1037	SQPAAAFLF	8074	MDALSAQLY	12389
B5101	LPMDNALSI	13	LPPFSLSPV	11	LPLLLSPSL	26	LPTTLFQPA	42	LPLLLSPSL	26	TPWPTSQGL	445
B5401	MPVFTLSPV	4	FPFSLLVDA	2	LPITMRFPA	2	LPTTLFQPA	2	LPITMRFPA	2	FPPPSPGPS	59
B5801	LTNCHKTRW	32	YSLYLFPHW	13	ITMRFPARW	7	RVIGSALQF	32	ITMRFPARW	7	LSACTSTSF	74

Table B.6: The top ranking MHC class I - peptide pairings for the alleles of the Kagoshima Cohort, according to the rank method (Section 5.2.4).

	HBZ		Gag		Pro		Tof		P13		P21	
alleles	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity	peptide	affinity
A0101	MQELGIDGY	6029	WLNFLQAAY	2442	LQQCQGVLY	6429	FSSSFLFKY	32	YRLSSTVPY	20283	LSACTSTSF	6727
A02	AVLDGLLSL	19.9557	FMQTIRLAV	8.75862	ALFSSNTPL	18.7783	LIISPLPRV	42.4385	LIISPLPRV	42.4385	AQLYSSLSL	352.256
A0201	AVLDGLLSL	31	FMQTIRLAV	7	ALFSSNTPL	11	LTMLIISPL	73	LIISPLPRV	63	EMDTWNPPL	86
A0203	ELVDGLLSL	19	FMQTIRLAV	5	RLPFRTTPI	3	LTMLIISPL	21	RLVPHLWGT	11	ALSAQLYSS	61
A0206	AVLDGLLSL	3	WQMKDLQAI	2	IQAPAVLGL	29	RLVPHLWGT	2	RLVPHLWGT	2	AQLYSSLSL	7
A0301	KAKQHSARK	145	SLLASLLPK	18	GITQYSQLK	77	SLSFNSSSK	19	SSFRIPSLR	87	PTFHPPSSR	1782
A1101	KAADVARRK	55	SSYDFHQLK	4	GITQYSQLK	34	SSFSRSFFR	3	SSFRIPSLR	14	PTFHPPSSR	482
A2402	YWQGRLEAM	4434	GYPGRVNEI	40	PFRTTPIVL	2930	SFSRSFFRF	12	VWTESSFRI	35	QSLIQPPTF	2171
A26	ELVDGLLSL	52.8654	ETPARICPI	182.25	DTKNNWAII	190.25	LVPHLWGTM	1229.29	LVPHLWGTM	1229.29	CTPSGEAPL	9512.65
A2601	ELVDGLLSL	61	ETPARICPI	216	DTKNNWAII	215	FSSSFLFKY	2530	LVPHLWGTM	1478	CTPSGEAPL	11277
A2602	ELVDGLLSL	14	HTNSPLGDM	13	MTVLPIALF	4	LVPHLWGTM	41	LVPHLWGTM	41	EMDTWNPPL	908
A3001	KAKQHSARK	6	KVLVVQPKK	5	LFSSNTPLK	57	AFSSSFLFK	43	MLIISPLPR	683	RSLPRQSLI	1219
A3101	RQRRAEEKR	43	WSRDCTQPR	61	NTWSGRPWR	16	SSFSRSFFR	2	RVWRLCTRR	4	PTFHPPSSR	427
A3301	EVESLEAER	23	EPYHAFVER	89	NTWSGRPWR	11	SSFSRSFFR	7	HLGPHRWTR	12	PTFHPPSSR	1332
B0702	LPVSCPEDL	171	RPPPGPCPL	9	LPVLIRLPF	7	GPRRSRPRL	9	VPYPSTPLL	33	PPSPPREPL	50
B1501	MQELGIDGY	230	YQQLWLAAF	53	LQQCQGVLY	147	FLLATSAAF	65	YRLSSTVPY	1909	LSACTSTSF	221
B3501	MQELGIDGY	492	DPILRSLAY	10	LPVLIRLPF	9	VPHLWGTMF	12	VPHLWGTMF	12	MDALSAQLY	401
B3901	DKEEEKAVL	1540	FVERLNIAL	330	SHPKTIEAL	219	RRAFSSSFL	2708	PHRWTRYRL	5300	EMDTWNPPL	507
B40	EEKQIAEYL	203.775	LEEPYHAFV	434.978	PEAKRPPVI	4188.22	RAFSSSFLF	13314	YRLSSTVPY	35018.8	AQLYSSLSL	8671.56
B4001	VEELVDGLL	260	EEDALLLDL	112	GQTQDHFKL	406	RAFSSSFLF	17075	YRLSSTVPY	36290	AQLYSSLSL	636
B4002	EEKQIAEYL	148	REYQQLWLA	248	PEAKRPPVI	2102	RAFSSSFLF	10989	VPHLWGTMF	21041	GEAPLSACT	2863
B4402	EEEKQIAEY	295	AETRGITGY	231	LQQCQGVLY	16483	SFSRSFFRF	8791	SDHLGPHRW	7122	MDALSAQLY	12389
B5101	LPVSCPEDL	246	LPKGYPGRV	185	LPFRTTPIV	21	VPYPSTPLL	204	VPYPSTPLL	204	LPRQSLIQP	375
B5401	LPVSCPEDL	2433	LPVMHPHGA	7	LPVIPLDPA	2	RPTGHLSRA	1054	RPTGHLSRA	1054	FPPPSPGPS	59
B5801	DLMGEVNYW	1488	QAAPGSPQF	70	NASRPCNTW	140	RAFSSSFLF	9	IISPLPRVW	31	LSACTSTSF	74

Table B.6: Continued

Peptide	A0201	B0702	A2402	B3501	Peptide	A0201	B0702	A2402	B3501
AAGAALIPV	YES				GLLSLEEEL	YES			
AAHHWLNFL	YES	YES		YES	GYPGRVNEI			YES	
AASGLFRCL		YES		YES	GYTRQLEGE			YES	
ALLGEIQWV	YES				HPGQLGAFL		YES		YES
APLPHTSQC		YES			HQLKKFLKI	YES		YES	
APPPPSSPT		YES			IALETPARI	YES		YES	
ASGLFRCLP	YES	YES	YES	YES	ICPINYSLL			YES	
AVLDGLLSL	YES	YES	YES	YES	IFSRSASPI			YES	
AWQNGLLPF			YES		ILIQTQAQI	YES			
CPINYSLLA				YES	ILPEDCLPT	YES			
CPLCQDPTH				YES	IPPSFLQAM		YES		YES
DLMGEVNYW			YES		IPRLPSFPT		YES		
DPILRSLAY				YES	IPRPPRGLA		YES		
DPISRLNAL		YES		YES	IQYSSFHSL	YES	YES	YES	
EEEKQIAEY				YES	IWQGDITHF			YES	
EKAVLDGLL			YES		KALMPVFTL	YES	YES	YES	YES
ELVDGLLSL	YES			YES	KARRRRRAE		YES		
EPEEDALLL				YES	KISLTTGAL	YES	YES		
EPEPEEDAL				YES	KLLQEKEDL	YES			
EPGPSSYDF				YES	KQIAEYLKR			YES	
EYLKRKEEE			YES		KYKNTLYRL			YES	
EYQQLWLAA			YES		KYLYHYLRT			YES	
EYTNIPISL			YES		KYTLQSYGL			YES	
FMQTIRLAV	YES				LAAHHWLNF		YES	YES	YES
FPGFGQSLL		YES		YES	LASLLPKGY				YES
FPQCTILQY				YES	LEAERRKLL		YES		
FPTQRTSKT				YES	LLFGYPVYV	YES			
FVERLNIAL	YES	YES		YES	LLITPVLQL	YES			
GFGQSLLFG	YES	YES	YES	YES	LLLDLPADI	YES			
GIDGYTRQL	YES				LLQEKEDLM	YES			YES

Table B.7: The HTLV-I peptides that were selected for the REVEAL$^{\text{TM}}$ MHC-peptide binding assay, for each allele. These were compared against the predicted binding affinities of Metaserver and Epipred.

Peptide	A0201	B0702	A2402	B3501	Peptide	A0201	B0702	A2402	B3501
LLQYLCSSL	YES				QPARAPVTL		YES		YES
LLYKISLTT	YES				QPIPETRSL		YES		
LMGEVNYWQ	YES				QPRPPPGPC		YES		
LPEDCLPTT				YES	QYLCSSLVA			YES	
LPFHSTLTT		YES		YES	RAEKKAADV		YES		
LPGLNSRQW		YES			RDRQRRAEE		YES		
LPTTLFQPA				YES	RGRLRRGPP		YES		
LPVMHPHGA		YES			RICPINYSL	YES	YES	YES	
LPVSCPEDL		YES		YES	RPAPPPPSS		YES		
LQYLCSSLV	YES				RPPPGPCPL		YES		
LSPPITWPL	YES		YES	YES	RPPRGLAAH		YES		YES
LTPPITHTT	YES			YES	RRRAEKKAA		YES		
LVEELVDGL	YES			YES	RVIGSALQF		YES	YES	YES
LVLQSSSFI	YES	YES			RVNEILHIL	YES	YES		YES
LWLAAFAAL			YES		SAQWIPWRL	YES	YES	YES	YES
MPVFTLSPV		YES		YES	SARLHRHAL		YES		
MQELGIDGY				YES	SFHSLHLLF			YES	
NFLQAAYRL			YES	YES	SFLLSHGLI			YES	
NYSLLASLL			YES		SLVQLRQAL	YES	YES		YES
PPNHRPWQM		YES			STLTTPGLI	YES		YES	
PYHAFVERL			YES		SWASILQGL			YES	
PYKRIEELL			YES		SYGLLCQTI			YES	
PYNPTSSGL			YES		TFLKTAAPL		YES	YES	
QAAPGSPQF				YES	TLGQHLPTL	YES			
QAMRKYSPF		YES	YES	YES	TLSFPDPGL	YES			
QLDSLISEA	YES				TLTAWQNGL	YES			
QLEGEVESL	YES				TLYRLHVWV	YES			
QLGAFLTNV	YES				TPKDKTKVL		YES		
QLLASAVLL	YES				TPNIPPSFL		YES		
QLWLAAFAA	YES				TTPGLIWTF			YES	

Table B.7: Continued

Peptide	A0201	B0702	A2402	B3501
TTPNIPPSF			YES	
TWPLLPHVI			YES	
VFTLSPVII			YES	
VLQSSSFIF			YES	
VPIRSRWAL		YES		
VPYKRIEEL		YES		
VSCPEDLLV			YES	
WALPELQAL				YES
WPLLPHVIF				YES
WQGRLEAMW			YES	
WQMKDLQAI	YES		YES	
WTFTDGTPM				YES
WTINHLNVL				YES
YILWDKQIL	YES			YES
YISQDFLNM	YES			YES
YLCSSLVAS	YES			
YLYQLSPPI	YES			
YPGRVNEIL		YES		YES
YWQGRLEAM			YES	YES

Table B.7: Continued

Gag MGQIFSRSASPIPRPPRGLAAHHWLNFLQAAYRLEPGPSSYDFHQLKKFL KIALETPARICPINYSLLASLLPKGYPGRVNEILHILIQTQAQIPSRPAP PPPSSPTHDPPDSDPQIPPPYVEPTAPQVLPVMHPHGAPPNHRPWQMKDL QAIKQEVSQAAPGSPQFMQTIRLAVQQFDPTAKDLQDLLQYLCSSLVASL HHQQLDSLISEAETRGITGYNPLAGPLRVQANNPQQQGLRREYQQLWLAA FAALPGSAKDPSWASILQGLEEPYHAFVERLNIALDNGLPEGTPKDPILR SLAYSNANKECQKLLQARGHTNSPLGDMLRACQTWTPKDKTKVLVVQPKK PPPNQPCFRCGKAGHWSRDCTQPRPPPGPCPLCQDPTHWKRDCPRLKPTI PEPEPEEDALLLDLPADIPHPKNFIGGEV
Env MGKFLATLILFFQFCPLIFGDYSPSCCTLTIGVSSYHSKPCNPAQPVCSW TLDLLALSADQALQPPCPNLVSYSSYHATYSLYLFPHWTKKPNRNGGGYY SASYSDPCSLKCPYLGCQSWTCPYTGAVSSPYWKFQHDVNFTQEVSRLNI NLHFSKCGFPFSLLVDAPGYDPIWFLNTEPSQLPPTAPPLLPHSNLDHIL EPSIPWKSKLLTLVQLTLQSTNYTCIVCIDRASLSTWHVLYSPNVSVPSS SSTPLLYPSLALPAPHLTLPFNWTHCFDPQIQAIVSSPCHNSLILPPFSL SPVPTLGSRSRRAVPVAVWLVSALAMGAGVAGGITGSMSLASGKSLLHEV DKDISQLTQAIVKNHKNLLKIAQYAAQNRRGLDLLFWEQGGLCKALQEQC RFPNITNSHVPILQERPPLENRVLTGWGLNWDLGLSQWAREALQTGITLV ALLLLVILAGPCILRQLRHLPSRVRYPHYSLIKPESSL
Pro HPTPKKLHRGGGLTSPPTLQQVLPNQDPASILPVIPLDPARRPVIKAQVD TQTSHPKTIEALLDTGADMTVLPIALFSSNTPLKNTSVLGAGGQTQDHFK LTSLPVLIRLPFRTTPIVLTSCLVDTKNNWAIIGRDALQQCQGVLYLPEA KRPPVILPIQAPAVLGLEHLPRPPEISQFPLNQNASRPCNTWSGRPWRQA ISNPTPGQGITQYSQLKRPMEPGDSSTTCGPLTL
Pol GKKAACNLANTGASRPWARTPPKAPRNQPVPFKPERLQALQHLVRKALEA GHIEPYTGPGNNPVFPVKKANGTWRFIHDLRATNSLTIDLSSSSPGPPDL SSLPTTLAHLQTIDLRDAFFQIPLPKQFQPYFAFTVPQQCNYGPGTRYAW KVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDDILLASPSHED LLLLSEATMASLISHGLPVSENKTQQTPGTIKFLGQIISPNHLTYDAVPT VPIRSRWALPELQALLGEIQWVSKGTPTLRQPLHSLYCALQRHTDPRDQI YLNPSQVQSLVQLRQALSQNCRSRLVQTLPLLGAIMLTLTGTTTVVFQSK EQWPLVWLHAPLPHTSQCPWGQLLASAVLLLDKYTLQSYGLLCQTIHHNI STQTFNQFIQTSDHPSVPILLHHSHRFKNLGAQTGELWNTFLKTAAPLAP VKALMPVFTLSPVIINTAPCLFSDGSTSRAAYILWDKQILSQRSFPLPPP HKSAQRAELLGLLHGLSSARSWRCLNIFLDSKYLYHYLRTLALGTFQGRS SQAPFQALLPRLLSRKVVYLHHVRSHTNLPDPISRLNALTDALLITPVLQ LSPAELHSFTHCGQTALTLQGATTTEASNILRSCHACRGGNPQHQMPRGH IRRGLLPNHIWQGDITHFKYKNTLYRLHVWVDTFSGAISATQKRKETSSE AISSLLQAIAHLGKPSYINTDNGPAYISQDFLNMCTSLAIRHTTHVPYNP TSSGLVERSNGILKTLLYKYFTDKPDLPMDNALSIALWTINHLNVLTNCH KTRWQLHHSPRLQPIPETRSLSNKQTHWYYFKLPGLNSRQWKGPQEALQE AAGAALIPVSASSAQWIPWRLLKRAACPRPVGGPADPKEKDLQHHG

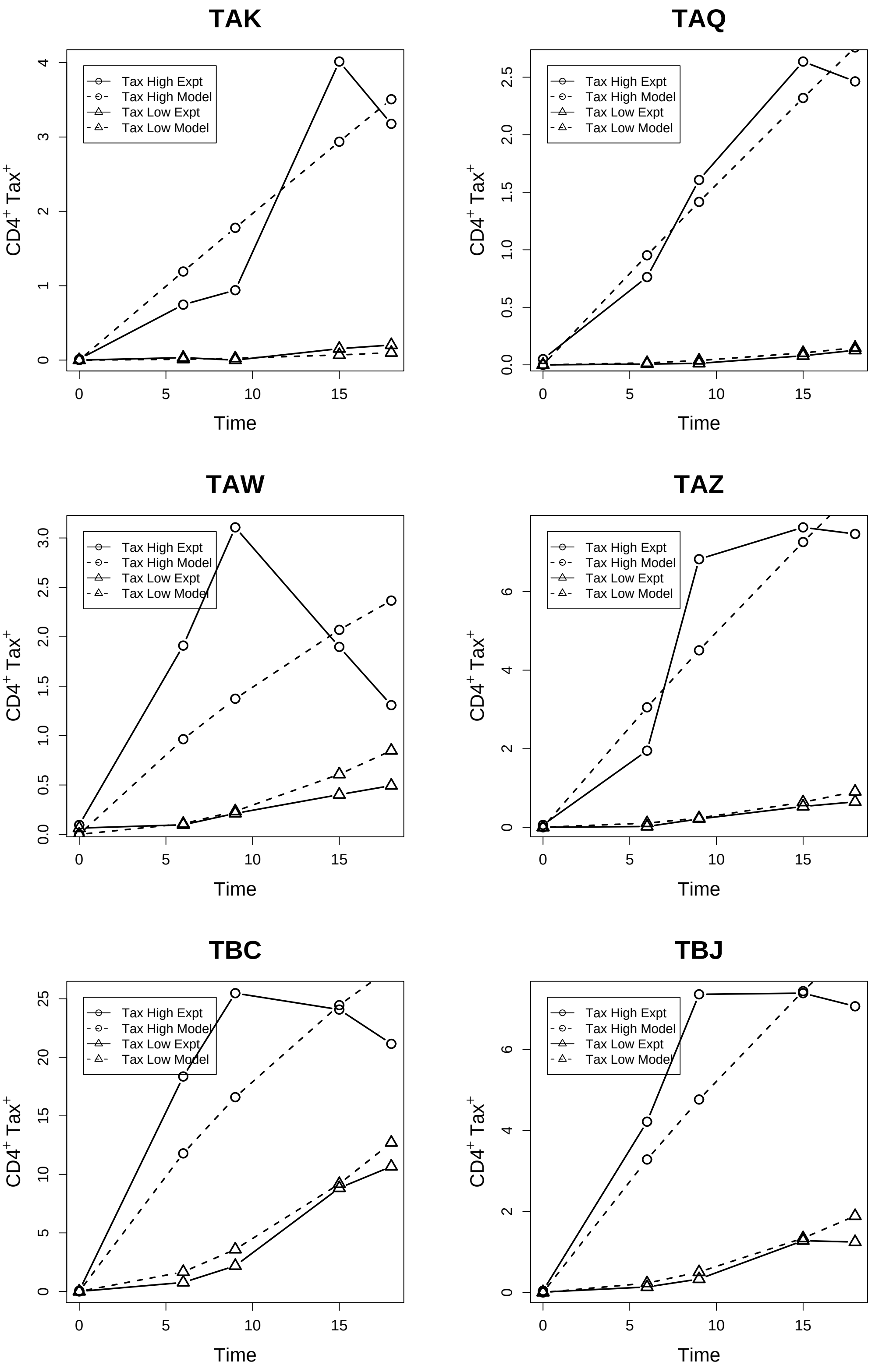
Table B.8: The HTLV-I proteome. This reference strain is from [154], with the exception of HBZ, which was identified more recently and described in [155].

Rof MPKTRRRPRRSQRKRPPTPWQLPPFSLQGLHLAFQLSSIAINPQLLHFFFPSTMLFRLLSPLSPLALTALLLFLLPPSDVSGLLLRPPPAPCLLLFLPFQILSGLLFLLFLPLFFSLPLLLSPSLPITMRFPARWRFLPWRAPSQPAAAFLF
P12 MLFRLLSPLSPLALTALLLFLLPPSDVSGLLLRPPPAPCLLLFLPFQILSGLLFLLFLPLFFSLPLLLSPSLPITMRFPARWRFLPWRAPSQPAAAFLF
Tof MALCCFAFSAPCLHLRSRRSCSSCFLLATSAAFFSARLLRRAFSSSFLFKYSAVCFSSSFSRSFFRFLFSSARRCRSRCVSPRGGAFSPGGPRRSRPRLSSSKDSKPSSTASSSSLSFNSSSKDNSPSTNSSTSRSSGHDTGKHRNSPADTKLTMLIISPLPRVWTESSFRIPSLRVWRLCTRRLVPHLWGTMFGPPTSSRPTGHLSRASDHLGPHRWTRYRLSSTVPYPSTPLLPHPENL
P13 MLIISPLPRVWTESSFRIPSLRVWRLCTRRLVPHLWGTMFGPPTSSRPTGHLSRASDHLGPHRWTRYRLSSTVPYPSTPLLPHPENL
Rex MPKTRRRPRRSQRKRPPTPWPTSQGLDRVFFSDTQSTCLETVYKATGAPSLGDYVRPAYIVTPYWPPVQSIRSPGTPSMDALSAQLYSSLSLDSPPSPPREPLRPSRSLPRQSLIQPPTFHPPSSRPCANTPPSEMDTWNPPLGSTSQPCLFQTPDSGPKTCTPSGEAPLSACTSTSFPPPSPGPSCPT
P21 MDALSAQLYSSLSLDSPPSPPREPLRPSRSLPRQSLIQPPTFHPPSSRPCANTPPSEMDTWNPPLGSTSQPCLFQTPDSGPKTCTPSGEAPLSACTSTSFPPPSPGPSCPT
Tax MAHFPGFGQSLLFGYPVYVFGDCVQGDWCPISGGLCSARLHRHALLATCPEHQITWDPIDGRVIGSALQFLIPRLPSFPTQRTSKTLKVLTPPITHTTPNIPPSFLQAMRKYSPFRNGYMEPTLGQHLPTLSFPDPGLRPQNLYTLWGGSVVCMYLYQLSPPITWPLLPHVIFCHPGQLGAFLTNVPYKRIEELLYKISLTTGALIILPEDCLPTTLFQPARAPVTLTAWQNGLLPFHSTLTTPGLIWTFTDGTPMISGPCPKDGQPSLVLQSSSFIFHKFQTKAYHPSFLLSHGLIQYSSFHSLHLLFEEYTNIPISLLFNEKEADDNDHEPQISPGGLEPPSEKHFRETEV
HBZ MAASGLFRCLPVSCPEDLLVEELVDGLLSLEEELKDKEEEKAVLDGLLSLEEESRGRLRRGPPGEKAPPRGETHRDRQRRAEEKRKRKKEREKEEEKQIAEYLKRKEEEKARRRRRAEKKAADVARRKQEEQERRERKWRQGAEKAKQHSARKEKMQELGIDGYTRQLEGEVESLEAERRKLLQEKEDLMGEVNYWQGRLEAMWLQ

Table B.8: Continued

Appendix C

Supplementary Data for Chapter 6



TAK

TAQ

TAW

TAZ

TBC

TBJ

FIGURE C.1: The time course of Tax expression as the proportion of CD4$^+$ lymphocytes that were Tax$^{\text{high}}$ or Tax$^{\text{low}}$. The supplementary data from Figure 6.1.

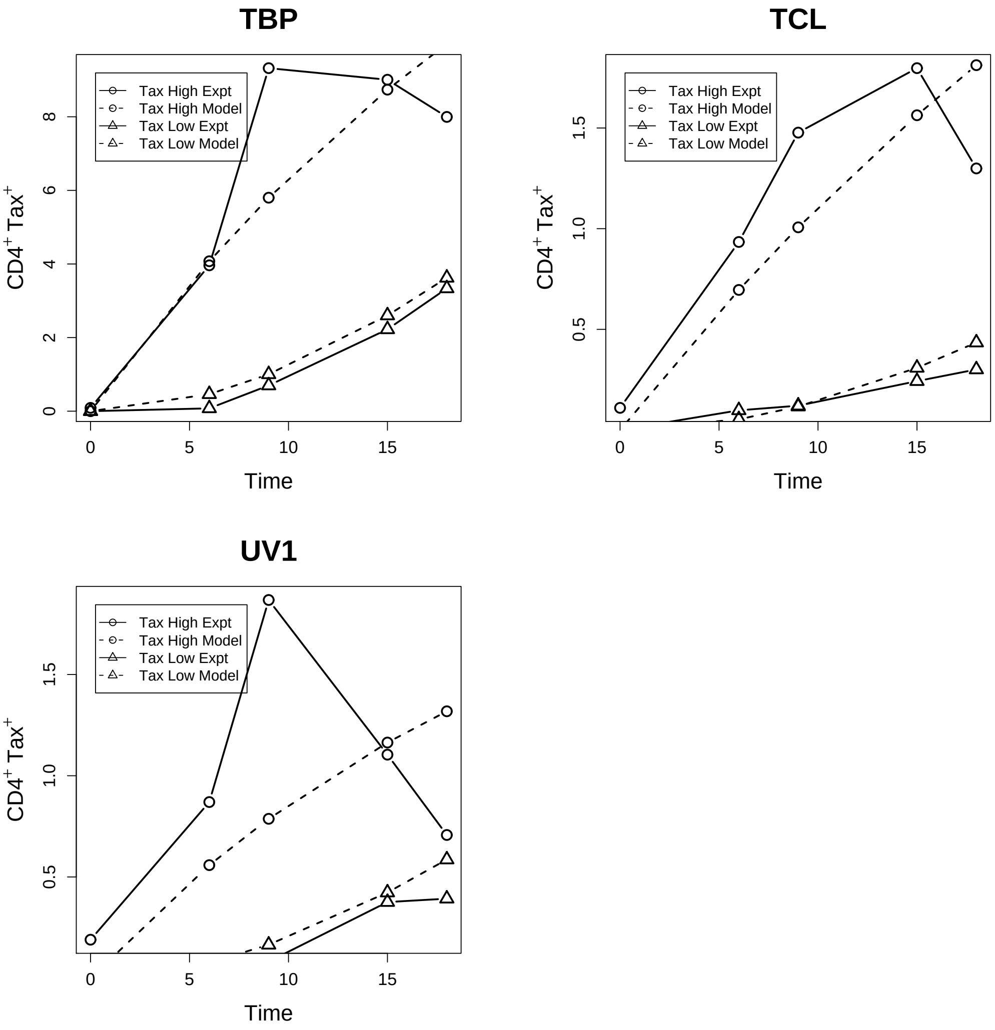


FIGURE C.1: Continued

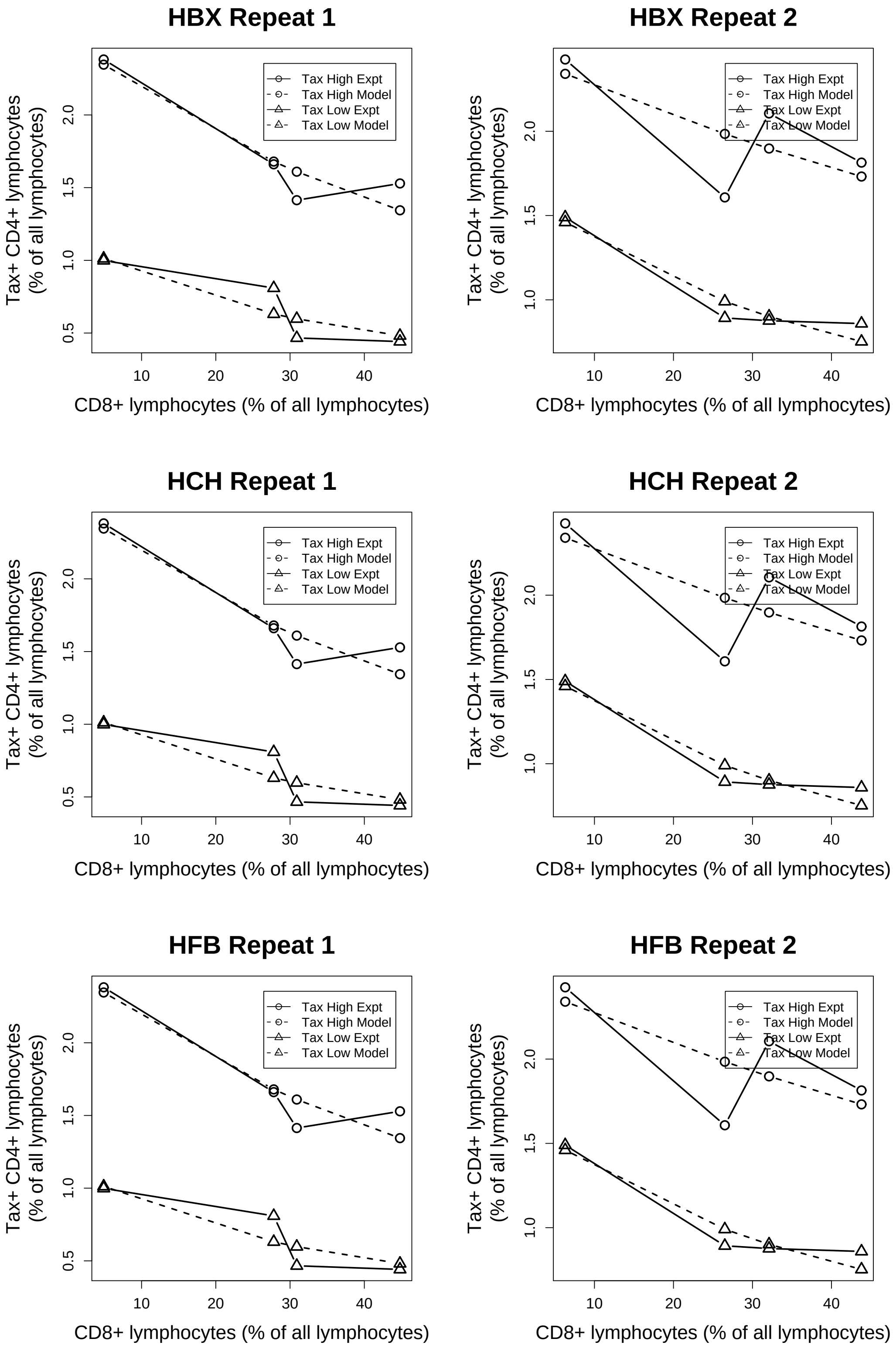


FIGURE C.2: The proportion of CD4$^+$ lymphocytes that were Tax$^{\text{high}}$ and Tax$^{\text{low}}$ following 18 h co-culture with different proportions of CD8$^+$ lymphocytes. The supplementary data from Figure 6.2.

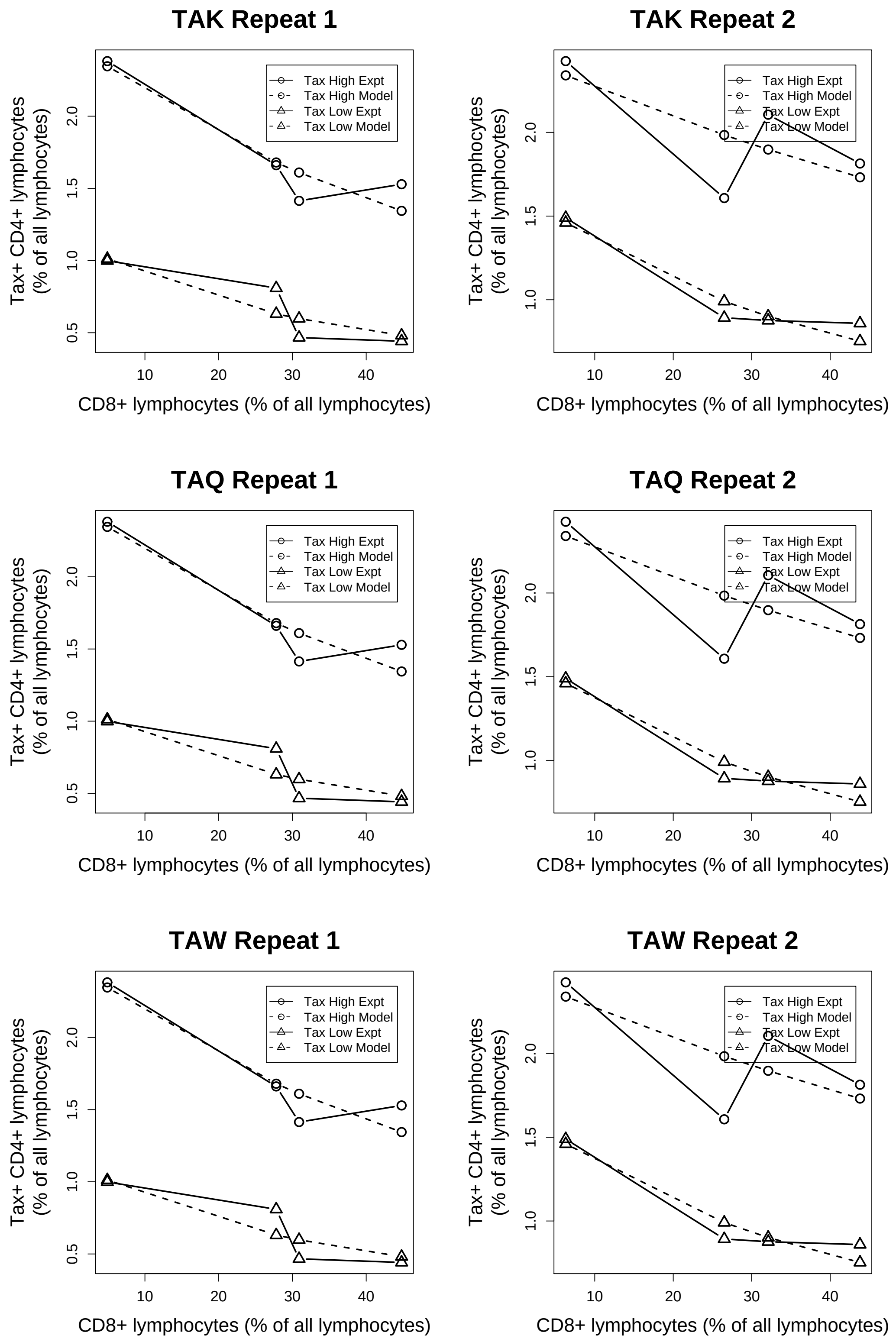


Figure C.2: Continued

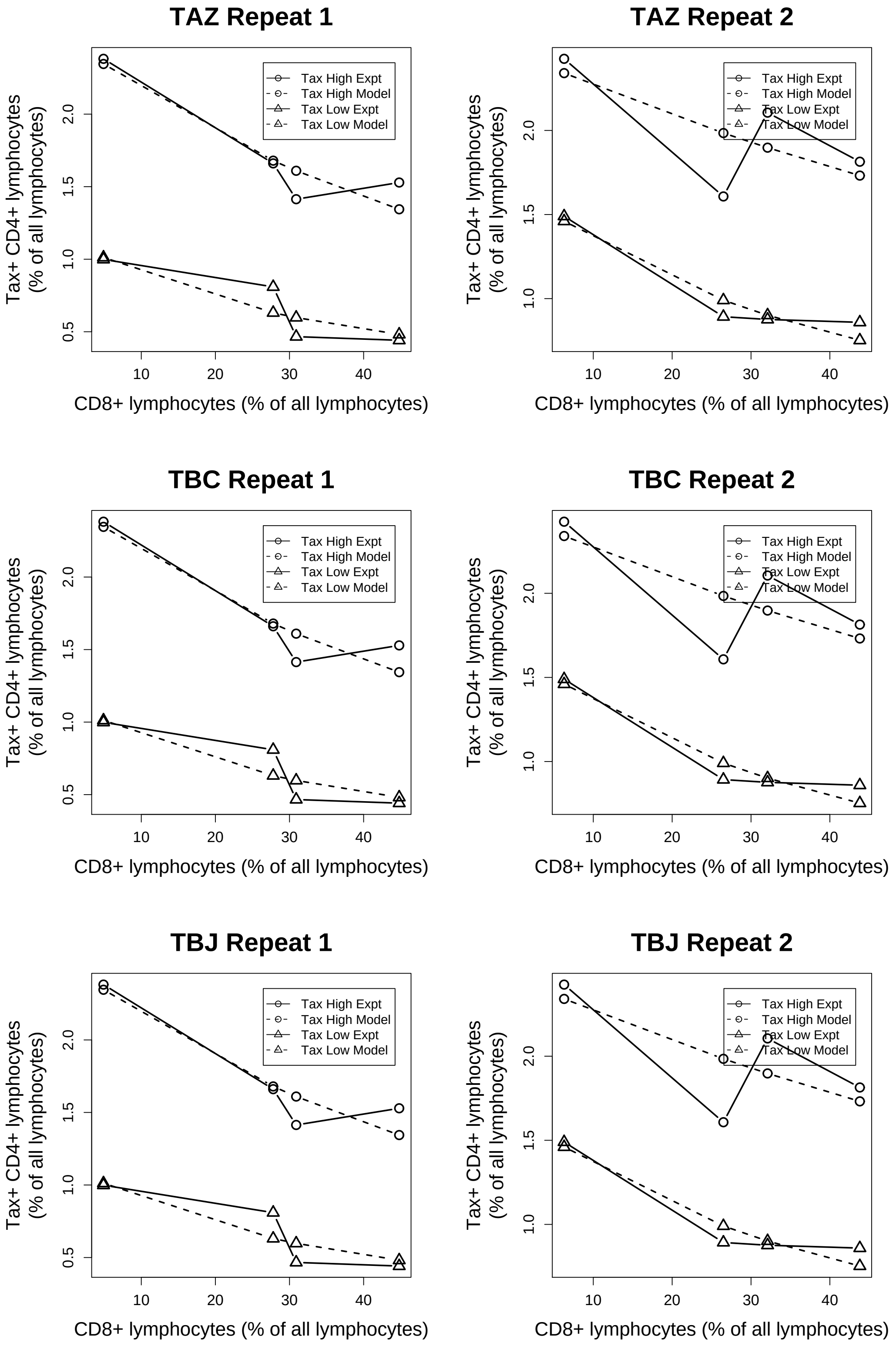
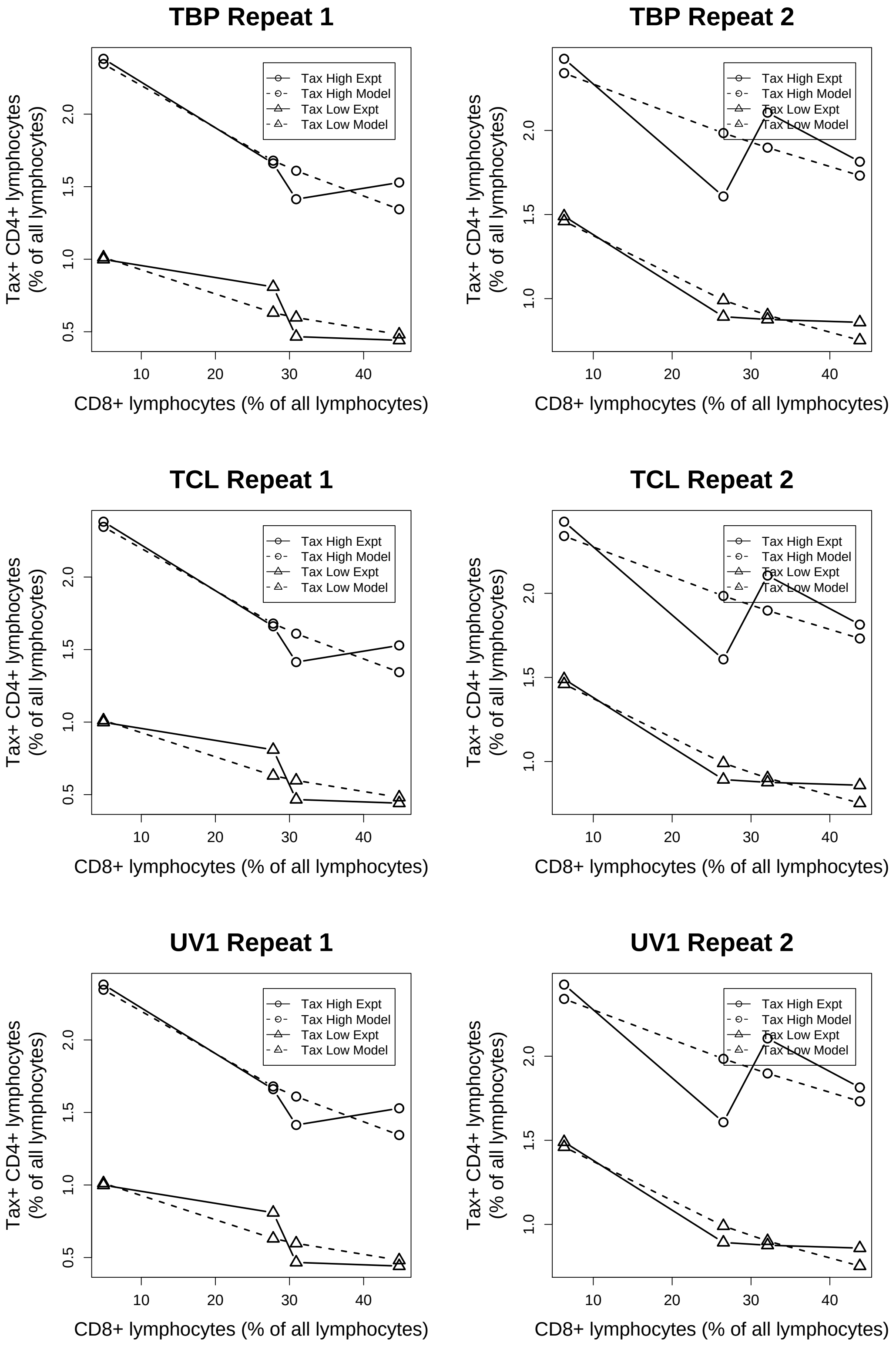


FIGURE C.2: Continued



TBP Repeat 1

TBP Repeat 2

TCL Repeat 1

TCL Repeat 2

UV1 Repeat 1

UV1 Repeat 2

Figure C.2: Continued

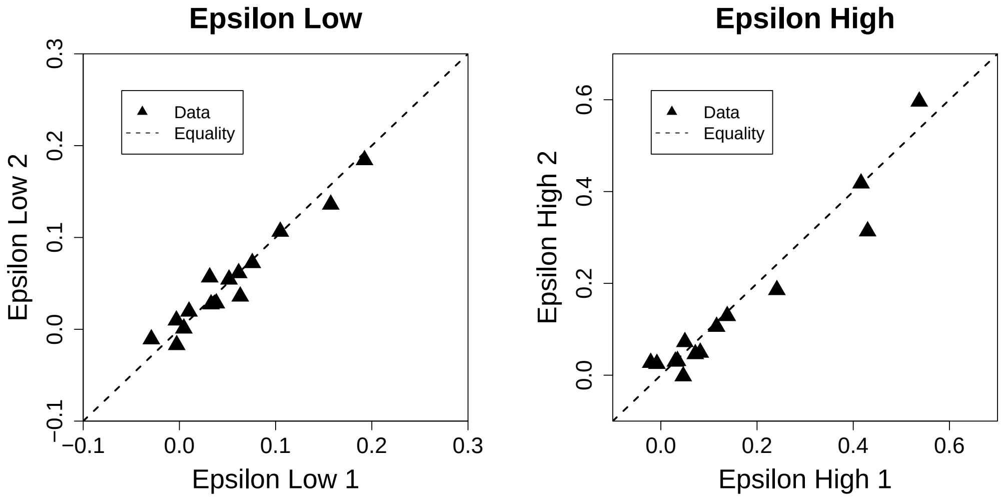


FIGURE C.3: A comparison of repeat measures of ϵ^{low} and ϵ^{high}. Both showed high agreement across repeats (ϵ^{low}: $R^2 = 0.9492$, $P < 0.001$. ϵ^{high}: $R^2 = 0.890$, $P < 0.001$).

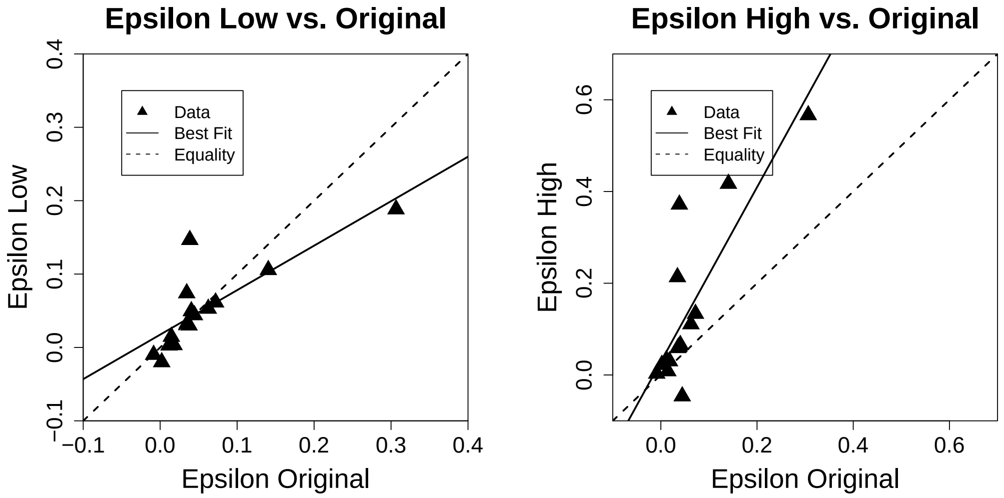


FIGURE C.4: A comparison of parameter ϵ with ϵ^{low} and ϵ^{high}. The killing rate ϵ of the original lysis model Equation 6.1 compared to ϵ^{high} ($R^2 = 0.686$, $P < 0.001$) and ϵ^{low} ($R^2 = 0.658$, $P < 0.001$).

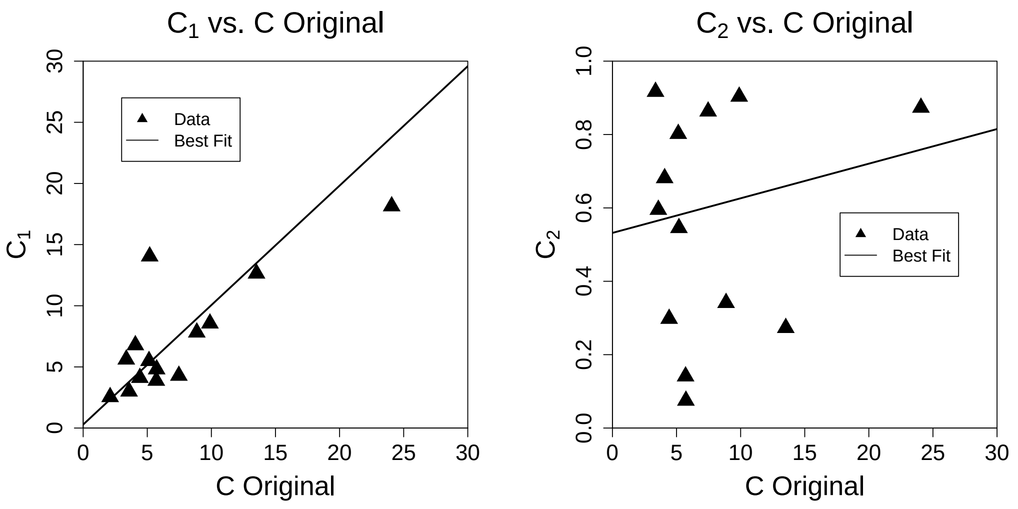


FIGURE C.5: A comparison of parameter c with c_1 and c_2. The parameter c of the original lysis model Equation 6.1 compared to the rate of increase of Tax$^{\text{low}}$ c_1 ($R^2 = 0.934$, $P < 0.001$) and Tax$^{\text{high}}$ c_2 ($R^2 = 0.129$, $P = 0.189$).

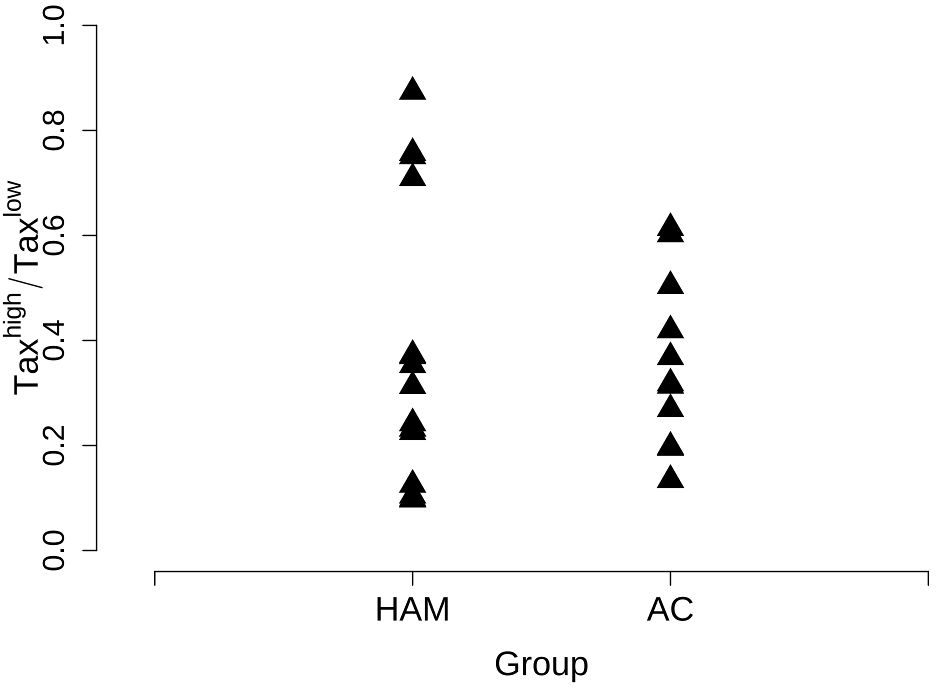


FIGURE C.6: The ratio of Tax expression for HAM/TSP and AC patients. We found no difference in the Tax$^{\text{high}}$/Tax$^{\text{low}}$ ratio between 8 HAM/TSP patients and 6 AC patients ($P = 0.871$, Wilcoxon-Mann-Whitney, HAM/TSP: $n = 16$, AC: $n = 12$). Both repeats for each patient were used.

Associated Publications

These are the papers associated with this thesis that have either been published or submitted.

Published:

1. Aidan Macnamara, Ulrich Kadolsky, Charles R M Bangham, and Becca Asquith. T-Cell Epitope Prediction: Rescaling Can Mask Biological Variation between MHC Molecules[1]. PLoS Comput Biol, 5(3):e1000327, Mar 2009.
2. Tarek Kattan[2], Aidan MacNamara[2], Aileen G Rowan[2], Hirohisa Nose, Angelina J Mosley, Yuetsu Tanaka, Graham P Taylor, Becca Asquith, and Charles R M Bangham. The avidity and lytic efficiency of the CTL response to HTLV-1. J Immunol, 182(9):5723-5729, May 2009.

Submitted:

1. Aidan MacNamara, Aileen Rowan, Silva Hilburn, Ulrich Kadolsky, Hiroshi Fujiwara, Koichiro Suemori, Masaki Yasukawa, Graham Taylor, Charles R M Bangham, and Becca Asquith. HLA Class I Binding of HBZ determines outcome in HTLV-I infection. PLoS Pathogens, *under review.*

[1] A pubcast of this manuscript is available at `http://www.scivee.tv/node/10738`.

[2] Joint first authors.

Bibliography

[1] K. J. Jeffery, K. Usuku, S. E. Hall, W. Matsumoto, G. P. Taylor, J. Procter, M. Bunce, G. S. Ogg, K. I. Welsh, J. N. Weber, A. L. Lloyd, M. A. Nowak, M. Nagai, D. Kodama, S. Izumo, M. Osame, and C. R. Bangham. HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load and the risk of HTLV-I-associated myelopathy. *Proc Natl Acad Sci U S A*, 96(7):3848–3853, Mar 1999.

[2] Nobubelo G Ngandu, Helba Bredell, Clive M Gray, Carolyn Williamson, Cathal Seoighe, , and The Hivnet028 Study Team. CTL Response to HIV Type 1 Subtype C Is Poorly Predicted by Known Epitope Motifs. *AIDS Res Hum Retroviruses*, 23 (8):1033–1041, Aug 2007. doi: 10.1089/aid.2007.0024. URL http://dx.doi.org/10.1089/aid.2007.0024.

[3] Charles R M Bangham and Mitsuhiro Osame. Cellular immune response to HTLV-1. *Oncogene*, 24(39):6035–6046, Sep 2005. doi: 10.1038/sj.onc.1208970. URL http://dx.doi.org/10.1038/sj.onc.1208970.

[4] K. J. Jeffery, A. A. Siddiqui, M. Bunce, A. L. Lloyd, A. M. Vine, A. D. Witkover, S. Izumo, K. Usuku, K. I. Welsh, M. Osame, and C. R. Bangham. The influence of HLA class I alleles and heterozygosity on the outcome of human T cell lymphotropic virus type I infection. *J Immunol*, 165(12):7278–7284, Dec 2000.

[5] B. J. Poiesz, F. W. Ruscetti, A. F. Gazdar, P. A. Bunn, J. D. Minna, and R. C. Gallo. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc Natl Acad Sci U S A*, 77(12):7415–7419, Dec 1980.

[6] G. de Th and R. Bomford. An HTLV-I vaccine: why, how, for whom? *AIDS Res Hum Retroviruses*, 9(5):381–386, May 1993.

[7] S. Jacobson, H. Shida, D. E. McFarlin, A. S. Fauci, and S. Koenig. Circulating CD8+ cytotoxic T lymphocytes specific for HTLV-I pX in patients with HTLV-I

associated neurological disease. *Nature*, 348(6298):245–248, Nov 1990. doi: 10.1038/348245a0. URL http://dx.doi.org/10.1038/348245a0.

[8] M. Kannagi, S. Harada, I. Maruyama, H. Inoko, H. Igarashi, G. Kuwashima, S. Sato, M. Morita, M. Kidokoro, and M. Sugimoto. Predominant recognition of human T cell leukemia virus type I (HTLV-I) pX gene products by human CD8+ cytotoxic T cells directed against HTLV-I-infected cells. *Int Immunol*, 3 (8):761–767, Aug 1991.

[9] C. E. Parker, S. Daenke, S. Nightingale, and C. R. Bangham. Activated, HTLV-1-specific cytotoxic T-lymphocytes are found in healthy seropositives as well as in patients with tropical spastic paraparesis. *Virology*, 188(2):628–636, Jun 1992.

[10] C. E. Parker, S. Nightingale, G. P. Taylor, J. Weber, and C. R. Bangham. Circulating anti-Tax cytotoxic T lymphocytes from human T-cell leukemia virus type I-infected people, with and without tropical spastic paraparesis, recognize multiple epitopes simultaneously. *J Virol*, 68(5):2860–2868, May 1994.

[11] Peter K C Goon, Alix Biancardi, Noam Fast, Tadahiko Igakura, Emmanuel Hanon, Angelina J Mosley, Becca Asquith, Keith G Gould, Sara Marshall, Graham P Taylor, and Charles R M Bangham. Human T cell lymphotropic virus (HTLV) type-1-specific CD8+ T cells: frequency and immunodominance hierarchy. *J Infect Dis*, 189(12):2294–2298, Jun 2004. doi: 10.1086/420832. URL http://dx.doi.org/10.1086/420832.

[12] P. Heger, O. Rosorius, J. Hauber, and R. H. Stauber. Titration of cellular export factors, but not heteromultimerization, is the molecular mechanism of trans-dominant HTLV-1 rex mutants. *Oncogene*, 18(28):4080–4090, Jul 1999. doi: 10.1038/sj.onc.1202762. URL http://dx.doi.org/10.1038/sj.onc.1202762.

[13] Christophe Nicot, Robert L Harrod, Vincenzo Ciminale, and Genoveffa Franchini. Human T-cell leukemia/lymphoma virus type 1 nonstructural genes and their functions. *Oncogene*, 24(39):6026–6034, Sep 2005. doi: 10.1038/sj.onc.1208977. URL http://dx.doi.org/10.1038/sj.onc.1208977.

[14] Luc Willems. The 14th International Conference on Human Retrovirology: HTLV and related retroviruses (July 1-4, 2009; Salvador, Brazil). *Retrovirology*, 6: 77, 2009. doi: 10.1186/1742-4690-6-77. URL http://dx.doi.org/10.1186/1742-4690-6-77.

[15] Kristien Verdonck, Elsa Gonzlez, Sonia Van Dooren, Anne-Mieke Vandamme, Guido Vanham, and Eduardo Gotuzzo. Human T-lymphotropic virus 1: recent knowledge about an ancient infection. *Lancet Infect Dis*, 7(4):266–281, Apr

2007. doi: 10.1016/S1473-3099(07)70081-6. URL http://dx.doi.org/10.1016/S1473-3099(07)70081-6.

[16] K. Novak. Ancient HTLV-1. *Nat Med*, 5(12):1357, Dec 1999. doi: 10.1038/70923. URL http://dx.doi.org/10.1038/70923.

[17] A. Gessain, J. Pecon-Slattery, L. Meertens, and R. Mahieux. Origins of HTLV-1 in South America. *Nat Med*, 6(3):232; author reply 233, Mar 2000. doi: 10.1038/73020. URL http://dx.doi.org/10.1038/73020.

[18] A. M. Vandamme, W. W. Hall, M. J. Lewis, P. Goubau, and M. Salemi. Origins of HTLV-1 in South America. *Nat Med*, 6(3):232–233, Mar 2000. doi: 10.1038/73023. URL http://dx.doi.org/10.1038/73023.

[19] H. C. Li, T. Fujiyoshi, H. Lou, S. Yashiki, S. Sonoda, L. Cartier, L. Nunez, I. Munoz, S. Horai, and K. Tajima. The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. *Nat Med*, 5(12):1428–1432, Dec 1999. doi: 10.1038/71006. URL http://dx.doi.org/10.1038/71006.

[20] S. Sonoda, H. C. Li, L. Cartier, L. Nunez, and K. Tajima. Ancient HTLV type 1 provirus DNA of Andean mummy. *AIDS Res Hum Retroviruses*, 16(16):1753–1756, Nov 2000. doi: 10.1089/08892220050193263. URL http://dx.doi.org/10.1089/08892220050193263.

[21] A. Gessain, E. Boeri, R. Yanagihara, R. C. Gallo, and G. Franchini. Complete nucleotide sequence of a highly divergent human T-cell leukemia (lymphotropic) virus type I (HTLV-I) variant from melanesia: genetic and phylogenetic relationship to HTLV-I strains from other geographical regions. *J Virol*, 67(2):1015–1023, Feb 1993.

[22] A. Gessain, R. C. Gallo, and G. Franchini. Low degree of human T-cell leukemia/-lymphoma virus type I genetic drift in vivo as a means of monitoring viral transmission and movement of ancient human populations. *J Virol*, 66(4):2288–2295, Apr 1992.

[23] N. K. Saksena, M. P. Sherman, R. Yanagihara, D. K. Dube, and B. J. Poiesz. LTR sequence and phylogenetic analyses of a newly discovered variant of HTLV-I isolated from the Hagahai of Papua New Guinea. *Virology*, 189(1):1–9, Jul 1992.

[24] M. Gasmi, B. Farouqi, M. d'Incan, and C. Desgranges. Long terminal repeat sequence analysis of HTLV type I molecular variants identified in four north African patients. *AIDS Res Hum Retroviruses*, 10(10):1313–1315, Oct 1994.

[25] T. Miura, T. Fukunaga, T. Igarashi, M. Yamashita, E. Ido, S. Funahashi, T. Ishida, K. Washio, S. Ueda, and K. Hashimoto. Phylogenetic subtypes of human T-lymphotropic virus type I and their relations to the anthropological background. *Proc Natl Acad Sci U S A*, 91(3):1124–1127, Feb 1994.

[26] A. U. Vidal, A. Gessain, M. Yoshida, R. Mahieux, K. Nishioka, F. Tekaia, L. Rosen, and G. De Th. Molecular epidemiology of HTLV type I in Japan: evidence for two distinct ancestral lineages with a particular geographical distribution. *AIDS Res Hum Retroviruses*, 10(11):1557–1566, Nov 1994.

[27] Y. Furukawa, M. Yamashita, K. Usuku, S. Izumo, M. Nakagawa, and M. Osame. Phylogenetic subgroups of human T cell lymphotropic virus (HTLV) type I in the tax gene and their association with different risks for HTLV-I-associated myelopathy/tropical spastic paraparesis. *J Infect Dis*, 182(5):1343–1349, Nov 2000.

[28] E. Hanon, R. E. Asquith, G. P. Taylor, Y. Tanaka, J. N. Weber, and C. R. Bangham. High frequency of viral protein expression in human T cell lymphotropic virus type 1-infected peripheral blood mononuclear cells. *AIDS Res Hum Retroviruses*, 16(16):1711–1715, Nov 2000. doi: 10.1089/08892220050193191. URL http://dx.doi.org/10.1089/08892220050193191.

[29] J. H. Richardson, A. J. Edwards, J. K. Cruickshank, P. Rudge, and A. G. Dalgleish. In vivo cellular tropism of human T-cell leukemia virus type 1. *J Virol*, 64(11): 5682–5687, Nov 1990.

[30] M. Nagai, Y. Yamano, M. B. Brennan, C. A. Mora, and S. Jacobson. Increased HTLV-I proviral load and preferential expansion of HTLV-I Tax-specific CD8+ T cells in cerebrospinal fluid from patients with HAM/TSP. *Ann Neurol*, 50(6): 807–812, Dec 2001.

[31] E. Hanon, J. C. Stinchcombe, M. Saito, B. E. Asquith, G. P. Taylor, Y. Tanaka, J. N. Weber, G. M. Griffiths, and C. R. Bangham. Fratricide among CD8(+) T lymphocytes naturally infected with human T cell lymphotropic virus type I. *Immunity*, 13(5):657–664, Nov 2000.

[32] Charles R M Bangham. The immune control and cell-to-cell spread of human T-lymphotropic virus type 1. *J Gen Virol*, 84(Pt 12):3177–3189, Dec 2003.

[33] H. Hoshino, M. Shimoyama, M. Miwa, and T. Sugimura. Detection of lymphocytes producing a human retrovirus associated with adult T-cell leukemia by syncytia induction assay. *Proc Natl Acad Sci U S A*, 80(23):7337–7341, Dec 1983.

[34] K. Nagy, P. Clapham, R. Cheingsong-Popov, and R. A. Weiss. Human T-cell leukemia virus type I: induction of syncytia and inhibition by patients' sera. *Int J Cancer*, 32(3):321–328, Sep 1983.

[35] M. A. Sommerfelt, B. P. Williams, P. R. Clapham, E. Solomon, P. N. Goodfellow, and R. A. Weiss. Human T cell leukemia viruses use a receptor determined by human chromosome 17. *Science*, 242(4885):1557–1559, Dec 1988.

[36] Fernando A Proietti, Anna Brbara F Carneiro-Proietti, Bernadette C Catalan-Soares, and Edward L Murphy. Global epidemiology of HTLV-I infection and associated diseases. *Oncogene*, 24(39):6058–6068, Sep 2005. doi: 10.1038/sj.onc.1208968. URL http://dx.doi.org/10.1038/sj.onc.1208968.

[37] R. F. Edlich, J. A. Arnette, and F. M. Williams. Global epidemic of human T-cell lymphotropic virus type-I (HTLV-I). *J Emerg Med*, 18(1):109–119, Jan 2000.

[38] Y. Hinuma, H. Komoda, T. Chosa, T. Kondo, M. Kohakura, T. Takenaka, M. Kikuchi, M. Ichimaru, K. Yunoki, I. Sato, R. Matsuo, Y. Takiuchi, H. Uchino, and M. Hanaoka. Antibodies to adult T-cell leukemia-virus-associated antigen (ATLA) in sera from patients with ATL and controls in Japan: a nation-wide sero-epidemiologic study. *Int J Cancer*, 29(6):631–635, Jun 1982.

[39] S. Hino, K. Yamaguchi, S. Katamine, H. Sugiyama, T. Amagasaki, K. Kinoshita, Y. Yoshida, H. Doi, Y. Tsuji, and T. Miyamoto. Mother-to-child transmission of human T-cell leukemia virus type-I. *Jpn J Cancer Res*, 76(6):474–480, Jun 1985.

[40] S. Hino, H. Sugiyama, H. Doi, T. Ishimaru, T. Yamabe, Y. Tsuji, and T. Miyamoto. Breaking the cycle of HTLV-I transmission via carrier mothers' milk. *Lancet*, 2(8551):158–159, Jul 1987.

[41] A. Komuro, M. Hayami, H. Fujii, S. Miyahara, and M. Hirayama. Vertical transmission of adult T-cell leukaemia virus. *Lancet*, 1(8318):240, Jan 1983.

[42] S. O. Stuver, N. Tachibana, A. Okayama, S. Shioiri, Y. Tsunetoshi, K. Tsuda, and N. E. Mueller. Heterosexual transmission of human T cell leukemia/lymphoma virus type I among married couples in southwestern Japan: an initial report from the Miyazaki Cohort Study. *J Infect Dis*, 167(1):57–65, Jan 1993.

[43] E. L. Murphy, J. P. Figueroa, W. N. Gibbs, A. Brathwaite, M. Holding-Cobham, D. Waters, B. Cranston, B. Hanchard, and W. A. Blattner. Sexual transmission of human T-lymphotropic virus type I (HTLV-I). *Ann Intern Med*, 111(7):555–560, Oct 1989.

[44] J. E. Kaplan, R. F. Khabbaz, E. L. Murphy, S. Hermansen, C. Roberts, R. Lal, W. Heneine, D. Wright, L. Matijas, R. Thomson, D. Rudolph, W. M. Switzer, S. Kleinman, M. Busch, and G. B. Schreiber. Male-to-female transmission of human T-cell lymphotropic virus types I and II: association with viral load. The Retrovirus Epidemiology Donor Study Group. *J Acquir Immune Defic Syndr Hum Retrovirol*, 12(2):193–201, Jun 1996.

[45] K. Okochi, H. Sato, and Y. Hinuma. A retrospective study on transmission of adult T cell leukemia virus by blood transfusion: seroconversion in recipients. *Vox Sang*, 46(5):245–253, 1984.

[46] A. Manns, R. J. Wilks, E. L. Murphy, G. Haynes, J. P. Figueroa, M. Barnett, B. Hanchard, and W. A. Blattner. A prospective study of transmission by transfusion of HTLV-I and risk factors associated with seroconversion. *Int J Cancer*, 51(6):886–891, Jul 1992.

[47] H. Strachan. On a form of multiple neuritis prevalent in the West Indies. *Practitioner*, 59:477, 1888.

[48] A. Gessain, F. Barin, J. C. Vernant, O. Gout, L. Maurs, A. Calender, and G. de Th. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. *Lancet*, 2(8452):407–410, Aug 1985.

[49] M. Osame, K. Usuku, S. Izumo, N. Ijichi, H. Amitani, A. Igata, M. Matsumoto, and M. Tara. HTLV-I associated myelopathy, a new clinical entity. *Lancet*, 1 (8488):1031–1032, May 1986.

[50] J. E. Kaplan, M. Osame, H. Kubota, A. Igata, H. Nishitani, Y. Maeda, R. F. Khabbaz, and R. S. Janssen. The risk of development of HTLV-I-associated myelopathy/tropical spastic paraparesis among persons infected with HTLV-I. *J Acquir Immune Defic Syndr*, 3(11):1096–1101, 1990.

[51] G. P. Taylor. *Principles and Practice of Clinical Virology*, chapter Human T-cell Lymphotropic Viruses, pages 695–710. Chichester, John Wiley and sons, Ltd., 2000.

[52] M. M. Aye, E. Matsuoka, T. Moritoyo, F. Umehara, M. Suehara, Y. Hokezu, H. Yamanaka, Y. Isashiki, M. Osame, and S. Izumo. Histopathological analysis of four autopsy cases of HTLV-I-associated myelopathy/tropical spastic paraparesis: inflammatory changes occur simultaneously in the entire central nervous system. *Acta Neuropathol*, 100(3):245–252, Sep 2000.

[53] Stphane Olindo, Agns Lzin, Philippe Cabre, Harold Merle, Martine Saint-Vil, Mireille Edimonana Kaptue, Assatou Signate, Raymond Csaire, and Didier

Smadja. HTLV-1 proviral load in peripheral blood mononuclear cells quantified in 100 HAM/TSP patients: a marker of disease progression. *J Neurol Sci*, 237(1-2):53–59, Oct 2005. doi: 10.1016/j.jns.2005.05.010. URL http://dx.doi.org/10.1016/j.jns.2005.05.010.

[54] J. A. Sakai, M. Nagai, M. B. Brennan, C. A. Mora, and S. Jacobson. In vitro spontaneous lymphoproliferation in patients with human T-cell lymphotropic virus type I-associated neurologic disease: predominant expansion of CD8+ T cells. *Blood*, 98(5):1506–1511, Sep 2001.

[55] Peter K C Goon, Tadahiko Igakura, Emmanuel Hanon, Angelina J Mosley, Becca Asquith, Keith G Gould, Graham P Taylor, Jonathan N Weber, and Charles R M Bangham. High circulating frequencies of tumor necrosis factor alpha- and interleukin-2-secreting human T-lymphotropic virus type 1 (HTLV-1)-specific CD4+ T cells in patients with HTLV-1-associated neurological disease. *J Virol*, 77(17):9716–9722, Sep 2003.

[56] P. A. Montanheiro, P. A. Montanheito, A. C Penalva de Oliveira, M. P. Posada-Vergara, A. C. Milagres, C. Tauil, P. E. Marchiori, A. J S Duarte, and J. Casseb. Human T-cell lymphotropic virus type I (HTLV-I) proviral DNA viral load among asymptomatic patients and patients with HTLV-I-associated myelopathy/tropical spastic paraparesis. *Braz J Med Biol Res*, 38(11):1643–1647, Nov 2005. doi: /S0100-879X2005001100011. URL http://dx.doi.org//S0100-879X2005001100011.

[57] Alison M Vine, Aviva D Witkover, Alun L Lloyd, Katie J M Jeffery, Asna Siddiqui, Sara E F Marshall, Mike Bunce, Nobutaka Eiraku, Shuji Izumo, Koichiro Usuku, Mitsuhiro Osame, and Charles R M Bangham. Polygenic control of human T lymphotropic virus type I (HTLV-I) provirus load and the risk of HTLV-I-associated myelopathy/tropical spastic paraparesis. *J Infect Dis*, 186(7):932–939, Oct 2002.

[58] Ryuji Kubota, Samantha S Soldan, Roland Martin, and Steven Jacobson. Selected cytotoxic T lymphocytes with high specificity for HTLV-I in cerebrospinal fluid from a HAM/TSP patient. *J Neurovirol*, 8(1):53–57, Feb 2002.

[59] Paolo A Muraro, Klaus-Peter Wandinger, Bibiana Bielekova, Bruno Gran, Adriana Marques, Ursula Utz, Henry F McFarland, Steve Jacobson, and Roland Martin. Molecular tracking of antigen-specific T cell clones in neurological immune-mediated disorders. *Brain*, 126(Pt 1):20–31, Jan 2003.

[60] Michael C Levin, Sang Min Lee, Franck Kalume, Yvette Morcos, F. Curtis Dohan, Karen A Hasty, Joseph C Callaway, Joseph Zunt, Dominic Desiderio, and John M

Stuart. Autoimmunity due to molecular mimicry as a cause of neurological disease. *Nat Med*, 8(5):509–513, May 2002. doi: 10.1038/nm0502-509. URL http://dx.doi.org/10.1038/nm0502-509.

[61] F. Garca-Vallejo, M. C. Domnguez, and O. Tamayo. Autoimmunity and molecular mimicry in tropical spastic paraparesis/human T-lymphotropic virus-associated myelopathy. *Braz J Med Biol Res*, 38(2):241–250, Feb 2005. doi: /S0100-879X2005000200013. URL http://dx.doi.org//S0100-879X2005000200013.

[62] G. C. Romn and L. N. Romn. Tropical spastic paraparesis. A clinical study of 50 patients from Tumaco (Colombia) and review of the worldwide features of the syndrome. *J Neurol Sci*, 87(1):121–138, Oct 1988.

[63] M. Nakagawa, K. Nakahara, Y. Maruyama, M. Kawabata, I. Higuchi, H. Kubota, S. Izumo, K. Arimura, and M. Osame. Therapeutic trials in 200 patients with HTLV-I-associated myelopathy/ tropical spastic paraparesis. *J Neurovirol*, 2(5): 345–355, Oct 1996.

[64] M. Osame, A. Igata, M. Matsumoto, M. Kohka, K. Usuku, and S. Izumo. HTLV-I-associated myelopathy (HAM): treatment trials, retrospective survey and clinical and laboratory findings. *Hematology Reviews*, 3:271–284, 1990.

[65] Graham P Taylor, Peter Goon, Yoshitaka Furukawa, Hannah Green, Anna Barfield, Angelina Mosley, Hirohisa Nose, Abdel Babiker, Peter Rudge, Koichiro Usuku, Mitsuhiro Osame, Charles R M Bangham, and Jonathan N Weber. Zidovudine plus lamivudine in Human T-Lymphotropic Virus type-I-associated myelopathy: a randomised trial. *Retrovirology*, 3:63, 2006. doi: 10.1186/1742-4690-3-63. URL http://dx.doi.org/10.1186/1742-4690-3-63.

[66] M. Yoshida. Multiple viral strategies of HTLV-1 for dysregulation of cell growth control. *Annu Rev Immunol*, 19:475–496, 2001. doi: 10.1146/annurev.immunol.19.1.475. URL http://dx.doi.org/10.1146/annurev.immunol.19.1.475.

[67] Graham P Taylor and Masao Matsuoka. Natural history of adult T-cell leukemia/-lymphoma and approaches to therapy. *Oncogene*, 24(39):6047–6057, Sep 2005. doi: 10.1038/sj.onc.1208979. URL http://dx.doi.org/10.1038/sj.onc.1208979.

[68] M. Shimoyama. Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. A report from the Lymphoma Study Group (1984-87). *Br J Haematol*, 79(3):428–437, Nov 1991.

[69] M. Roudier, I. Lamaury, and M. Strobel. Human T cell leukemia/lymphoma virus type I (HTLV-I) and Pneumocystis carinii associated with T cell proliferation and haemophagocytic syndrome. *Leukemia*, 11(3):453–454, Mar 1997.

[70] Kunihiro Tsukasaki, Olivier Hermine, Ali Bazarbachi, Lee Ratner, Juan Carlos Ramos, William Harrington, Deirdre O'Mahony, John E Janik, Achila L Bittencourt, Graham P Taylor, Kazunari Yamaguchi, Atae Utsunomiya, Kensei Tobinai, and Toshiki Watanabe. Definition, prognostic factors, treatment, and response criteria of adult T-cell leukemia-lymphoma: a proposal from an international consensus meeting. *J Clin Oncol*, 27(3):453–459, Jan 2009. doi: 10.1200/JCO.2008.18.2428. URL http://dx.doi.org/10.1200/JCO.2008.18.2428.

[71] K. Eguchi, T. Origuchi, H. Takashima, K. Iwata, S. Katamine, and S. Nagataki. High seroprevalence of anti-HTLV-I antibody in rheumatoid arthritis. *Arthritis Rheum*, 39(3):463–466, Mar 1996.

[72] Edward L Murphy, Baoguang Wang, Ronald A Sacher, Joy Fridey, James W Smith, Catharie C Nass, Bruce Newman, Helen E Ownby, George Garratty, Shelia T Hutching, and George B Schreiber. Respiratory and urinary tract infections, arthritis, and asthma associated with HTLV-I and HTLV-II infection. *Emerg Infect Dis*, 10(1):109–116, Jan 2004.

[73] Y. Iwakura, M. Tosu, E. Yoshida, M. Takiguchi, K. Sato, I. Kitajima, K. Nishioka, K. Yamamoto, T. Takeda, and M. Hatanaka. Induction of inflammatory arthropathy resembling rheumatoid arthritis in mice transgenic for HTLV-I. *Science*, 253 (5023):1026–1028, Aug 1991.

[74] Maria Yakova, Agns Lzin, Fabienne Dantin, Gisle Lagathu, Stphane Olindo, Georges Jean-Baptiste, Serge Arfi, and Raymond Csaire. Increased proviral load in HTLV-1-infected patients with rheumatoid arthritis or connective tissue disease. *Retrovirology*, 2:4, 2005. doi: 10.1186/1742-4690-2-4. URL http://dx.doi.org/10.1186/1742-4690-2-4.

[75] H. Aono, K. Fujisawa, T. Hasunuma, S. J. Marriott, and K. Nishioka. Extracellular human T cell leukemia virus type I tax protein stimulates the proliferation of human synovial cells. *Arthritis Rheum*, 41(11):1995–2003, Nov 1998. doi: 3.0.CO;2-4. URL http://dx.doi.org/3.0.CO;2-4.

[76] M. Mochizuki, T. Watanabe, K. Yamaguchi, K. Tajima, K. Yoshimura, S. Nakashima, M. Shirao, S. Araki, N. Miyata, and S. Mori. Uveitis associated with human T lymphotropic virus type I: seroepidemiologic, clinical, and virologic studies. *J Infect Dis*, 166(4):943–944, Oct 1992.

[77] K. Nakao, N. Ohba, M. Nakagawa, and M. Osame. Clinical course of HTLV-I-associated uveitis. *Jpn J Ophthalmol*, 43(5):404–409, 1999.

[78] E. M. Carvalho and A. Da Fonseca Porto. Epidemiological and clinical interaction between HTLV-1 and Strongyloides stercoralis. *Parasite Immunol*, 26(11-12):487–497, 2004. doi: 10.1111/j.0141-9838.2004.00726.x. URL http://dx.doi.org/10.1111/j.0141-9838.2004.00726.x.

[79] Luis A Marcos, Angelica Terashima, Herbert L Dupont, and Eduardo Gotuzzo. Strongyloides hyperinfection syndrome: an emerging global infectious disease. *Trans R Soc Trop Med Hyg*, 102(4):314–318, Apr 2008. doi: 10.1016/j.trstmh.2008.01.020. URL http://dx.doi.org/10.1016/j.trstmh.2008.01.020.

[80] S. Niewiesk, S. Daenke, C. E. Parker, G. Taylor, J. Weber, S. Nightingale, and C. R. Bangham. The transactivator gene of human T-cell leukemia virus type I is more variable within and between healthy carriers than patients with tropical spastic paraparesis. *J Virol*, 68(10):6778–6781, Oct 1994.

[81] S. Niewiesk, S. Daenke, C. E. Parker, G. Taylor, J. Weber, S. Nightingale, and C. R. Bangham. Naturally occurring variants of human T-cell leukemia virus type I Tax protein impair its recognition by cytotoxic T lymphocytes and the transactivation function of Tax. *J Virol*, 69(4):2649–2653, Apr 1995.

[82] S. Niewiesk and C. R. Bangham. Evolution in a chronic RNA virus infection: selection on HTLV-I tax protein differs between healthy carriers and patients with tropical spastic paraparesis. *J Mol Evol*, 42(4):452–458, Apr 1996.

[83] Ryuji Kubota, Kousuke Hanada, Yoshitaka Furukawa, Kimiyoshi Arimura, Mitsuhiro Osame, Takashi Gojobori, and Shuji Izumo. Genetic stability of human T lymphotropic virus type I despite antiviral pressures by CTLs. *J Immunol*, 178(9):5966–5972, May 2007.

[84] Yoshitaka Furukawa, Koichiro Usuku, Shuji Izumo, and Mitsuhiro Osame. Human T cell lymphotropic virus type I (HTLV-I) p12I is dispensable for HTLV-I transmission and maintenance of infection in vivo. *AIDS Res Hum Retroviruses*, 20(10):1092–1099, Oct 2004. doi: 10.1089/aid.2004.20.1092. URL http://dx.doi.org/10.1089/aid.2004.20.1092.

[85] R. E. Smith, S. Niewiesk, S. Booth, C. R. Bangham, and S. Daenke. Functional conservation of HTLV-1 rex balances the immune pressure for sequence variation in the rex gene. *Virology*, 237(2):397–403, Oct 1997. doi: 10.1006/viro.1997.8789. URL http://dx.doi.org/10.1006/viro.1997.8789.

[86] R. Kubota, M. Nagai, T. Kawanishi, M. Osame, and S. Jacobson. Increased HTLV type 1 tax specific CD8+ cells in HTLV type 1-asociated myelopathy/tropical spastic paraparesis: correlation with HTLV type 1 proviral load. *AIDS Res Hum Retroviruses*, 16(16):1705–1709, Nov 2000. doi: 10.1089/08892220050193182. URL http://dx.doi.org/10.1089/08892220050193182.

[87] Becca Asquith, Angelina J Mosley, Anna Barfield, Sara E F Marshall, Adrian Heaps, Peter Goon, Emmanuel Hanon, Yuetsu Tanaka, Graham P Taylor, and Charles R M Bangham. A functional CD8+ cell assay reveals individual variation in CD8+ cell antiviral efficacy and explains differences in human T-lymphotropic virus type 1 proviral load. *J Gen Virol*, 86(Pt 5):1515–1523, May 2005. doi: 10.1099/vir.0.80766-0. URL http://dx.doi.org/10.1099/vir.0.80766-0.

[88] Becca Asquith, Yan Zhang, Angelina J Mosley, Catherine M de Lara, Diana L Wallace, Andrew Worth, Lambrini Kaftantzi, Kiran Meekings, George E Griffin, Yuetsu Tanaka, David F Tough, Peter C Beverley, Graham P Taylor, Derek C Macallan, and Charles R M Bangham. In vivo T lymphocyte dynamics in humans and the impact of human T-lymphotropic virus 1 infection. *Proc Natl Acad Sci U S A*, 104(19):8035–8040, May 2007. doi: 10.1073/pnas.0608832104. URL http://dx.doi.org/10.1073/pnas.0608832104.

[89] Tarek Kattan, Aidan MacNamara, Aileen G Rowan, Hirohisa Nose, Angelina J Mosley, Yuetsu Tanaka, Graham P Taylor, Becca Asquith, and Charles R M Bangham. The avidity and lytic efficiency of the CTL response to HTLV-1. *J Immunol*, 182(9):5723–5729, May 2009. doi: 10.4049/jimmunol.0900069. URL http://dx.doi.org/10.4049/jimmunol.0900069.

[90] Amir H Sabouri, Koichiro Usuku, Daisuke Hayashi, Shuji Izumo, Yoshiro Ohara, Mitsuhiro Osame, and Mineki Saito. Impaired function of human T-lymphotropic virus type 1 (HTLV-1)-specific CD8+ T cells in HTLV-1-associated neurologic disease. *Blood*, 112(6):2411–2420, Sep 2008. doi: 10.1182/blood-2008-02-140335. URL http://dx.doi.org/10.1182/blood-2008-02-140335.

[91] N. E. Mueller and W. A. Blattner. *Viral Infections of Humans: Epidemiology and Control*, chapter Retroviruses: HTLV, pages 785–813. New York, Plenum Medical Press, 1997.

[92] Anne O'Garra and Paulo Vieira. Regulatory T cells and mechanisms of immune system control. *Nat Med*, 10(8):801–805, Aug 2004. doi: 10.1038/nm0804-801. URL http://dx.doi.org/10.1038/nm0804-801.

[93] Yoshihisa Yamano, Norihiro Takenouchi, Hong-Chuan Li, Utano Tomaru, Karen Yao, Christian W Grant, Dragan A Maric, and Steven Jacobson. Virus-induced

dysfunction of CD4+CD25+ T cells in patients with HTLV-I-associated neuroimmunological disease. *J Clin Invest*, 115(5):1361–1368, May 2005. doi: 10.1172/JCI200523913. URL http://dx.doi.org/10.1172/JCI200523913.

[94] Unsong Oh, Christian Grant, Caitlin Griffith, Kazunori Fugo, Norihiro Takenouchi, and Steven Jacobson. Reduced Foxp3 protein expression is associated with inflammatory disease during human t lymphotropic virus type 1 Infection. *J Infect Dis*, 193(11):1557–1566, Jun 2006. doi: 10.1086/503874. URL http://dx.doi.org/10.1086/503874.

[95] Robert S Fujinami. A tax on luxury: HTLV-I infection of CD4+CD25+ Tregs. *J Clin Invest*, 115(5):1144–1146, May 2005. doi: 10.1172/JCI200525130. URL http://dx.doi.org/10.1172/JCI200525130.

[96] Hiroki Yano, Takashi Ishida, Atsushi Inagaki, Toshihiko Ishii, Shigeru Kusumoto, Hirokazu Komatsu, Shinsuke Iida, Atae Utsunomiya, and Ryuzo Ueda. Regulatory T-cell function of adult T-cell leukemia/lymphoma cells. *Int J Cancer*, 120(9): 2052–2057, May 2007. doi: 10.1002/ijc.22536. URL http://dx.doi.org/10.1002/ijc.22536.

[97] Mineki Saito, Veronique M Braud, Peter Goon, Emmanuel Hanon, Graham P Taylor, Akiko Saito, Nobutaka Eiraku, Yuetsu Tanaka, Koichiro Usuku, Jonathan N Weber, Mitsuhiro Osame, and Charles R M Bangham. Low frequency of CD94/NKG2A+ T lymphocytes in patients with HTLV-1-associated myelopathy/tropical spastic paraparesis, but not in asymptomatic carriers. *Blood*, 102 (2):577–584, Jul 2003. doi: 10.1182/blood-2002-09-2855. URL http://dx.doi.org/10.1182/blood-2002-09-2855.

[98] F. Yu, Y. Itoyama, K. Fujihara, and I. Goto. Natural killer (NK) cells in HTLV-I-associated myelopathy/tropical spastic paraparesis-decrease in NK cell subset populations and activity in HTLV-I seropositive individuals. *J Neuroimmunol*, 33 (2):121–128, Aug 1991.

[99] K. Fujihara, Y. Itoyama, F. Yu, C. Kubo, and I. Goto. Cellular immune surveillance against HTLV-I infected T lymphocytes in HTLV-I associated myelopathy/tropical spastic paraparesis (HAM/TSP). *J Neurol Sci*, 105(1):99–107, Sep 1991.

[100] Alison M Vine, Adrian G Heaps, Lambrini Kaftantzi, Angelina Mosley, Becca Asquith, Aviva Witkover, Gillian Thompson, Mineki Saito, Peter K C Goon, Laura Carr, Francisco Martinez-Murillo, Graham P Taylor, and Charles R M Bangham. The role of CTLs in persistent viral infection: cytolytic gene expression in CD8+

lymphocytes distinguishes between individuals with a high or low proviral load of human T cell lymphotropic virus type 1. *J Immunol*, 173(8):5121–5129, Oct 2004.

[101] Peter K C Goon, Emmanuel Hanon, Tadahiko Igakura, Yuetsu Tanaka, Jonathan N Weber, Graham P Taylor, and Charles R M Bangham. High frequencies of Th1-type CD4(+) T cells specific to HTLV-1 Env and Tax proteins in patients with HTLV-1-associated myelopathy/tropical spastic paraparesis. *Blood*, 99(9):3335–3341, May 2002.

[102] J. W. Yewdell and J. R. Bennink. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol*, 17:51–88, 1999. doi: 10.1146/annurev.immunol.17.1.51. URL http://dx.doi.org/10.1146/annurev.immunol.17.1.51.

[103] Mette Voldby Larsen, Claus Lundegaard, Kasper Lamberth, Sren Buus, Sren Brunak, Ole Lund, and Morten Nielsen. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*, 35(8):2295–2303, Aug 2005. doi: 10.1002/eji.200425811. URL http://dx.doi.org/10.1002/eji.200425811.

[104] Huynh-Hoa Bui, John Sidney, Kenny Dinh, Scott Southwood, Mark J Newman, and Alessandro Sette. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 7:153, 2006. doi: 10.1186/1471-2105-7-153. URL http://dx.doi.org/10.1186/1471-2105-7-153.

[105] Alessandro Sette, Ward Fleri, Bjoern Peters, Muthuraman Sathiamurthy, Huynh-Hoa Bui, and Stephen Wilson. A roadmap for the immunomics of category A-C pathogens. *Immunity*, 22(2):155–161, Feb 2005.

[106] Alessandro Sette and Bjoern Peters. Immune epitope mapping in the post-genomic era: lessons for vaccine development. *Curr Opin Immunol*, 19(1):106–110, Feb 2007. doi: 10.1016/j.coi.2006.11.002. URL http://dx.doi.org/10.1016/j.coi.2006.11.002.

[107] Muthuraman Sathiamurthy, Bjoern Peters, Huynh-Hoa Bui, John Sidney, John Mokili, Stephen S Wilson, Ward Fleri, Deborah L McGuinness, Philip E Bourne, and Alessandro Sette. An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. *Immunome Res*, 1(1):2, Sep 2005. doi: 10.1186/1745-7580-1-2. URL http://dx.doi.org/10.1186/1745-7580-1-2.

[108] James T Snyder, Igor M Belyakov, Amiran Dzutsev, Franois Lemonnier, and Jay A Berzofsky. Protection against lethal vaccinia virus challenge in HLA-A2 transgenic mice by immunization with a single CD8+ T-cell peptide epitope of vaccinia and variola viruses. *J Virol*, 78(13):7052–7060, Jul 2004. doi: 10.1128/JVI.78.13.7052-7060.2004. URL http://dx.doi.org/10.1128/JVI.78.13.7052-7060.2004.

[109] David C Tscharke, Gunasegaran Karupiah, Jie Zhou, Tara Palmore, Kari R Irvine, S. M Mansour Haeryfar, Shanicka Williams, John Sidney, Alessandro Sette, Jack R Bennink, and Jonathan W Yewdell. Identification of poxvirus CD8+ T cell determinants to enable rational design and characterization of smallpox vaccines. *J Exp Med*, 201(1):95–104, Jan 2005. doi: 10.1084/jem.20041912. URL http://dx.doi.org/10.1084/jem.20041912.

[110] Ingo Drexler, Caroline Staib, Wolfgang Kastenmuller, Stefan Stevanovi?, Burkhard Schmidt, Franois A Lemonnier, Hans-Georg Rammensee, Dirk H Busch, Helga Bernhard, Volker Erfle, and Gerd Sutter. Identification of vaccinia virus epitope-specific HLA-A*0201-restricted T cells and comparative analysis of smallpox vaccines. *Proc Natl Acad Sci U S A*, 100(1):217–222, Jan 2003. doi: 10.1073/pnas.262668999. URL http://dx.doi.org/10.1073/pnas.262668999.

[111] Carla Oseroff, Ferdynand Kos, Huynh-Hoa Bui, Bjoern Peters, Valerie Pasquetto, Jean Glenn, Tara Palmore, John Sidney, David C Tscharke, Jack R Bennink, Scott Southwood, Howard M Grey, Jonathan W Yewdell, and Alessandro Sette. HLA class I-restricted responses to vaccinia recognize a broad array of proteins mainly involved in virulence and viral gene regulation. *Proc Natl Acad Sci U S A*, 102(39):13980–13985, Sep 2005. doi: 10.1073/pnas.0506768102. URL http://dx.doi.org/10.1073/pnas.0506768102.

[112] Valerie Pasquetto, Huynh-Hoa Bui, Rielle Giannino, Cindy Banh, Fareed Mirza, John Sidney, Carla Oseroff, David C Tscharke, Kari Irvine, Jack R Bennink, Bjoern Peters, Scott Southwood, Vincenzo Cerundolo, Howard Grey, Jonathan W Yewdell, and Alessandro Sette. HLA-A*0201, HLA-A*1101, and HLA-B*0702 transgenic mice recognize numerous poxvirus determinants from a wide variety of viral gene products. *J Immunol*, 175(8):5504–5515, Oct 2005.

[113] Mingjun Wang, Britta Johansen, Mogens H Nissen, Mette Thorn, Henrik Klverpris, Anders Fomsgaard, Sren Buus, and Mogens H Classon. Identification of an HLA-A*0201 restricted Bcl2-derived epitope expressed on tumors. *Cancer Lett*, 251(1):86–95, Jun 2007. doi: 10.1016/j.canlet.2006.11.004. URL http://dx.doi.org/10.1016/j.canlet.2006.11.004.

[114] Mette Thorn, Mingjun Wang, Henrik Klverpris, Esben G W Schmidt, Anders Fomsgaard, Lynn Wenandy, Annika Berntsen, Sren Brunak, Sren Buus, and Mogens H Claesson. Identification of a new hTERT-derived HLA-A*0201 restricted, naturally processed CTL epitope. *Cancer Immunol Immunother*, 56(11):1755–1763, Nov 2007. doi: 10.1007/s00262-007-0319-y. URL http://dx.doi.org/10.1007/s00262-007-0319-y.

[115] Mingjun Wang, Kasper Lamberth, Mikkel Harndahl, Gustav Rder, Anette Stryhn, Mette V Larsen, Morten Nielsen, Claus Lundegaard, Sheila T Tang, Morten H Dziegiel, Jrgen Rosenkvist, Anders E Pedersen, Sren Buus, Mogens H Claesson, and Ole Lund. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. *Vaccine*, 25(15):2823–2831, Apr 2007. doi: 10.1016/j.vaccine.2006.12.038. URL http://dx.doi.org/10.1016/j.vaccine.2006.12.038.

[116] Zabrina L Brumme, Chanson J Brumme, David Heckerman, Bette T Korber, Marcus Daniels, Jonathan Carlson, Carl Kadie, Tanmoy Bhattacharya, Celia Chui, James Szinger, Theresa Mo, Robert S Hogg, Julio S G Montaner, Nicole Frahm, Christian Brander, Bruce D Walker, and P. Richard Harrigan. Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. *PLoS Pathog*, 3(7):e94, Jul 2007. doi: 10.1371/journal.ppat.0030094. URL http://dx.doi.org/10.1371/journal.ppat.0030094.

[117] Jos A M Borghans, Anne Mlgaard, Rob J de Boer, and Can Ke?mir. HLA Alleles Associated with Slow Progression to AIDS Truly Prefer to Present HIV-1 p24. *PLoS ONE*, 2(9):e920, 2007. doi: 10.1371/journal.pone.0000920. URL http://dx.doi.org/10.1371/journal.pone.0000920.

[118] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovi? SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50 (3-4):213–219, Nov 1999.

[119] Morten Nielsen, Claus Lundegaard, Peder Worning, Christina Sylvester Hvid, Kasper Lamberth, Sren Buus, Sren Brunak, and Ole Lund. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, 20(9):1388–1397, Jun 2004. doi: 10.1093/bioinformatics/bth100. URL http://dx.doi.org/10.1093/bioinformatics/bth100.

[120] Huynh-Hoa Bui, John Sidney, Bjoern Peters, Muthuraman Sathiamurthy, Asabe Sinichi, Kelly-Anne Purton, Bianca R Moth, Francis V Chisari, David I Watkins, and Alessandro Sette. Automated generation and evaluation of specific MHC

binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5):304–314, Jun 2005. doi: 10.1007/s00251-005-0798-y. URL http://dx.doi.org/10.1007/s00251-005-0798-y.

[121] Bjoern Peters and Alessandro Sette. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6:132, 2005. doi: 10.1186/1471-2105-6-132. URL http://dx.doi.org/10.1186/1471-2105-6-132.

[122] David Heckerman, Carl Kadie, and Jennifer Listgarten. Leveraging information across HLA alleles/supertypes improves epitope prediction. *J Comput Biol*, 14 (6):736–746, 2007. doi: 10.1089/cmb.2007.R013. URL http://dx.doi.org/10.1089/cmb.2007.R013.

[123] S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P-M. Kloetzel, H-G. Rammensee, H. Schild, and H-G. Holzhtter. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci*, 62(9):1025–1037, May 2005. doi: 10.1007/s00018-005-4528-2. URL http://dx.doi.org/10.1007/s00018-005-4528-2.

[124] S. Buus, S. L. Lauemller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, and S. Brunak. Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens*, 62(5):378–384, Nov 2003.

[125] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemller, Kasper Lamberth, Sren Buus, Sren Brunak, and Ole Lund. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, 12(5):1007–1017, May 2003.

[126] B. Peters, H-H. Bui, J. Sidney, Z. Weng, J. T. Loffredo, D. I. Watkins, B. R. Moth, and A. Sette. A computational resource for the prediction of peptide binding to Indian rhesus macaque MHC class I molecules. *Vaccine*, 23(45):5212–5224, Nov 2005. doi: 10.1016/j.vaccine.2005.07.086. URL http://dx.doi.org/10.1016/j.vaccine.2005.07.086.

[127] M. Nakagawa, S. Izumo, S. Ijichi, H. Kubota, K. Arimura, M. Kawabata, and M. Osame. HTLV-I-associated myelopathy: analysis of 213 patients based on clinical features and laboratory findings. *J Neurovirol*, 1(1):50–61, Mar 1995.

[128] URL http://www.hiv.lanl.gov/content/immunology.

[129] T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M. P. Protti, F. Sinigaglia, and J. Hammer. Generation

of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*, 17(6):555–561, Jun 1999. doi: 10.1038/9858. URL http://dx.doi.org/10.1038/9858.

[130] Andrew J T George, Jaroslav Stark, and Cliburn Chan. Understanding specificity and sensitivity of T-cell recognition. *Trends Immunol*, 26(12):653–659, Dec 2005. doi: 10.1016/j.it.2005.09.011. URL http://dx.doi.org/10.1016/j.it.2005.09.011.

[131] Tim Elliott. The 'chop-and-change' of MHC class I assembly. *Nat Immunol*, 7 (1):7–9, Jan 2006. doi: 10.1038/ni0106-7. URL http://dx.doi.org/10.1038/ni0106-7.

[132] Anthony P Williams, Chen Au Peh, Anthony W Purcell, James McCluskey, and Tim Elliott. Optimization of the MHC class I peptide cargo is dependent on tapasin. *Immunity*, 16(4):509–520, Apr 2002.

[133] A. Sette and J. Sidney. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, 50(3-4):201–212, Nov 1999.

[134] Bette T. M. Korber, Christian Brander, Barton F. Haynes, Richard Koup, John P. Moore, Bruce D. Walker, and David I. Watkins, editors. *HIV Molecular Immunology 2005*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico, 2005.

[135] Bjoern Peters, Huynh-Hoa Bui, Sune Frankild, Morten Nielson, Claus Lundegaard, Emrah Kostem, Derek Basch, Kasper Lamberth, Mikkel Harndahl, Ward Fleri, Stephen S Wilson, John Sidney, Ole Lund, Soren Buus, and Alessandro Sette. A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol*, 2(6):e65, Jun 2006. doi: 10.1371/journal.pcbi.0020065. URL http://dx.doi.org/10.1371/journal.pcbi.0020065.

[136] Mette Larsen, Claus Lundegaard, Kasper Lamberth, Soren Buus, Ole Lund, and Morten Nielsen. Large-Scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, 8(1):424, Oct 2007. doi: 10.1186/1471-2105-8-424. URL http://dx.doi.org/10.1186/1471-2105-8-424.

[137] UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res*, 36(Database issue):D190–D195, Jan 2008.

[138] V. H. Engelhard. Structure of peptides associated with class I and class II MHC molecules. *Annu Rev Immunol*, 12:181–207, 1994. doi: 10.1146/annurev.iy.

12.040194.001145. URL http://dx.doi.org/10.1146/annurev.iy.12.040194.001145.

[139] Nicole Frahm, Karina Yusim, Todd J Suscovich, Sharon Adams, John Sidney, Peter Hraber, Hannah S Hewitt, Caitlyn H Linde, Daniel G Kavanagh, Tonia Woodberry, Leah M Henry, Kellie Faircloth, Jennifer Listgarten, Carl Kadie, Nebojsa Jojic, Kaori Sango, Nancy V Brown, Eunice Pae, M. Tauheed Zaman, Florian Bihl, Ashok Khatri, Mina John, Simon Mallal, Francesco M Marincola, Bruce D Walker, Alessandro Sette, David Heckerman, Bette T Korber, and Christian Brander. Extensive HLA class I allele promiscuity among viral CTL epitopes. *Eur J Immunol*, 37(9):2419–2433, Sep 2007. doi: 10.1002/eji.200737365. URL http://dx.doi.org/10.1002/eji.200737365.

[140] Samantha J Westrop, Nathali Grageda, and Nesrina Imami. Novel approach to recognition of predicted HIV-1 Gag B3501-restricted CD8 T-cell epitopes by HLA-B3501(+) patients: confirmation by quantitative ELISpot analyses and characterisation using multimers. *J Immunol Methods*, 341(1-2):76–85, Feb 2009. doi: 10.1016/j.jim.2008.11.003. URL http://dx.doi.org/10.1016/j.jim.2008.11.003.

[141] T. F. Greten, J. E. Slansky, R. Kubota, S. S. Soldan, E. M. Jaffee, T. P. Leist, D. M. Pardoll, S. Jacobson, and J. P. Schneck. Direct visualization of antigen-specific T cells: HTLV-1 Tax11-19- specific CD8(+) T cells are activated in peripheral blood and accumulate in cerebrospinal fluid from HAM/TSP patients. *Proc Natl Acad Sci U S A*, 95(13):7568–7573, Jun 1998.

[142] Steven Jacobson. Immunopathogenesis of human T cell lymphotropic virus type I-associated neurologic disease. *J Infect Dis*, 186 Suppl 2:S187–S192, Dec 2002. doi: 10.1086/344269. URL http://dx.doi.org/10.1086/344269.

[143] Roshni Sundaram, Yiping Sun, Christopher M Walker, Francois A Lemonnier, Steven Jacobson, and Pravin T P Kaumaya. A novel multivalent human CTL peptide construct elicits robust cellular immune responses in HLA-A*0201 transgenic mice: implications for HTLV-1 vaccine design. *Vaccine*, 21(21-22):2767–2781, Jun 2003.

[144] S. Hausmann, W. E. Biddison, K. J. Smith, Y. H. Ding, D. N. Garboczi, U. Utz, D. C. Wiley, and K. W. Wucherpfennig. Peptide recognition by two HLA-A2/Tax11-19-specific T cell clones in relationship to their MHC/peptide/TCR crystal structures. *J Immunol*, 162(9):5389–5397, May 1999.

[145] Carel A van Baalen, Christophe Guillon, Minus van Baalen, Esther J Verschuren, Patrick H M Boers, Albert D M E Osterhaus, and Rob A Gruters. Impact of

antigen expression kinetics on the effectiveness of HIV-specific cytotoxic T lymphocytes. *Eur J Immunol*, 32(9):2644–2652, Sep 2002. doi: 3.0.CO;2-R. URL http://dx.doi.org/3.0.CO;2-R.

[146] Daniel L Barber, E. John Wherry, David Masopust, Baogong Zhu, James P Allison, Arlene H Sharpe, Gordon J Freeman, and Rafi Ahmed. Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature*, 439(7077):682–687, Feb 2006. doi: 10.1038/nature04444. URL http://dx.doi.org/10.1038/nature04444.

[147] Mathias Lichterfeld, Xu G Yu, Stanley K Mui, Katie L Williams, Alicja Trocha, Mark A Brockman, Rachel L Allgaier, Michael T Waring, Tomohiko Koibuchi, Mary N Johnston, Daniel Cohen, Todd M Allen, Eric S Rosenberg, Bruce D Walker, and Marcus Altfeld. Selective depletion of high-avidity human immunodeficiency virus type 1 (HIV-1)-specific CD8+ T cells after early HIV-1 infection. *J Virol*, 81(8):4199–4214, Apr 2007. doi: 10.1128/JVI.01388-06. URL http://dx.doi.org/10.1128/JVI.01388-06.

[148] Hendrik Streeck, Zabrina L Brumme, Michael Anastario, Kristin W Cohen, Jonathan S Jolin, Angela Meier, Chanson J Brumme, Eric S Rosenberg, Galit Alter, Todd M Allen, Bruce D Walker, and Marcus Altfeld. Antigen load and viral sequence diversification determine the functional profile of HIV-1-specific CD8+ T cells. *PLoS Med*, 5(5):e100, May 2008. doi: 10.1371/journal.pmed.0050100. URL http://dx.doi.org/10.1371/journal.pmed.0050100.

[149] Charles R M Bangham. CTL quality and the control of human retroviral infections. *Eur J Immunol*, 39(7):1700–1712, Jul 2009. doi: 10.1002/eji.200939451. URL http://dx.doi.org/10.1002/eji.200939451.

[150] Aidan Macnamara, Ulrich Kadolsky, Charles R M Bangham, and Becca Asquith. T-Cell Epitope Prediction: Rescaling Can Mask Biological Variation between MHC Molecules. *PLoS Comput Biol*, 5(3):e1000327, Mar 2009. doi: 10.1371/journal.pcbi.1000327. URL http://dx.doi.org/10.1371/journal.pcbi.1000327.

[151] Boris Schmid, Can Ke?mir, and Rob J de Boer. The specificity and polymorphism of the MHC class I prevents the global adaptation of HIV-1 to the monomorphic proteasome and TAP. *PLoS ONE*, 3(10):e3525, 2008. doi: 10.1371/journal.pone.0003525. URL http://dx.doi.org/10.1371/journal.pone.0003525.

[152] Marie-Hlne Fortier, Etienne Caron, Marie-Pierre Hardy, Grgory Voisin, Sbastien Lemieux, Claude Perreault, and Pierre Thibault. The MHC class I peptide repertoire is molded by the transcriptome. *J Exp Med*, 205(3):595–610, Mar 2008. doi: 10.1084/jem.20071985. URL http://dx.doi.org/10.1084/jem.20071985.

[153] Charles R. M. Bangham. *Genetic susceptibility to infectious diseases*, chapter Human T-Lymphotropic Virus Type 1 (HTLV-1)Associated Diseases, pages 303–317. New York, Oxford University Press., 2008.

[154] M. Seiki, S. Hattori, Y. Hirayama, and M. Yoshida. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc Natl Acad Sci U S A*, 80(12):3618–3622, Jun 1983.

[155] Yorifumi Satou, Jun ichirou Yasunaga, Mika Yoshida, and Masao Matsuoka. HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. *Proc Natl Acad Sci U S A*, 103(3):720–725, Jan 2006. doi: 10.1073/pnas.0507631103. URL http://dx.doi.org/10.1073/pnas.0507631103.

[156] C. Pique, F. Connan, J. P. Levilain, J. Choppin, and M. C. Dokhlar. Among all human T-cell leukemia virus type 1 proteins, tax, polymerase, and envelope proteins are predicted as preferential targets for the HLA-A2-restricted cytotoxic T-cell response. *J Virol*, 70(8):4919–4926, Aug 1996.

[157] M. Nagai, K. Usuku, W. Matsumoto, D. Kodama, N. Takenouchi, T. Moritoyo, S. Hashiguchi, M. Ichinose, C. R. Bangham, S. Izumo, and M. Osame. Analysis of HTLV-I proviral load in 202 HAM/TSP patients and 243 asymptomatic HTLV-I carriers: high proviral load strongly predisposes to HAM/TSP. *J Neurovirol*, 4(6):586–593, Dec 1998.

[158] Masao Matsuoka and Patrick Green. The HBZ gene, a key player in HTLV-1 pathogenesis. *Retrovirology*, 6(1):71, Aug 2009. doi: 10.1186/1742-4690-6-71. URL http://dx.doi.org/10.1186/1742-4690-6-71.

[159] Gilles Gaudray, Frederic Gachon, Jihane Basbous, Martine Biard-Piechaczyk, Christian Devaux, and Jean-Michel Mesnard. The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. *J Virol*, 76(24):12813–12822, Dec 2002.

[160] Marie-Hlne Cavanagh, Sbastien Landry, Brigitte Audet, Charlotte Arpin-Andr, Patrick Hivin, Marie-Eve Par, Julien Thte, Eric Wattel, Susan J Marriott, Jean-Michel Mesnard, and Benoit Barbeau. HTLV-I antisense transcripts initiating in the 3'LTR are alternatively spliced and polyadenylated. *Retrovirology*, 3:

15, 2006. doi: 10.1186/1742-4690-3-15. URL http://dx.doi.org/10.1186/1742-4690-3-15.

[161] Jihane Basbous, Charlotte Arpin, Gilles Gaudray, Marc Piechaczyk, Christian Devaux, and Jean-Michel Mesnard. The HBZ factor of human T-cell leukemia virus type I dimerizes with transcription factors JunB and c-Jun and modulates their transcriptional activity. *J Biol Chem*, 278(44):43620–43627, Oct 2003. doi: 10.1074/jbc.M307275200. URL http://dx.doi.org/10.1074/jbc.M307275200.

[162] Min Li, Matthew Kesic, Han Yin, Lianbo Yu, and Patrick L Green. Kinetic Analysis of Human T-cell Leukemia Virus Type 1 Gene Expression in Cell Culture and Infected Animals. *J Virol*, Feb 2009. doi: 10.1128/JVI.02315-08. URL http://dx.doi.org/10.1128/JVI.02315-08.

[163] E. Hanon, S. Hall, G. P. Taylor, M. Saito, R. Davis, Y. Tanaka, K. Usuku, M. Osame, J. N. Weber, and C. R. Bangham. Abundant tax protein expression in CD4+ T cells infected with human T-cell lymphotropic virus type I (HTLV-I) is prevented by cytotoxic T lymphocytes. *Blood*, 95(4):1386–1392, Feb 2000.

[164] Robert A Seder, Patricia A Darrah, and Mario Roederer. T-cell quality in memory and protection: implications for vaccine design. *Nat Rev Immunol*, 8(4):247–258, Apr 2008. doi: 10.1038/nri2274. URL http://dx.doi.org/10.1038/nri2274.

[165] John Stambas, Peter C Doherty, and Stephen J Turner. An in vivo cytotoxicity threshold for influenza A virus-specific effector and memory CD8(+) T cells. *J Immunol*, 178(3):1285–1292, Feb 2007.

[166] Becca Asquith and Charles R M Bangham. Quantifying HTLV-I dynamics. *Immunol Cell Biol*, 85(4):280–286, Jun 2007. doi: 10.1038/sj.icb.7100050. URL http://dx.doi.org/10.1038/sj.icb.7100050.

[167] B. Lee, Y. Tanaka, and H. Tozawa. Monoclonal antibody defining tax protein of human T-cell leukemia virus type-I. *Tohoku J Exp Med*, 157(1):1–11, Jan 1989.

[168] Marco A Purbhoo, Darrell J Irvine, Johannes B Huppa, and Mark M Davis. T cell killing does not require the formation of a stable mature immunological synapse. *Nat Immunol*, 5(5):524–530, May 2004. doi: 10.1038/ni1058. URL http://dx.doi.org/10.1038/ni1058.

[169] Y. Sykulev, M. Joo, I. Vturina, T. J. Tsomides, and H. N. Eisen. Evidence that a single peptide-MHC complex on a target cell can elicit a cytolytic T cell response. *Immunity*, 4(6):565–571, Jun 1996.

[170] E. R. Christinck, M. A. Luscher, B. H. Barber, and D. B. Williams. Peptide binding to class I MHC on living cells and quantitation of complexes required for CTL lysis. *Nature*, 352(6330):67–70, Jul 1991. doi: 10.1038/352067a0. URL http://dx.doi.org/10.1038/352067a0.

[171] C. Pique, A. Ureta-Vidal, A. Gessain, B. Chancerel, O. Gout, R. Tamouza, F. Agis, and M. C. Dokhlar. Evidence for the chronic in vivo production of human T cell leukemia virus type I Rof and Tof proteins from cytotoxic T lymphocytes directed against viral peptides. *J Exp Med*, 191(3):567–572, Feb 2000.

[172] C. A. Biron, K. B. Nguyen, G. C. Pien, L. P. Cousens, and T. P. Salazar-Mather. Natural killer cells in antiviral defense: function and regulation by innate cytokines. *Annu Rev Immunol*, 17:189–220, 1999. doi: 10.1146/annurev.immunol.17.1.189. URL http://dx.doi.org/10.1146/annurev.immunol.17.1.189.

[173] David H Raulet. Interplay of natural killer cells and their receptors with the adaptive immune response. *Nat Immunol*, 5(10):996–1002, Oct 2004. doi: 10.1038/ni1114. URL http://dx.doi.org/10.1038/ni1114.

[174] Lewis L Lanier. NK cell recognition. *Annu Rev Immunol*, 23:225–274, 2005. doi: 10.1146/annurev.immunol.23.021704.115526. URL http://dx.doi.org/10.1146/annurev.immunol.23.021704.115526.

[175] Francesco Colucci, James P Di Santo, and Paul J Leibson. Natural killer cell activation in mice and men: different triggers for similar weapons? *Nat Immunol*, 3(9):807–813, Sep 2002. doi: 10.1038/ni0902-807. URL http://dx.doi.org/10.1038/ni0902-807.

[176] A. Harel-Bellan, A. Quillet, C. Marchiol, R. DeMars, T. Tursz, and D. Fradelizi. Natural killer susceptibility of human cells may be regulated by genes in the HLA region on chromosome 6. *Proc Natl Acad Sci U S A*, 83(15):5688–5692, Aug 1986.

[177] A. Moretta, G. Tambussi, C. Bottino, G. Tripodi, A. Merli, E. Ciccone, G. Pantaleo, and L. Moretta. A novel surface antigen expressed by a subset of human CD3- CD16+ natural killer cells. Role in cell activation and regulation of cytolytic function. *J Exp Med*, 171(3):695–714, Mar 1990.

[178] M. Uhrberg, N. M. Valiante, B. P. Shum, H. G. Shilling, K. Lienert-Weidenbach, B. Corliss, D. Tyan, L. L. Lanier, and P. Parham. Human diversity in killer cell inhibitory receptor genes. *Immunity*, 7(6):753–763, Dec 1997.

[179] N. M. Valiante, M. Uhrberg, H. G. Shilling, K. Lienert-Weidenbach, K. L. Arnett, A. D'Andrea, J. H. Phillips, L. L. Lanier, and P. Parham. Functionally and

structurally distinct NK cell receptor repertoires in the peripheral blood of two human donors. *Immunity*, 7(6):739–751, Dec 1997.

[180] Mary Carrington and Paul Norman. *The KIR Gene Cluster*. National Library of Medicine (US), NCBI, 2009. URL http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=mono_003.

[181] Salim I Khakoo and Mary Carrington. KIR and disease: a model system or system of models? *Immunol Rev*, 214:186–201, Dec 2006. doi: 10.1111/j.1600-065X.2006.00459.x. URL http://dx.doi.org/10.1111/j.1600-065X.2006.00459.x.

[182] Salim I Khakoo, Chloe L Thio, Maureen P Martin, Collin R Brooks, Xiaojiang Gao, Jacquie Astemborski, Jie Cheng, James J Goedert, David Vlahov, Margaret Hilgartner, Steven Cox, Ann-Margeret Little, Graeme J Alexander, Matthew E Cramp, Stephen J O'Brien, William M C Rosenberg, David L Thomas, and Mary Carrington. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*, 305(5685):872–874, Aug 2004. doi: 10.1126/science.1097670. URL http://dx.doi.org/10.1126/science.1097670.

[183] Maureen P Martin, Xiaojiang Gao, Jeong-Hee Lee, George W Nelson, Roger Detels, James J Goedert, Susan Buchbinder, Keith Hoots, David Vlahov, John Trowsdale, Michael Wilson, Stephen J O'Brien, and Mary Carrington. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet*, 31(4):429–434, Aug 2002. doi: 10.1038/ng934. URL http://dx.doi.org/10.1038/ng934.

[184] S. Gaudieri, D. DeSantis, E. McKinnon, C. Moore, D. Nolan, C. S. Witt, S. A. Mallal, and F. T. Christiansen. Killer immunoglobulin-like receptors and HLA act both independently and synergistically to modify HIV disease progression. *Genes Immun*, 6(8):683–690, Dec 2005. doi: 10.1038/sj.gene.6364256. URL http://dx.doi.org/10.1038/sj.gene.6364256.

[185] D. Wodarz, S. E. Hall, K. Usuku, M. Osame, G. S. Ogg, A. J. McMichael, M. A. Nowak, and C. R. Bangham. Cytotoxic T-cell abundance and virus load in human immunodeficiency virus type 1 and human T-cell leukaemia virus type 1. *Proc Biol Sci*, 268(1473):1215–1221, Jun 2001. doi: 10.1098/rspb.2001.1608. URL http://dx.doi.org/10.1098/rspb.2001.1608.

[186] M. A. Nowak and C. R. Bangham. Population dynamics of immune responses to persistent viruses. *Science*, 272(5258):74–79, Apr 1996.

[187] Frederic Toulza, Adrian Heaps, Yuetsu Tanaka, Graham P Taylor, and Charles R M Bangham. High frequency of CD4+FoxP3+ cells in HTLV-1 infection: inverse

correlation with HTLV-1-specific CTL response. *Blood*, 111(10):5047–5053, May 2008. doi: 10.1182/blood-2007-10-118539. URL http://dx.doi.org/10.1182/blood-2007-10-118539.

[188] Thomas Stranzl, Mette Voldby Larsen, Claus Lundegaard, and Morten Nielsen. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, 62(6):357–368, Jun 2010. doi: 10.1007/s00251-010-0441-4. URL http://dx.doi.org/10.1007/s00251-010-0441-4.

[189] Koichiro Suemori, Hiroshi Fujiwara, Toshiki Ochi, Taiji Ogawa, Masao Matsuoka, Tadashi Matsumoto, Jean-Michel Mesnard, and Masaki Yasukawa. HBZ is an immunogenic protein, but not a target antigen for human T-cell leukemia virus type 1-specific cytotoxic T lymphocytes. *J Gen Virol*, 90(Pt 8):1806–1811, Aug 2009. doi: 10.1099/vir.0.010199-0. URL http://dx.doi.org/10.1099/vir.0.010199-0.

[190] Tal Vider-Shalit, Vered Fishbain, Shai Raffaeli, and Yoram Louzoun. Phase-dependent immune evasion of herpesviruses. *J Virol*, 81(17):9536–9545, Sep 2007. doi: 10.1128/JVI.02636-06. URL http://dx.doi.org/10.1128/JVI.02636-06.

[191] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Rder, Bjoern Peters, Alessandro Sette, Ole Lund, and Sren Buus. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence. *PLoS ONE*, 2(8):e796, 2007. doi: 10.1371/journal.pone.0000796. URL http://dx.doi.org/10.1371/journal.pone.0000796.