Limits of viral adaptation to the antigen presentation pathway

Grenzen van virale adaptatie aan de antigeen presentatie route

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op dinsdag 8 september 2009 des middags te 4.15 uur

door

Boris Valentijn Schmid

geboren op 16 juli 1978 te Leiderdorp

Promotor: Prof. dr. R. J. De Boer
Co-promotor: Dr. C. Keşmir

The work in this thesis was financially supported by the Netherlands Organisation for Scientific Research (NWO) (VICI grant 016.048.603, Open Program 812.07.003), and a High Potential grant (2007) from the Utrecht University.

Leescommissie:

Prof. dr. Rolf F. Hoekstra
Prof. dr. Paul Klenerman
Prof. dr. Bette T. Korber
Prof. dr. Jacques J. Neefjes

The printing of this thesis was financially supported by the Utrecht University, and the Eijkman Graduate School for Infectious Diseases.

Contents

General Introduction

1

"Most young children go through a 'why' phase, questioning the world. Some of us never leave this phase."

This introduction is mainly targeted to the non-immunologist/biologist reader, and hopefully will provide them with the background information required to understand the introductions of chapters 2-5. Readers that are familiar with the human immune system and the antigen presentation pathway can start reading at section 1.4; the origin of the MHC polymorphism.

The frequent use of italicized questions in this introduction is inspired by the contributions of the late Ernst Mayrs (Mayr, 1961, 1997) to the philosophy of biological science. In this thesis I provide a possible explanation both in proximate ('how/why does it work') and ultimate ('why is it the way it is') terms for the current shape of the antigen presentation pathway.

1.1 UNDER SIEGE. HOST-PATHOGEN COEVOLUTION

Since the onset of cellular life (Holland and Domingo, 1998; Hendrix et al., 2000), and possibly even before that (Takeuchi and Hogeweg, 2008), hosts have been plagued by pathogens. While at first pathogens and host were comparable in size and evolutionary speed, the difference in generation time between the two classes increased as the hosts became larger (Gaillard et al., 2005). After millions of years of evolution, a virus like HIV-1 has a generation time of +/- 36 hours (Perelson et al., 1996; Rodrigo et al., 1999), whereas humans have a generation time of +/- 28 years (Fenner, 2005), which is nearly a 7000-fold difference. It is therefore essential for slow-evolving hosts to have some kind of generic immune system that can deal with fast-changing pathogens.

1.2 THE HUMAN IMMUNE SYSTEM

Hosts have evolved a variety of defensive mechanisms that cannot easily be escaped by pathogens. For the Gnathostomata (jawed vertebrata, which includes *Homo sapiens*, the immune system can roughly be described as a three-layered system.

The first layer of defense is the physical edge of our beings. The intact human skin is impermeable for most pathogens (Elias, 2007), and vastly reduces number of pathogens that enter the body. Surface areas that are not covered by a horn layer, but are still exposed to the outside world (e.g. the digestive tract, eyes, air cavities, the urine-duct and the vagina) are coated with a mu-

cus that contain antibiotic or acidic molecules that makes survival and entry of pathogens less likely (Huttner and Bevins, 1999).

Once pathogens penetrate this first layer, they enter the extracellular fluid (i.e. the blood plasma and tissue fluid). In this fluid, pathogens are exposed to two immune systems, namely the innate immune system and the adaptive immune system. The **innate immune system** is a fast-responding and evolutionary conserved method of dealing with pathogens, and consists mainly of phagocytes, and granulocytes[1]. These immune cells recognize conserved parts of pathogens such as the bacterial flagella (Beatson et al., 2006), or viral capsids, and respond by engulfing the pathogen, and/or locally releasing toxic molecules. Although some pathogens do carry mutated versions of these evolutionary conserved parts (Andersen-Nissen et al., 2005), most avoid the innate immune response by other means (e.g. by covering their cell wall or capsid with host-derived proteins, or by delivering inhibitory signals to the phagocytes (Finlay and McFadden, 2006)).

The **adaptive immune system** operates on the same extracellular level as the innate immune response, but deals with pathogens in a different way. The adaptive immune system consists mainly of T cells and B cells. These cells do not specifically target the evolutionary conserved signatures of pathogens, but are selected during their development to target material of foreign origin (i.e. foreign antigens). The adaptive immune system can be further subdivided in the humoral immune system and the cellular immune system.

The **B cells** of the **humoral immune system** have a receptor on their cell surface that upon binding to its cognate antigen will either trigger the cell to start producing antibodies by itself, or after activation by a T helper cell (reviewed in LeBien and Tedder (2008)). These antibodies are released into the extracellular fluid. As the antibodies bind to the pathogens, they hinder the normal functioning of a pathogen directly, but also serve as marker for cells of the innate and the adaptive immune system[2] to destroy that pathogen.

However, the effectivity of antibodies, and thus of the humoral immune system is limited to the extracellular fluid domain. Many bacteria and all viruses spend a large portion of their lifetime within host cells, where they are hidden from the humoral and innate immune system. A third layer of defense, the intracellular **Antigen presentation pathway**, and its extracellular component, the **cellular immune system**, are specialized in detecting intracellular pathogens[3].

[1] NK cells are part of the innate immune system, but are also involved in the cellular immune system, where they monitor the expression of MHC class I alleles on the cell surface. Intracellular pathogens can downregulate MHC class I expression to avoid the adaptive immune system, but cannot do so completely without drawing the attention of NK cells (Waldhauer and Steinle, 2008).

[2] Antibodies can also serve as a marker for the complement system.

[3] There are several other intracellular defense systems in place, which are unjustly left out of this summary of the immune system in exchange for brevity. Primarily amongst them are the siRNAs (Carthew and Sontheimer, 2009).

1.3 THE ANTIGEN PRESENTATION PATHWAY

1.3.1 *Purpose*

"What does the Antigen presentation Pathway do?"

The Antigen presentation pathway makes it possible for the immune system to monitor the intracellular content of cells in the body from the surface of these cells. The pathway displays small samples of all the proteins in a cell on the cell surface with specialized proteins called **Major Histocompatibility complex (MHC) molecules**[4]. Viruses need the host cell to synthesize the proteins neccesary for replication, and intracellular bacteria typically excrete proteins into the cytosol to manipulate the host cell (Pamer et al., 1997). For both types of pathogens, protein samples will inevitably appear on the cell surface.

Once these protein samples are displayed on the cell surface, infected cells can be recognized by the **$CD8^+$ T cells** of the cellular immune system. $CD8^+$ T cells learn to recognize foreign peptides during their development in the thymus. Here, immature T cells rearrange certain genes that define the binding pattern of the T cell receptor. Only those T cells with a T cell receptor that can bind with enough, but not too much affinity to MHC class I (for $CD8^+$ T cells) or class II molecules[5] (for CD4+ T cells) mature into competent naive T cells (Boehmer, 2008; Borghans et al., 2003). This process ensures that naive T cells will only recognize non-human peptide fragments with enough affinity to start an immune responsee. Humans have a large repertoire of naive T cells, and roughly 50% of the presented foreign peptides will be recognized by one or more T cells (Yewdell and Bennink, 1999). The other half could be too similar to a self-antigen or not be recognized at all by any of the T cells.

1.3.2 *Implementation*

"How does the Antigen presentation Pathway work."

The antigen presentation pathway starts with a large protein complex called the **proteasome**, whose main function is the degradation of intracellular proteins. The proteasome cuts existing cytosolic proteins into peptide fragments, and releases these fragments back into the cytosol (Fig. 1.1). Here, the peptide fragments are further degraded by amino-peptidases to single amino acids. From this collection of degrading peptide fragments, the transporter associated with antigen processing (**TAP**) binds to peptide fragments of 10-12 amino acids long, and transports these fragments into the Endoplasmic Reticulum (ER). Once in the ER, these peptide fragments, or **epitope precursor**, are loaded onto

[4] Aside from the two MHC classes, there is a third antigen presentation pathway which has specialized in presenting bacterial and autosomal lipids (Lawton and Kronenberg, 2004)

[5] The second class of MHC molecules, which can be found on certain immune cells, is specialized in presenting epitopes derived from extracellular content to CD4+ T helper cells.

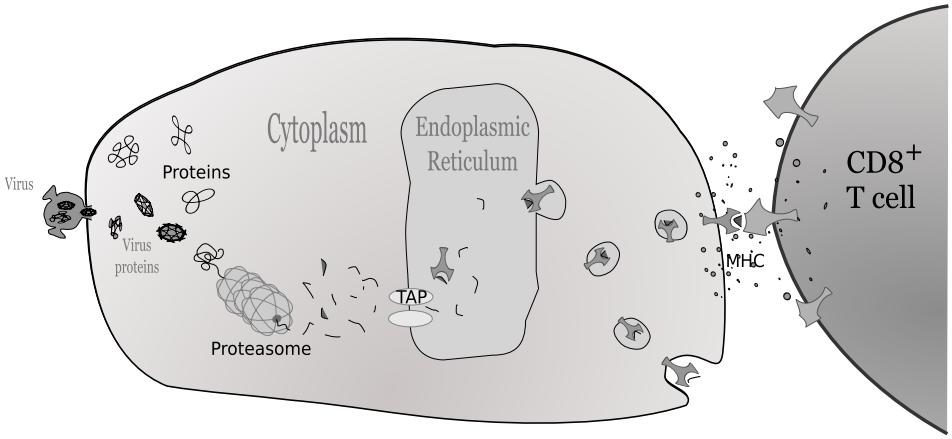

Figure 1.1 Schematic of the classical antigen presentation pathway. Both virus and host proteins are cleaved by the proteasome into small peptides. Some of these peptides are transported by TAP into the endoplasmic reticulum, where they bind to MHC class I alleles, whereas others are reduced in the cytosol to individual amino acids by N-terminal amino-peptidases. The MHC-peptide complexes are transported to the cell surface, where $CD8^+$ T cells can scan the peptides, and thus attack cells that contain virus proteins.

MHC class I molecules, while others are further degraded by amino-peptidases to single amino acids (Craiu et al., 1997). The MHC-epitope complexes are transported to the cell surface, where the **CTL epitopes** are exposed to the T cell receptors of $CD8^+$ T cells.

There are several intermediate steps in the antigen presentation pathway that trim protein fragments. These peptidases predominantly determine the rate with which protein fragments are degraded in the cytosol and the endoplasmic reticulum. Most (<99%) of the protein fragments in the cytosol are destroyed by amino-peptidases before they can be transported by TAP (Reits et al., 2004). As these peptidases appear to be aspecific to the amino acid composition of the protein fragment, and degrade all protein fragments at relatively the same speed (Reits et al., 2004), their role in determining the CTL epitope repertoire of viruses appears to be limited.

1.3.3 *Specificity*

Not every possible protein fragment can be presented by the pathway. Each of the three steps in the antigen presentation pathway has a certain specificity, a limited 'search image' of what amino acid patterns are suitable for proteasomal cleavage, TAP transport, or MHC binding (Burroughs et al., 2004). Given that a typical host is heterozygous for both MHC class I loci in humans that are associated with antigen presentation to the cellular immune system (i.e. HLA-

A and HLA-B[6]), the antigen presentation pathway presents 4-8% of all possible peptides in a protein. This percentage should be large enough to guarantee that even the smallest viruses generates several CTL epitopes[7]

Because the 'search image' of the pathway is constant, pathogens can escape the presentation of CTL epitopes by substituting some of the amino acids in their proteins. The only constraint for pathogens is that escape mutations possibly affects protein function, and therefore pathogen fitness. However, viruses like HIV-1 tolerate a large amount of amino acid variation in their proteins (Brander et al., 2006), and it seems possible that a pathogen could escape all of its important CTL epitopes in a particular host.

1.3.4 *Functional Polymorphism*

"Why does the Antigen presentation Pathway keep working?"

If all human hosts carried an identical antigen presentation pathway, then pathogens could have escaped the cellular immune system of all humans by accumulating a limited number of CTL epitope escape mutations. However, many variants of the MHC alleles of the antigen presentation pathway exist in the human population[8], each of them with a different binding preference. Different hosts 'sample' different parts of a pathogen, and as a consequence, each time the pathogen is transmitted to a new host, the CTL epitopes that are under selection pressure change as well. Escape mutations that were made in previous hosts would typically not be under selection pressure in the next host, and are expected to revert back to original protein sequence to restore optimal protein function (e.g. Leslie et al. (2004)).

1.4 POLYMORPHISM

1.4.1 *Origin of the MHC Polymorphism*

It seems sensible from the perspective of a host population to prefer an MHC polymorphism in the antigen presentation pathway. However, it is not neccesary to invoke the often complex-to-prove group selection arguments, as there are direct fitness advantages for individual hosts to carry a mutant MHC allele (Penn et al., 2002).

Two fitness advantages associated with expressing a new mutant MHC alleles are the Heterozygote Advantage (HA) and the Rare Allele Advantage

[6] Although there are CTL epitopes described for HLA-C, the MHC molecule is poorly expressed on the cell surface (Neisig et al., 1998), and (one of) its functions is to act as a ligand for NK-cells (Romero et al., 2008).

[7] Even for Hepatitis B, which has a genome of just over 3000 bp and is one of the smallest known animal viruses (Kay and Zoulim, 2007), 79 CTL epitopes have been reported (see http://www.immuneepitope.org)

[8] As of June 2009, there are 767 known variants of HLA-A, and 1178 variants of HLA-B (IMGT/HLA database)

(RAA). The **Heterozygote Advantage** is the advantage that a heterozygous host has over a host that inherited identical MHC alleles from his parents. Having two different MHC alleles for binding epitope precursors, the heterozygous host can present a wider range of CTL epitopes on its cell surface. This more or less doubles the number of immune responses that can be mounted against pathogens, and increases the chance of targeting an immunodominant epitope (Doherty and Zinkernagel, 1975; Carrington et al., 1999). The **rare allele advantage** comes into play when pathogens are adapting to the MHC alleles that they encounter in the population. Carrying a rare MHC allele makes it less likely for the host that it will be infected with a pathogen that carries escape mutations specific for that particular MHC allele (Slade and McCallum, 1992; Langefors et al., 2001; Borghans et al., 2004). The HA and especially the RAA provide sufficient selection pressure to drive the evolution of a high degree of MHC polymorphism (de Boer et al., 2004).

1.4.2 *Proteasome and TAP*

The two other steps of the antigen presentation pathway, i.e. the proteasome and TAP, are monomorphic. Both steps exhibit genetic variation in the form of single nucleotide polymorphisms, but this genetic variation appears not to result in functionally different alleles (Gomez et al., 2006; Alvarado-Guerri et al., 2005; Faucz et al., 2000). The lack of polymorphism in the human proteasome and TAP does not appear to be a restriction imposed by the structure and task of these proteins: for both the proteasome and TAP, functional variants exist.

Several different variants of the proteasome already exist in every human; the most notable variants are the default constituitive proteasome, and the immunoproteasome, which is present in cells exposed to an immunomodulatory cytokine IFN-gamma (Tanaka and Kasahara, 1998). A more localized variant of the proteasome is the thymoproteasome, expressed in the human thymus (Murata et al., 2008), and an hypothesized testis-specific proteasome (Tanaka, 2009). All these variants cleave intracellular proteins, but with different preferences for the type of amino acids after which to cleave proteins (Toes et al., 2001; Kesmir et al., 2003). As the C-terminal of most CTL epitopes is determined by the proteasome, these different cleavage patterns of the proteasome have a large impact on the epitopes that are presented on the cell surface (Craiu et al., 1997; Rock et al., 2002). In general, cells in an inflamed environment, such as at the location of a virus infection, will predominantly express the immunoproteasome. Despite all these functionally different proteasomes exist, there does not appear to be a functional polymorphism in humans. Everyone carries the same set of proteasomes.

TAP does not have a functional polymorphism in humans, (Gomez et al., 2006; Alvarado-Guerri et al., 2005) but is speculated to be functionally polymorph in birds (Sironi et al., 2008), trouts (Jensen et al., 2008), and is confirmed to be functionally polymorph in rats (Heemels et al., 1993; Gubler et al., 1998).

1.4.3 *Why did only the MHC become polymorphic?*

With an explanation for the MHC polymorphism, the story of the classical antigen presentation pathway seems complete. The proteasome cleaves proteins, TAP transports some of the resulting peptide fragments to the ER, where the epitope precursors form a complex with MHC class I molecules. The complex is transported to the cell surface, where T cells can inspect the epitopes presented by the MHC. Pathogen adaptation to this pathway is twarted by the extensive polymorphism found in the MHC alleles, which evolved due to the heterozygote and rare allele advantage.

However, the existence of two monomorphic steps in antigen presentation pathway is poorly explained. More specifically, the following two questions have not been answered:

"Why do pathogens not adapt to the monomorphic proteasome and TAP?"

"Why did only the MHC become polymorphic?"

From a pathogen perspective: recent studies of HIV-1 escape mutations (Yokomaku et al., 2004; Brander et al., 1999) had shown that pathogens could escape CTL epitopes by escaping proteasomal cleavage or TAP transport. Pathogens that accumulated epitope precursor escapes would in theory not be affected by the MHC polymorphism in the human population, which would be a major weakness of the antigen presentation pathway (Yusim et al., 2002).

From the a perspective: the heterozygote advantage applies not only to the MHC; a host heterozygous for proteasome or TAP would present a wider range of CTL epitopes, and therefore, in analogy with MHC heterozygosity, would have had a fitness advantage over hosts with a homozygous pathway. Furthermore, if pathogens are not only escaping MHC binding, but also proteasomal cleavage and TAP transport, the rare allele advantage would apply to all of the steps in the antigen presentation pathway. The expected result would be an antigen presentation pathway in which all three steps are polymorphic.

1.5 HOW TO ANSWER?

"How can the Antigen presentation pathway do its job, despite the rapid evolution of pathogens?"

There is ample information about the current state of the Antigen presentation pathway. However, data on the evolution of the pathway is scarce (Lawlor et al., 1990; Danchin et al., 2004). There is evidence that the different steps of the pathway are coevolved to optimize epitope presentation (Toes et al., 2001; Kesmir et al., 2003), and that some MHC alleles can persist in a population for millions of years (Mayer et al., 1988), whereas others are relatively young (Watkins et al., 1992).

In the 1980's a new disease, named AIDS (Acquired Immune Deficiency Syndrome) caught the attention of the medical and scientific community. Once it became clear that a virus was the causing agent of AIDS, blood samples were routinely taken from AIDS patients, and the virus in the blood was isolated and its protein amino acid sequence analysed. As a result of this practice, there is now more than 30 years of data available on the evolution and adaptation of HIV-1 to its new human host.

Combining this data set with recently developed algorithms that can predict the CTL epitopes in a protein (Peters and Sette, 2005; Nielsen et al., 2004), it was possible to study how a pathogen that is new to the human host adapts itself to the human antigen presentation pathway.

We discovered that the epitope and epitope precursor density in HIV-1 had remained constant in the last 3 decades, and described a mechanism by which the monomorphic proteasome and TAP components of the pathway would be (partially) protected from adaptation by the MHC polymorphism (Chapter 2). Subsequently, we tested a theory on the large-scale adaptation of HIV-1 prior to the 1980s (Yusim et al., 2002), but found no conclusive evidence for this to be the case (Chapter 3).

We tested the hypothesized mechanism by which proteasome and TAP were protected in an individual-based model, in which a virus modeled after HIV-1 was allowed to adapt to a host population. We studied the effect of different degrees of MHC polymorphisms on the maximum level of adaptation that the virus could reach to that population, and showed in more detail why pathogens could not exploit the monomorphic proteasome and TAP (Chapter 4).

Finally, we extended the model such that the host population could evolve all components of its antigen presentation pathway in response to multiple endemic pathogens. This allowed us to study whether the current structure of our antigen presentation pathway could be explained just in terms of host-pathogen coevolution (Chapter 5).

The specificity and polymorphism of the MHC class I prevents the global adaptation of HIV-1 to the monomorphic proteasome and TAP.

2

Boris V. Schmid[1], Can Keşmir[1,2], Rob J. de Boer[1]

1 Theoretical Biology, Utrecht University, The Netherlands.
2 Academic Biomedical Centre, Utrecht University, The Netherlands.

PLoS ONE **3(10)**: e3525 (2008)

2.1 ABSTRACT

The large diversity in MHC class I molecules in a population lowers the chance that a virus infects a host to which it is pre-adapted to escape the MHC binding of CTL epitopes. However, viruses can also lose CTL epitopes by escaping the monomorphic antigen processing components of the pathway (proteasome and TAP) that create the epitope precursors. If viruses were to accumulate escape mutations affecting these monomorphic components, they would become pre-adapted to all hosts regardless of the MHC polymorphism. To assess whether viruses exploit this apparent vulnerability, we study the evolution of HIV-1 with bioinformatic tools that allow us to predict CTL epitopes, and quantify the frequency and accumulation of antigen processing escapes. We found that within hosts, proteasome and TAP escape mutations occur frequently. However, on the population level these escapes do not accumulate: the total number of predicted epitopes and epitope precursors in HIV-1 clade B has remained relatively constant over the last 30 years. We argue that this lack of adaptation can be explained by the combined effect of the MHC polymorphism and the high specificity of individual MHC molecules. Because of these two properties, only a subset of the epitope precursors in a host are potential epitopes, and that subset differs between hosts. We estimate that upon transmission of a virus to a new host 2/3rd of the mutations that caused epitope precursor escapes are released from immune selection pressure.

2.2 INTRODUCTION

Antigen presentation allows $CD8^+$ T cells to monitor the protein content of a cell and detect the presence of intracellular viruses (Paulsson, 2004). The clas-

sical antigen presentation pathway consists of three main steps: the (immuno-) proteasome, which cleaves cytoplasmic proteins into peptide fragments; the transporter associated with antigen processing (TAP), which transports peptide fragments into the endoplasmic reticulum; and the major histocompatibility complex (MHC) class I, which binds a small fraction of these endoplasmic peptide fragments (Assarsson et al., 2007), and transports them to the cell surface (Craiu et al., 1997; Rock et al., 2002; Groothuis et al., 2005). The peptide fragments that are processed by the proteasome and transported by TAP are commonly called 'epitope precursors'.

Of these three steps in the antigen presentation pathway it is only the MHC that is highly polymorphic, which is thought to have evolved because of a *rare allele advantage* (Snell, 1968; Bodmer, 1972; Borghans et al., 2004): hosts that carry rare MHC alleles are less likely to be infected by viruses that are adapted to escape the host's MHC alleles than hosts with common MHC alleles, because it is less likely that these viruses come from a host with the same rare MHC alleles. Therefore hosts with rare MHC alleles are thought to have a fitness advantage. Indeed, hosts that were infected with pre-adapted variants of the human immunodeficiency virus 1 (HIV-1) were found to progress rapidly to AIDS (Goulder et al., 2001; Leslie et al., 2004; Chopera et al., 2008). However, if viruses adapt to escape the epitope precursors (Bergmann et al., 1994; Beekman et al., 2000; Allen et al., 2004; Milicic et al., 2005), which are created by the monomorphic proteasome and TAP, the protective effect of the MHC polymorphism and the fitness advantage of hosts with rare MHC alleles would be lost.

We studied the ability of HIV to generate and accumulate epitope and epitope precursor escapes, using algorithms that can reliably predict the likelihood of proteasomal cleavage, TAP transport, and MHC binding of amino acid sequences (see Material & Methods). We discovered that there is no accumulation of epitope precursor escapes on the population level: the total number of epitope precursors (as well as that of epitopes) has remained relatively constant over the last 30 years. We explore several possible causes for this lack of adaptation to the antigen processing machinery, and postulate a mechanism by which the specificity and polymorphism of the MHC prevents the adaptation of viruses to the monomorphic parts of the antigen presentation pathway.

2.3 MATERIAL & METHODS

2.3.1 *CTL epitope predictions*

Currently, a wide variety of algorithms (Peters and Sette, 2005; Nielsen et al., 2004; Parker et al., 1994; Doytchinova et al., 2006) are available to predict MHC-peptide binding. The capacity of these algorithms to identify new epitopes has routinely been benchmarked on experimental data (Peters et al., 2006; Larsen et al., 2007), and their accuracy has increased over time to such an extent that the correlation between predicted an measured binding affinity is as good as the correlation between measurements from different laboratories (Peters et al.,

2006). A further increase in accuracy of identifying Cytotoxic T lymphocytes (CTL) epitopes is achieved by combining the MHC binding predictors with predictors trained to mimic the specificity of the proteasome and TAP, thus creating a model of the complete antigen presentation pathway (Larsen et al., 2005; Tenzer et al., 2005; Doytchinova et al., 2006). These pathway models come in two types: those that sum the scores of the independent steps of the antigen processing pathway, and use a threshold on the summed score (e.g. MHC-pathway (Tenzer et al., 2005) and NetCTL (Larsen et al., 2005)), and those that eliminate epitope candidates at each step (e.g. EpiJen (Doytchinova et al., 2006), MAPPP (Hakenberg et al., 2003) and the alternative implementation of MHC-pathway (Tenzer et al., 2005)).

In this study we use the alternative implementation of the MHC-pathway model (Tenzer et al., 2005). We screen all possible peptide fragments of 14 amino acids within a particular protein, and eliminate those fragments that cannot be correctly processed by either the proteasome, TAP or the MHC class I molecules (see Fig. 2.1). This approach allows us to distinguish between adaptation of a virus to antigen processing and adaptation to MHC binding. The threshold values for the proteasome and TAP predictors (see Fig. 2.1) were derived by applying the MHC-pathway model to a large bacterial protein data set and selecting threshold values which correspond to the estimated specificity of the proteasome (33%) and TAP (76%) (Burroughs et al., 2004). For the MHC-binding predictions we used the default threshold of -2.7, which corresponds to an IC50 threshold of 500 nM (Peters et al., 2006; Assarsson et al., 2007). As a result of using 500 nM as the threshold for MHC binding our analysis focused on the medium to strong HIV-1 epitopes, and disregarded the weaker CTL epitopes in the 500-5000 nM range in favor of a higher specificity (i.e. less false positives) of the MHC-pathway model. The dependency of our results on the selected thresholds and the selected predictor was tested by repeating the population and ancestor analysis for the HIV-1 clade B ENV, GAG and NEF proteins, using a more relaxed MHC binding criteria (5000 nM), as well as using another prediction algorithm, NetCTL(Larsen et al., 2005). The predictors used in this paper are available through a web interface (http://tools.immuneepitope.org/analyze/html/MHC_binding.html 2006-01-01 version). Note that we excluded 2 of the 34 available MHC predictors. The A*3002 predictor was very non-specific at our thresholds, predicting MHC binding in as much as 9944 out of 50.000 HIV-1 derived 9mers (20%). The B*0801 MHC predictor appeared to be very specific, and predicted no MHC binding in 50.000 HIV-1 derived 9mers at the default threshold.

2.3.2 *Prediction quality*

Epitope predictors are routinely tested on large sets of epitopes derived from various pathogens (Lundegaard et al., 2006; Peters et al., 2006). More recently, Larsen et al. (2007) specifically tested the performance of four widely used predictors on a data set of only HIV-1 epitopes. In that study, NetCTL and

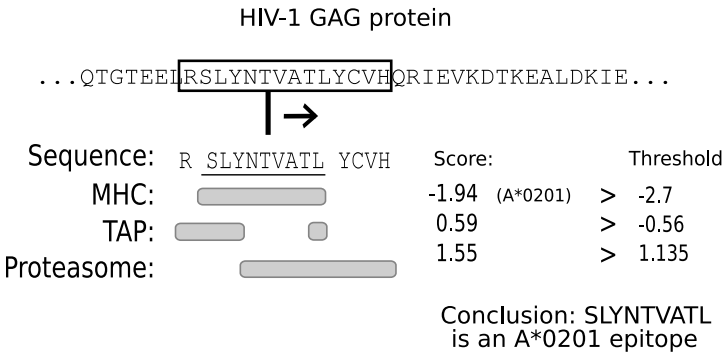

Figure 2.1 Schematic of the MHC-pathway model. A window of 14 amino acids is slided across a protein. Each of these '14mers' consist of a N-terminal flanking region of 1 amino acid, a 9mer epitope candidate and a C-terminal flanking region of 4 amino acids. Beneath the 14mer it is marked which parts of the peptide are used by the MHC, TAP or proteasome predictors respectively. Applying the 14mer to the MHC, TAP and proteasome predictors results in three different scores. If each of these scores is higher than a fixed threshold, then the 9mer embedded in the 14mer is predicted to be a CTL epitope for MHC allele tested (in this case A*0201). If a 14mer passes at least the proteasome and TAP predictors then the 9mer embedded in the 14mer is predicted to be an epitope precursor. In the analysis of longitudinal within-host data sets, CTL epitopes are scored as escapes if mutations in the 14-mer lower one of the three scores below the corresponding threshold.

MHC-pathway came out as the best performing algorithms. MHC-pathway is estimated to recover 80% of the known epitopes at a specificity of 90%, and recover 30% of the known epitopes at a specificity of 99.3%. However, Larsen et al. (2007) stressed that these specificity ratings were underestimates, as the test data set was build with the assumption that any peptide that isn't a known CTL epitope must be a non-epitope, due to the lack of confirmed non-epitopes. As a result, many of the correctly predicted but experimentally not yet verified epitopes were scored as erroneous predictions. A second issue that makes the exact estimation of prediction quality difficult is that many of the experimentally confirmed epitopes are based on CTL responses measured against overlapping peptide pools, and are often defined as the best responding amino acid substring within a peptide that elicits a T cell response, regardless of whether or not this substring is the peptide that can be naturally processed. A more reliable way to estimate the specificity of the predictors is to predict a set of CTL epitopes and subsequently verify $CD8^+$ T cell responses against these epitopes experimentally. Schellens et al. (2008) identified 18 new CTL epitopes out of a set of 22 predicted CTL epitopes in this manner (using NetCTL). This suggests that the specificity of the predictors is far higher than the benchmark estimates, and places the amount of false positive predictions at 20%. Prez

et al. (2008) identified 114 out of 184 predicted epitopes (38% false positives) in a similar manner, but predicted CTL epitopes for the MHC supertypes rather than genotypes, which may explain their higher rate of false positives. A more direct approach to measuring predictor quality is the use of mild acid elution and mass spectrometry to determine MHC-binding peptides. Using these techniques (Fortier et al., 2008) estimated the false positive rate of the MHC-binding predictors of the MHC-pathway model to be less than 2%.

2.3.3 *HIV-1 longitudinal within-host data*

In December 2007, we performed an exhaustive search on the HIV Sequence Database (http://www.hiv.lanl.gov) for longitudinal within-host sequences from 4 digit HLA-genotyped patients that had not received antiretroviral therapy, and for which at least 3 matching MHC predictors were available. This resulted in a data set of 13 patients for which GAG, NEF and POL protein sequences were available (see Table 2.2 and Table 2.3 for sampling dates and accession numbers). All patients were infected with HIV-1 clade B, and their sample HIV-1 sequences spanned a time period of at least two years. The time between infection with HIV-1 and extraction of the early sequence sample was in all cases less than a year. Surprisingly, 8 out of 13 patients carried HLA-B5701 and/or HLA-B2705, two rare and protective alleles (Klein et al., 1998; Kaslow et al., 2001), which probably reflects an observation bias in the data base. All longitudinal within-host sequences were translated from nucleotide to protein sequences with the GeneCutter tool (http://www.hiv.lanl.gov). For patient PIC1362(1052829) multiple sequences per protein per timepoint were available, with small differences between each sequence. Not knowing which of the early timepoints (if any) was the ancestral sequence of the late timepoints complicated some of the within-host analysis. We took a prudent approach by excluding from the analysis the amino acid positions and CTL epitopes for which population dynamics effects could not be ruled out. For example, if a particular epitope was present in the majority of the early timepoint sequences, but not in any of the late timepoint sequences there are two possibilities: the early timepoint sequences that still contained the epitope escaped it, or the early timepoint sequences that did not contain the epitope outcompeted those sequence that did contain the epitope. In such cases the epitope was excluded from the analysis.

2.3.4 *HIV-1 population data*

The HIV-1 population data set used in this paper is the HIV-1 clade B subset of the aligned HIV-1 Sequence Compendium 2002 (Dec 2007 version) (Kuiken et al., 2003). This data set was pruned of sequences for which the sampling date was unknown. The sequence compendium consists of 9 aligned fasta files, one for each of HIV-1's proteins. The number of available HIV-1 clade B sequences in the compendium differs per protein and ranges from 96 to 386 sequences (see Table 2.4 for details). The correlations in Fig. 2.2 and Fig. 2.3 were determ-

ined with the Kendall Tau rank correlation test (Kendall, 1938) of the statistical package R (Ihaka and Gentleman, 1996). The predicted HIV-1 clade B ancestor sequence (Korber et al., 2000) (available at http://www.hiv.lanl.gov) was aligned to the population data set with HMMER 2.3.2 (Eddy, 1998), a profile hidden Markov model.

2.4 RESULTS

2.4.1 *Adaptation to the human population*

To determine whether HIV has exploited the lack of polymorphism of the proteasome and TAP, we predicted the number of epitope precursors in a HIV-1 clade B sequence population data set sampled between 1980 to 2005 (see Material & Methods for details on the HIV-1 Sequence Compendium data set (Kuiken et al., 2003) and the quality of the MHC-pathway model (Larsen et al., 2007)). We plotted the predicted epitope precursor density of each HIV-1 sequence against its sampling date to study the changes over time (Fig. 2.2, first column & Table 2.4). Using 32 MHC peptide binding predictors, the same procedure was performed for the average density of MHC-binding peptides, and for the average density of CTL epitopes. In all three cases there was no sign of any large-scale adaptation of HIV-1 clade B over the last 30 years: the number of epitope precursors, MHC-binding peptides and CTL epitopes per HIV-1 sequence remained constant over time. Differences existed mainly between proteins: the envelope protein (ENV) seemed more immunogenic and had a higher density of precursors, MHC-binders and CTL epitopes than the other proteins, and the NEF protein showed a far greater variability between sequences than the other proteins. In addition to the proteins shown in Fig. 2.2, the same analysis was performed for the other proteins of HIV-1 (Table 2.4), for two other HIV-1 clades (clade C (Table 2.6) and clade A1 (Table 2.7)), as well as for human subpopulations (Kroatia, UK & USA (Table 2.5)) within the HIV-1 clade B population data set. This resulted in a total of 102 tests, of which 14 were found to be significant at a p-value of < 0.01 (Kendall Tau rank correlation test). However, in 6 out of these 14 significant cases HIV-1 was gaining epitope precursors, MHC-binding peptides, or CTL epitopes over time. The 7 cases in which the density significantly decreased over time were not consistently occurring in the same proteins when comparing different HIV-1 clades, nor were they consistently affecting the same step in the antigen presentation pathway. This makes it unlikely that these 7 correlations reflect any adaptation of HIV-1 to its human host. We verified this result by repeating the analysis for ENV, GAG and NEF using a more relaxed MHC binding threshold (5000 nM), as well as with a different prediction algorithm (Larsen et al., 2007) (data not shown). This did not result in a qualitative difference, except that according to the NetCTL predictions (Larsen et al., 2007) the epitope precursor density in HIV-1 NEF decreased slowly, but significantly over time (from 49 to 46 precursors over a 25y period, Kendall Tau rank correlation test, $p < 0.01$).

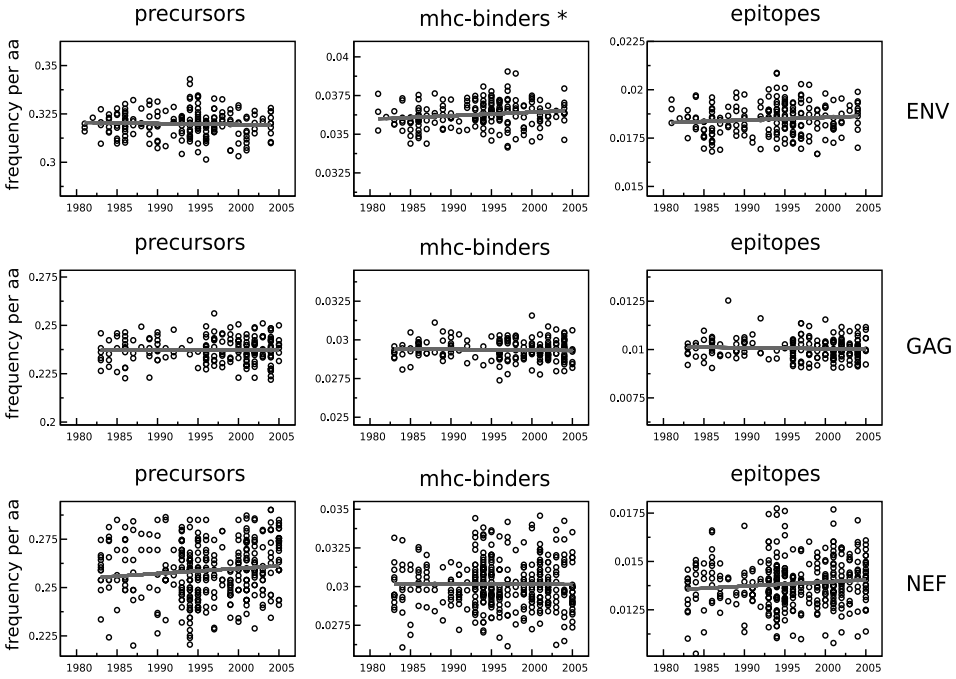

Figure 2.2 The predicted density of epitope precursors, MHC-binding peptides and CTL epitopes in ENV, GAG and NEF stay constant over time. The density (expressed as frequency per amino acid) is plotted on the y-axis with the same scale factor within each column, which makes it possible to compare differences between proteins. The densities for MHC-binding and CTL epitopes are averaged over the 32 MHC-binding predictors. *: significant increase over time (Kendall Tau rank correlation test, $p < 0.01$).

That the number of MHC-binding peptides in HIV-1 remained constant over time (Fig. 2.2, middle column) was to be expected, based on the theory that the MHC polymorphism prevents pathogens like HIV-1 from escaping MHC binding on a population level (Slade and McCallum, 1992; Borghans et al., 2004), and earlier reports that the virulence of HIV-1 had not changed over time (Müller et al., 2006; Herbeck et al., 2008). However, other studies reported that HIV-1 was capable of adapting to common MHC alleles (Moore et al., 2002; Leslie et al., 2005; Poon et al., 2007), and suggested that HIV-1 was adapting to its new human host population. Follow-up studies on Moore et al. (2002) showed that their analysis was sensitive to founder effects in the viral lineage (Bhattacharya et al., 2007) and that without these effects the adaptation of HIV-1 to the population could only be detected on a small number of amino acids (Brumme et al., 2007).

In line with our *a priori* expectations, Yusim et al. (2002) proposed that the clustering of epitopes in HIV-1 proteins was a result of adaptation of the virus to the monomorphic proteasome and TAP. However, our results (Fig. 2.2) refute

that expectation: HIV-1 has not accumulated epitope precursor escapes over the last 30 years. In this paper we explore possible reasons for the apparent lack of adaptation of HIV-1 to the monomorphic antigen processing machinery.

2.4.2 *Selection pressure by CD8$^+$ T cells*

A simple explanation why we find that HIV-1 is not accumulating CTL epitope and epitope precursor escapes (Fig. 2.2) would be that CD8$^+$ T cells exert too little selection pressure on the virus. The role of CD8$^+$ T cells in controlling a chronic HIV-1 infection has been under debate (Zhang et al., 2003; Asquith et al., 2006), but the strongest evidence that the virus is under selection pressure of CD8$^+$ T cells is that certain immune escape mutations are rapidly reverted to the wildtype upon entering an HLA-mismatched host (Leslie et al., 2004; Furutsuki et al., 2004; Li et al., 2007; Frater et al., 2007). Additional evidence comes from CD8$^+$ depletion studies of chronic SIV infections in monkeys (Matano et al., 1998; Jin et al., 1999; Schmitz et al., 1999), from studies that show that MHC-heterozygous hosts progress slower to AIDS than homozygous hosts (Carrington et al., 1999), and from correlates between HIV-1 disease progression and the presence or absence of certain MHC class I molecules (Klein et al., 1998; Kaslow et al., 2001).

In addition to the strong evidence on the selection pressure imposed by CD8$^+$ from the current literature, we studied CD8$^+$ T-cell mediated immune selection pressure on HIV-1 by testing whether amino acid replacement mutations happen preferentially in CTL epitopes and their flanking regions. For this purpose we data-mined the Los Alamos HIV database for longitudinal within-host HIV-1 sequence data from MHC genotyped and treatment-naive patients (see Material & Methods, and the Supplemental Materials). This resulted in 13 patients for which GAG, NEF and POL protein sequences were available (see Table 2.2, Table 2.3). We compared for each of these proteins the number of amino acid replacements that occurred within predicted epitopes or their flanking regions to the expected number of mutations. This expected number of mutations is based on the fraction of the protein that the epitopes and their flanking regions covered ('epitope cover'). We found a trend towards mutations occurring within CTL epitopes or their flanking regions for the three HIV-1 proteins that were tested (Wilcoxon signed rank test: $p = 0.09$, with 15 out of 21 samples following the trend).

A surprising observation based on our immune selection pressure study was that the predicted CTL epitopes within a single host can cover a large fraction of the viral proteome. For those samples in the longitudinal within-host data set where predictors were available for all four of the Human Leukocyte Antigen A (HLA-A) and HLA-B alleles of the host, the epitope cover ranged from 12% to 74%. The average epitope cover of a single MHC allele for the HIV-1 clade B HXB2 reference sequence was 17%, and all 32 MHC predictors together covered 94% of the HXB2 proteome.

That we find no significant correlation between the location of mutations and predicted CTL epitopes might be due to differences between CTL epitopes in the strength of the immune selection pressure imposed on them, which would reduce the detection power of our method of testing for selection pressure. Yewdell and Bennink (1999) indicated that only half of all CTL epitopes can trigger a $CD8^+$ T cell response. The underlying mechanism is poorly understood, but possibly involves self-tolerance (Rolland et al., 2007; Frankild et al., 2008). Zafiropoulos et al. (2004) and Frater et al. (2007) showed that the selection pressure imposed on the virus differs between CTL epitopes. A possible cause for this variation is whether an epitope is presented early or late during the infection of a cell (van Baalen et al., 2002; Sacha et al., 2007a,b). Since the trend we find is confirmed by the current literature, we assert that the lack of adaptation in Fig. 2.2 is not likely to be due to a lack of immune selection pressure.

2.4.3 *Short timespan of population data set*

Another possible reason why we find that HIV-1 is not accumulating CTL epitope and epitope precursor escapes (Fig. 2.2) would be that the timespan of our population data set (30 years - from 1976 to 2006) is too short to detect an evolutionary process like the adaptation of a virus to its host. To test this, we predicted the epitope precursors of the putative HIV-1 clade B ancestor sequence (Korber et al., 2000) and plotted the fraction of ancestral epitope precursors contained in each sequence of our population data set against the sampling date (Fig. 2.3). In this way the 'immunological similarity' of a sequence with the ancestor sequence can be visualized. This similarity is expected to decline over time, based on the destruction of ancestral epitope precursors by neutral amino acid substitutions (Whitney et al., 1985; Kimura, 1991; Yokomaku et al., 2004), and by the accumulation of escapes from $CD8^+$ T cell responses within hosts. If the time covered by our population data set is sufficient, we should see a decrease over time in the immunological similarity of current-day sequences to the ancestral HIV-1 sequence.

Indeed, for ancestral epitope precursors as well as for ancestral MHC-binding peptides and CTL epitopes, we found that the density declined significantly over time in the six largest proteins of HIV-1 (Kendall Tau rank correlation test, $p < 0.01$ in 17/18 tests). The only exception was a non-significant decrease in the number of predicted ancestral epitopes in HIV-1 NEF (Fig. 2.3, bottom right panel). Analysis of the NEF protein subset of the population data set revealed that in the Kroatian population the number of ancestral epitopes in NEF was increasing over time. Whether this increase reflects a particular adaptation of the virus, or is due to a founder effect in the Kroatian subpopulation that was oversampled in the HIV Sequence Compendium data set is not known. The three smallest proteins of HIV-1 (TAT, VPR and VPU) yielded no significant results, which indicates that at protein sizes of less than 100 amino acids our method becomes insensitive.

The HIV-1 clade B consensus sequence (Fig. 2.3, open square on the right-hand side of each panel) is more similar to the predicted ancestral sequence than most of the HIV-1 sample sequences themselves are, which indicates that HIV-1 is undergoing divergent evolution. This is inconsistent with the idea that HIV-1 is undergoing a large-scale global adaptation to the human host (which would imply a convergent evolution process). Based on the results presented thus far, we conclude that the evolution of HIV-1 seems largely determined by the loss of ancestral epitopes due to antigenic drift, by local adaptation of the virus to each individual host, and by the reversion of earlier adaptations.

Finally, the rate at which ancestral epitopes and epitope precursors disappear (Fig. 2.3, grey lines) gives a novel way to estimate in what year HIV-1 clade B was introduced into the human population. The age of the ancestral HIV-1 B sequence can be predicted by extrapolating the regression line back to where the fraction of ancestral epitopes or epitope precursors in HIV-1 sequences becomes one, assuming that the loss over time has been linear. Each protein and each category (precursor, MHC binding, epitopes) generates a separate prediction for the age of the ancestral sequence. For the larger genes ENV, GAG and POL, the estimated ancestral age is 1939 ± 13, whereas for the smaller genes it is 1900 ± 54 years. The estimate for the larger genes concurs with the findings of Korber et al. (2000), who dated the ancestral sequence on 1920-1940.

Summarizing: the analysis of the loss of ancestral epitope precursors shows that our method is sensitive enough to pick up evolutionary processes in the larger proteins of HIV-1 ($>$ 100 amino acids). Therefore, the lack of adaptation to epitope precursors in Fig. 2.2 should not be attributed to the relatively short time span of the population data set.

2.4.4 *Rarity of precursor escapes*

Brander et al. (1999) hypothesized that the proteasome and TAP should be rather non-specific for their substrate in order to fulfil their intracellular functions. Therefore, most mutations should not affect antigen processing, and as a result epitope precursor escapes would be harder to generate than MHC binding escapes. Although several studies have clearly shown that antigen processing escapes do exist (Brander et al., 1999; Yokomaku et al., 2004), the frequency of successful antigen processing escapes *in vivo* could be so low that these kind of escapes play no role in the evolution of HIV-1, which would explain why HIV-1 is not accumulating epitope precursor escapes (Fig. 2.2).

We used the MHC-pathway model to determine the frequency of antigen processing escapes in a longitudinal within-host HIV-1 sequence data set (27 HIV-1 proteins from a total of 13 different patients, see Material & Methods, and Table 2.2 and Table 2.3). We found that 38 out of a total of 375 predicted CTL epitopes were escaped by the virus (10.1%) during the time spanned by the longitudinal within-host data set. Of these 38 escaped CTL epitopes, 34 (89%) contained one or more mutations that prevented the peptide from binding to its associated MHC molecule, and 6 (16%) contained one or more mutations in

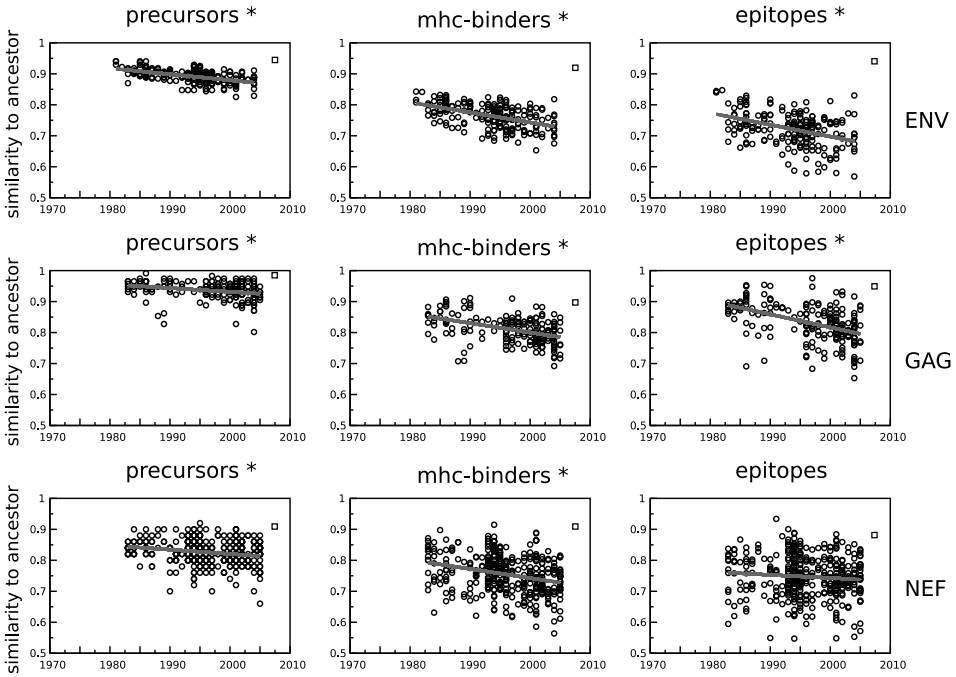

Figure 2.3 The predicted fraction of ancestral epitope precursors, MHC-binding peptides and CTL epitopes present in the HIV-1 clade B sequences declines over time. The predicted fraction is plotted per sequence as a dot on the y-axis. Ancestral epitope precursors are defined by their C-terminal position (Fig. 2.1, 10th amino acid from the left) in the aligned sequences . Similarly, MHC-binding peptides and CTL epitopes are defined by their C-terminal, but also by the MHC that they are predicted to bind to. *: significant decrease over time (Kendall Tau rank correlation test, $p < 0.01$). The consensus sequence is plotted as a square on the right-hand side of each panel.

the epitope or the epitope's flanking region that prevented antigen processing of the epitope precursor.

A second way to study the frequency of epitope precursor escapes is to study the predicted effect of a single amino acid substitution on the number of CTL epitopes in a HIV-1 protein. While this approach completely ignores functional constraints on proteins, it has the advantage that we can calculate the average effect of a single mutation on the escape of CTL epitopes. This procedure was repeated a large number of times, until on average each amino acid in the HXB2 reference sequence had been mutated five times (previous mutations were reversed before a new one was generated). In 3.8% of the cases this procedure resulted in the loss of one or more epitopes per MHC allele. In 38% of these escape mutations, it was the epitope precursor that was no longer processed correctly by the proteasome or TAP (Table 2.1).

The amino acid substitution simulations on the HIV-1 HXB2 reference sequence showed that new CTL epitopes are predicted to be readily created by random mutations (Table 2.1). Similarly, comparing the early and late timepoints of the longitudinal within-host data set showed that 56 new CTL epitopes were predicted to have arisen, which is in good agreement with the simulation results in Table 2.1. Although it seems counter-intuitive, the generation of new CTL epitopes has been shown to occur in reality (Allen et al., 2005; Karlsson et al., 2007). There are several reasons why new CTL epitopes could come about, despite immune selection pressure: 1) a single mutation could escape an epitope against which a strong immune response was directed, while at the same time create a new epitope with a weaker response (Allen et al., 2005), 2) if a small number of strong immune responses determine most of the fitness of the virus, adding a single weak response to the existing weak responses has a negligible effect on the fitness of the virus, 3) the new epitopes might not be recognized by any of the $CD8^+$ T cell receptors of the host (Yewdell and Bennink, 1999), and 4) the time between the generation of a new epitope and the expansion of a $CD8^+$ T cell response against this new epitope provides a time window during which new CTL epitopes in a HIV-1 sequence are not penalized (Barouch et al., 2005).[1]

The analysis of the longitudinal within-host data set and the simulated HIV-1 HXB2 reference sequence mutations established that antigen processing escapes occur relatively frequently, and thus that the predicted lack of antigen processing adaptation of HIV-1 shown in Fig. 2.2 is not because precursor escapes are too hard to generate. The analysis also showed that new CTL epitopes are frequently generated during the within-host evolution of the virus.

2.4.5 *Polymorphism and Specificity*

In the previous sections we investigated three possible explanations for the predicted lack of adaptation of HIV-1 to the monomorphic antigen processing pathway (Fig. 2.2), but found no compelling evidence for any of them. Here we propose an alternative explanation: as each MHC class I allele utilizes only a small

[1] At the population level, MHC-mismatched CTL epitopes will also frequently be generated as a by-product of escape or compensatory mutations in a host. The virus will only experience a selection pressure of these new epitopes when it is transmitted to a MHC-matched host.

Table 2.1 Effect of random amino acid substitutions on the average loss and gain of CTL epitopes per MHC allele in the HIV-1 HXB2 reference sequence.

	loss	*gain*
Mutations affecting epitope count	3.8%	4.3%
due to processing	38%	40%
due to MHC binding	80%	81%

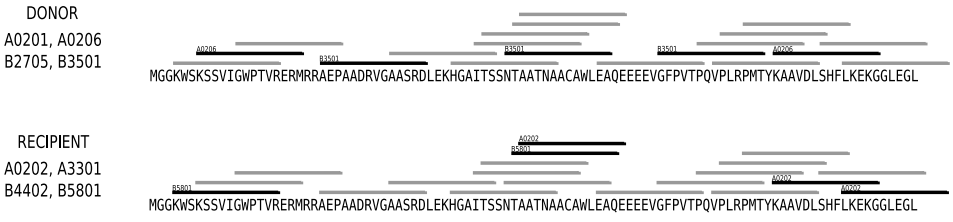

Figure 2.4 The effect of MHC specificity and polymorphism on the selection pressure on epitope precursors. This example shows the first 100 amino acids of the NEF protein of the HIV-1 HXB2 reference sequence, and the precursors and epitopes for two fictitious hosts. The lines represent the epitope precursors generated by the monomorphic proteasome and TAP. The black lines depict those epitope precursors that are used by the MHC alleles. In this example, 4 out of 5 epitope precursors were released from selection pressure, as were 32 out of 70 amino acids.

fraction of the available HIV-1 epitope precursors, not all of the epitope precursors in a host are under selection pressure. When a virus is transmitted from one host to a new host with a different set of MHC molecules a large number of the epitope precursors that were previously under immune selection pressure are no longer so. Escape mutations in those epitope precursors can subsequently revert to the wildtype sequence. A visual example of this mechanism is depicted in Fig. 2.4, in which a HIV-1 protein is passed from one fictitious host to another.

While the proposed mechanism is straightforward and plausible, its protective effect depends on the fraction of epitope precursors that is under selection pressure in the donor host, but no longer in the recipient host. This is directly influenced by the specificity and promiscuity of the MHC alleles of both host and donor: the more specific the MHC binding is, the smaller the subset of epitope precursors is that is used by the MHCs of the host, and therefore the larger the typical fraction of epitope precursors is that is released from selection pressure when the virus changes from one host to the other. We estimated this fraction with a simple model in which we create fictitious hosts with random sets of MHC alleles, and transmit the HIV-1 HXB2 reference sequence from one host to another. Each time the virus is transmitted, we calculate the fraction of the epitope precursors that were used by the MHC alleles of the donor host, but are not utilized in the recipient host. In this way we estimated that on average 18% of the epitope precursors are under selection pressure in a random host, i.e. are an actual epitope in that host. 66% of these actual epitopes will be released from selection pressure in the next host (see Fig. 2.5). Alternatively, the protective effect can also be calculated at the level of amino acid positions rather than precursors. By doing so we predicted that on average 49% of the amino acid positions are under selection pressure in a random host, and that 39% of this group of 49% is released from selection pressure upon transfer of the virus to a new host. These two estimates represent the extreme ends of how

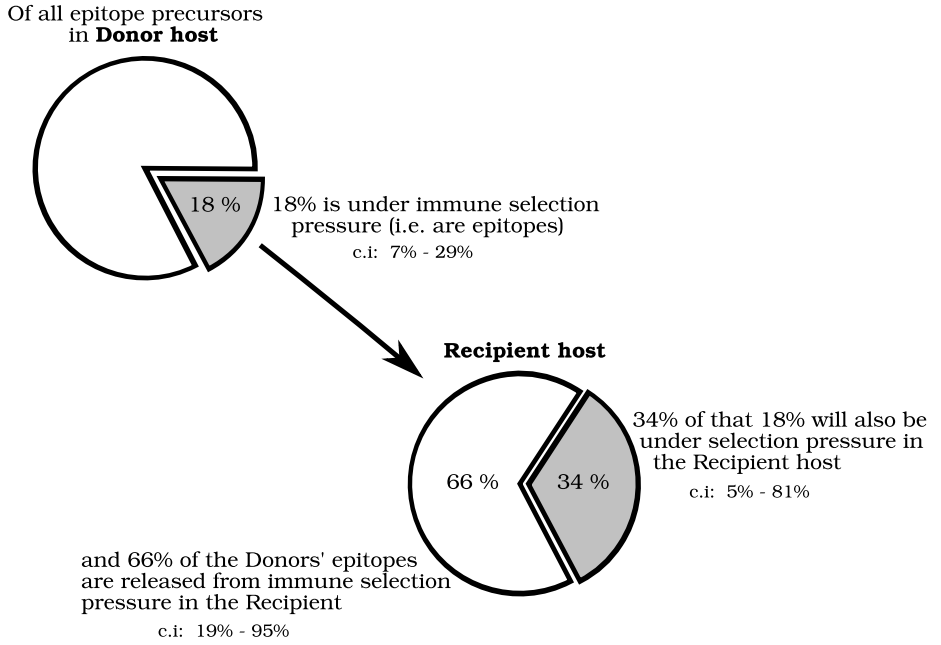

Figure 2.5 The number of epitope precursors that bind to MHC alleles in the donor host but no longer in the recipient host is calculated based on a 1000 simulated passages of the HIV-1 HXB2 sequence between two hosts. The hosts are randomly created from a set of 18 different HLA-A molecules and 14 different HLA-B molecules. On average, the 4 MHC alleles of a host bind 18% (145) of the 818 predicted epitope precursors. Of these 145 epitope precursors, an average of 34% can bind to one of the MHC alleles of a recipient host, whereas 66% is no longer under selection pressure. 95% confidence intervals (c.i) are shown in the figure. The overlap in epitope precursors used by the donor and recipient is partially due to overlapping MHC molecules (by chance in 42% of the cases the donor and recipient shared one or more MHC alleles), and partially due to the promiscuity of MHC alleles.

much escape mutations in one epitope precursor influence the processing and presentation of another epitope precursor. The true fraction of epitope precursors that is released from selection pressure when a virus travels from one host to the next should lie somewhere in between this range of 39% - 66%. Note that the range depends on our chosen threshold for MHC binding of IC50 500 nM (Peters et al., 2006; Assarsson et al., 2007). Increasing the specificity of the MHC binding to 50 nM increases the fraction of released epitopes to a range of 76%-83%, and lowers the average number of predicted CTL epitopes from 145 to 18 epitopes per host per viral sequence. Decreasing the specificity of the MHC binding to 5000 nM decreases the fraction of released epitopes to a range of 6%-31%, but with this threshold we predict an unrealistically large number of CTL epitopes (514) per viral sequence.

Based on these predictions, we argue that the magnitude of this MHC specificity and polymorphism-dependent release mechanism is large enough to play an essential role in slowing down the adaptation of HIV-1 to the proteasome and TAP. Its exact effect on the evolution of HIV-1 will also depend on other characteristics of the virus, such as the transmission rate, and the balance between the rate of escape mutations and the rate of escape reversion (Poon et al., 2007). Combined with our proposed mechanism, these factors will determine the eventual degree of adaptation to the antigen presentation pathway that viruses like HIV-1 can reach.

2.5 DISCUSSION

The total number of (predicted) epitope precursors and CTL epitopes in a large population data set of HIV-1 clade B sequences is not decreasing over time (Fig. 2.2). This is in contrast to our initial expectation that HIV-1 would be able to adapt to the monomorphic steps of the antigen presentation pathway (i.e. the proteasome and TAP) to evade presentation of its proteins on the cell surface. We investigated three possible factors that could explain why we did not detect adaptation; 1) possible lack of $CD8^+$ selection pressure, 2) the (evolutionary) short timespan of 30 years of our population data set (Fig. 2.3), and 3) the possible rarity of epitope precursor escapes (Table 2.1), but found no compelling evidence for any of them.

In the last section of the results we added and discussed a fourth possibility, namely that the adaptation of HIV-1 to epitope precursors is limited by frequent loss of the immune selection pressure on epitope precursor escapes as the virus passes from one host to another (Fig. 2.4). We proceeded by quantifying that a typical proteasome or TAP escape mutation is released from selection in 39% to 66% of the human hosts. We propose that this loss of selection pressure on epitope precursors is one of the main factors that determine the eventual degree of adaptation to epitope precursors that HIV-1 can reach on the population level (Fig. 2.5). Other factors are the transmission rate, the rate at which epitope precursor escapes are acquired, and the rate at which they are lost or reverted (Poon et al., 2007).

Based on our understanding of this mechanism, we speculate that only one of the steps in the antigen presentation pathway has to be polymorphic to prevent pathogens from adapting to any step in the pathway. The mechanism functions best when the polymorphy occurs at the most specific step in the pathway, as that increases the fraction of epitope precursors that is not under selection pressure. While in humans it is the MHC class I molecules that are highly polymorphic and specific, other solutions do appear to exist. The TAP molecules of rats are more specific than the human TAP, and have a limited functional polymorphism (Gubler et al., 1998), and the TAP and MHC genes of chickens are equally polymorphic on the nucleotide level (Walker et al., 2005). We are currently exploring the conditions that determine which of the steps of

the antigen presentation pathway become polymorphic using an agent-based host-pathogen model.

The lack of any large-scale adaptation of HIV-1 to reduce its number of CTL epitopes -as reported by this study- is not necessarily in contradiction with the possible fixation of certain CTL epitope escape mutations at the population level (Leslie et al., 2005; Brumme et al., 2007)[2], especially if these occur in combination with compensatory mutations (Navis et al., 2007). However, our analysis indicates that fixation of particular CTL epitope escapes is not occurring at a scale that makes it detectable amidst the constant destruction and generation of CTL epitopes due to neutral amino acid substitutions (see Fig. 2.3, Table 2.1). Furthermore, we have only studied the adaptation of HIV-1 to escape antigen presentation by means of amino acid substitutions. HIV-1 also influences epitope presentation by blocking TAP transport (Kutsch et al., 2002) and down-regulating MHC molecules (Baugh et al., 2008). Adaptation to the pathway could well be occurring at this level, rather than at the level of individual CTL epitopes. However, current findings that individual escape mutations can have a large impact on viral load during within-host evolution (Karlsson et al., 2007; Maurer et al., 2008), suggests that there is still a strong selection pressure on individual CTL epitopes.

In summary: the monomorphic parts of the antigen presentation pathway are protected from viral immune escape adaptations because only a subset of the epitope precursors can be presented by the MHC alleles of a particular host. Because of the MHC polymorphism, this subset differs per host. As a result, epitope precursor escape mutations are frequently released from immune selection pressure when pathogens spread through a population, and can revert to the wildtype sequence. The protective effect of this mechanism increases with the polymorphism of MHC and the specificity of the individual MHC class I alleles.

2.6 ACKNOWLEDGEMENTS

Part of this work was performed during a visit of Rob de Boer and Boris Schmid to the Los Alamos laboratories and the Santa Fe Institute. We would like to thank Alan Perelson for hosting us and Bette Korber and Karina Yusim for the productive discussions during the early stages of this research. We would also like to thank Becca Asquith for her feedback on this work.

2.7 SUPPLEMENTAL MATERIALS

[2]For a detailed discussion of Kawashima et al. (2009), see Chapter 4.5.

Table 2.2 Details longitudinal within-host data set (part 1).

Patient ID	Protein	Sampling date	Accession number(s)
005(10151393)	NEF	1985	AF129336
		1989	AF129395
PIC1362(10152829)	GAG	1998	DQ853466, DQ853467, DQ853468, DQ853469, DQ853470, DQ853471, DQ853472, DQ853473, DQ853475
		2002	DQ853439, DQ853440, DQ853441, DQ853442, DQ853443, DQ853444, DQ853445, DQ853446, DQ853447, DQ853448
	NEF	1998	DQ853427, DQ853430, DQ853432, DQ853605, DQ853607, DQ853608, DQ853611, DQ853612, DQ853614, DQ853615, DQ853616, DQ853618, DQ853620, DQ853622, DQ853624, DQ853625, DQ853626, DQ853627, DQ853628, DQ853629,DQ853630, DQ853632, DQ853633, DQ853634, DQ853635, DQ853636, DQ853637, DQ853642, DQ853645, DQ853646, DQ853647, DQ853648
		2002	DQ853439, DQ853440, DQ853441, DQ853442, DQ853443, DQ853444, DQ853445, DQ853446, DQ853448, DQ853486, DQ853487, DQ853488, DQ853489, DQ853490, DQ853491, DQ853492, DQ853493, DQ853494, DQ853495, DQ853496, DQ853497, DQ853498, DQ853499, DQ853500
	POL	1998	DQ853466, DQ853467, DQ853468, DQ853469, DQ853470, DQ853471, DQ853472, DQ853473, DQ853474
		2002	DQ853439, DQ853440, DQ853441, DQ853442, DQ853443, DQ853444, DQ853445, DQ853446, DQ853447, DQ853448

For patient PIC1362 additional sequence information is also available in the Los Alamos Database for ENV, REV, TAT, VIF, VPR and VPU. To have at least more than one patient per protein we have limited our analysis of the available sequence data of patient PIC1362 to GAG, NEF and POL.

Table 2.3 Details longitudinal within-host data set (part 2).

Patient ID	Protein	Sampling date	Accession number(s)
MACS1(10159224)	GAG	1985	EF525480
		1989	EF525482
	POL	1985	EF525481
		1989	EF525483
MACS2(10159231)	GAG	1991	EF525500
		1995	EF525502
	POL	1991	EF525501
		1995	EF525503
MACS3(10159232)	GAG	1987	EF525504
		1992	EF525505
MACS5(10159234)	GAG	1984	EF525512
		1989	EF525514
	POL	1984	EF525513
		1989	EF525515
MACS7(10159236)	GAG	1988	EF525520
		1992	EF525522
	POL	1988	EF525521
		1992	EF525523
MACS9(10159238)	GAG	1992	EF525526
		1997	EF525528
	POL	1992	EF525527
		1997	EF525529
MACS11(10159226)	GAG	1985	EF525488
		1990	EF525490
	POL	1985	EF525489
		1990	EF525491
MACS12(10159227)	GAG	1985	EF525492
		1992	EF525493
MACS13(10159228)	POL	1984	EF525494
		1989	EF525495
MACS14(10159229)	POL	1986	EF525496
		1988	EF525497
MACS16(10159230)	POL	1989	EF525498
		1994	EF525499

Table 2.4 Details HIV-1 Clade B population data set.

Protein (# samples)		*P*	*density per aa*	*2008→2032*	*half-life*
ENV (196)					
	Precursors	0.6552	0.319	273.1->272.0	3083 y
	MHC-binders	**0.0085**	0.037	31.3->31.8	
	Epitopes	0.2145	0.019	16.0->16.3	
GAG (186)					
	Precursors	0.9887	0.237	117.4->117.6	
	MHC-binders	0.4551	0.029	14.5->14.4	2924 y
	Epitopes	0.5673	0.010	4.9->4.9	1033 y
NEF (368)					
	Precursors	0.0188	0.266	55.2->56.6	
	MHC-binders	0.6518	0.030	6.2->6.2	83727 y
	Epitopes	0.0231	0.014	2.9->3.0	
POL (97)					
	Precursors	0.0825	0.257	257.4->254.4	1042 y
	MHC-binders	0.6235	0.033	32.6->32.5	4384 y
	Epitopes	0.0424	0.013	12.6->12.3	526 y
REV (96)					
	Precursors	0.3933	0.216	25.5->25.1	858 y
	MHC-binders	0.3137	0.027	3.2->3.1	359 y
	Epitopes	0.1570	0.009	1.0->0.9	127 y
TAT (99)					
	Precursors	0.9730	0.211	18.3->18.4	
	MHC-binders	0.7891	0.017	1.5->1.5	
	Epitopes	0.2038	0.004	0.3->0.4	
VIF (180)					
	Precursors	0.0388	0.271	52.1->53.8	
	MHC-binders	**0.0037**	0.037	7.1->6.9	428 y
	Epitopes	0.3610	0.013	2.6->2.6	1986 y
VPR (160)					
	Precursors	0.1176	0.326	31.3->30.4	444 y
	MHC-binders	0.0206	0.035	3.4->3.3	352 y
	Epitopes	**0.0037**	0.016	1.5->1.4	124 y
VPU (147)					
	Precursors	0.2528	0.337	27.7->28.7	
	MHC-binders	0.5678	0.052	4.3->4.2	1426 y
	Epitopes	0.3480	0.025	2.1->2.0	464 y

The density of epitope precursors, MHC-binding 9mers and CTL epitopes is expressed per amino-acid and where applicable averaged over the 32 available MHC-binding predictors. The '2008→32' column gives the estimated current (2008) number of precursors, average number of MHC-binders and average number of CTL epitopes and projects 25 years into the future, based on linear regression. 'Half-life' is an estimate of the number of years it will take at which half of the precursors, MHC-binders or CTL epitopes have been lost, assuming a linear decline. Statistical test: Kendall Tau rank correlation test, with p-values < 0.001 in bold face.

Table 2.5 Details HIV-1 Clade B sub-population data set.

Protein (# samples)		*P*	*density per aa*	*2008→2032*	*half-life*
HIV-1 Clade B Kroatia					
NEF (153)					
	Precursors	**0.0033**	0.272	56.2->59.2	
	MHC-binders	0.8802	0.030	6.2->6.3	
	Epitopes	**0.0090**	0.014	3.0->3.2	
VIF (70)					
	Precursors	0.3064	0.271	52.0->53.6	
	MHC-binders	0.0749	0.036	7.0->6.7	292 y
	Epitopes	0.1858	0.013	2.5->2.4	203 y
HIV-1 Clade B Great Brittain					
NEF (60)					
	Precursors	0.6408	0.257	53.1->52.4	854 y
	MHC-binders	0.5277	0.029	6.1->5.7	169 y
	Epitopes	0.7420	0.013	2.8->2.6	177 y
HIV-1 Clade B USA					
ENV (81)					
	Precursors	0.1717	0.321	275.1->277.2	
	MHC-binders	**0.0079**	0.037	31.7->32.6	
	Epitopes	0.2258	0.019	16.1->16.7	
GAG (56)					
	Precursors	0.5596	0.236	116.9->116.0	1587 y
	MHC-binders	0.9887	0.029	14.5->14.5	4273 y
	Epitopes	0.3232	0.010	4.8->4.6	248 y
NEF (62)					
	Precursors	0.5100	0.270	56.0->57.3	
	MHC-binders	0.3156	0.031	6.4->6.7	
	Epitopes	0.9902	0.014	2.9->2.9	
VPU (56)					
	Precursors	0.4164	0.329	27.0->26.7	1335 y
	MHC-binders	0.4952	0.051	4.2->4.1	490 y
	Epitopes	0.3067	0.025	2.1->1.9	151 y

See HIV-1 Clade B population data set table for an explanation of the columns. Removed proteins with less than 50 samples. Statistical test: Kendall Tau rank correlation test, with p-values < 0.001 in bold face.

Table 2.6 Details HIV-1 Clade C population data set.

Protein (# samples)		*P*	*density per aa*	*2008→2032*	*half-life*
ENV (346)					
	Precursors	0.8204	0.319	273.5->271.9	2108 y
	MHC-binders	0.0933	0.037	31.3->31.2	3580 y
	Epitopes	0.2452	0.019	15.9->15.7	978 y
GAG (577)					
	Precursors	0.2472	0.239	118.3->116.8	914 y
	MHC-binders	0.0357	0.030	14.6->14.5	2222 y
	Epitopes	**0.0003**	0.009	4.6->4.3	150 y
NEF (374)					
	Precursors	**0.0007**	0.269	55.7->51.5	160 y
	MHC-binders	0.1038	0.030	6.2->5.9	227 y
	Epitopes	0.0152	0.015	3.1->2.8	121 y
POL (279)					
	Precursors	0.7939	0.257	257.9->256.7	2582 y
	MHC-binders	0.9384	0.032	32.5->32.5	180095 y
	Epitopes	0.5383	0.012	12.4->12.5	
REV (295)					
	Precursors	0.1672	0.193	22.7->21.2	181 y
	MHC-binders	0.7996	0.027	3.2->3.2	18684 y
	Epitopes	0.6159	0.006	0.8->0.7	171 y
TAT (286)					
	Precursors	0.4717	0.187	16.3->16.6	
	MHC-binders	**0.0059**	0.017	1.5->1.3	81 y
	Epitopes	0.4379	0.004	0.3->0.3	649 y
VIF (295)					
	Precursors	0.6219	0.286	54.9->55.0	
	MHC-binders	0.0593	0.037	7.1->7.4	
	Epitopes	0.6649	0.013	2.6->2.5	1053 y
VPR (298)					
	Precursors	0.9737	0.313	30.1->30.0	8964 y
	MHC-binders	0.2668	0.035	3.3->3.3	656 y
	Epitopes	0.4666	0.016	1.5->1.6	
VPU (293)					
	Precursors	0.8715	0.376	30.8->30.8	
	MHC-binders	0.6532	0.053	4.3->4.2	461 y
	Epitopes	0.4382	0.023	1.9->1.8	223 y

See HIV-1 Clade B population data set table for an explanation of the columns. Removed proteins with less than 50 samples. Statistical test: Kendall Tau rank correlation test, with p-values < 0.001 in bold face.

Table 2.7 Details HIV-1 Clade A1 population data set.

Protein (# samples)		*P*	*density per aa*	*2008→2032*	*half-life*
ENV (51)					
	Precursors	**0.0002**	0.303	259.1->240.7	169 y
	MHC-binders	0.4856	0.036	30.8->30.4	951 y
	Epitopes	0.3581	0.018	15.5->15.1	470 y
GAG (118)					
	Precursors	0.9002	0.233	115.5->117.2	
	MHC-binders	**0.0026**	0.030	14.7->15.4	
	Epitopes	0.0156	0.010	4.7->5.1	
NEF (87)					
	Precursors	0.5081	0.254	52.6->50.4	291 y
	MHC-binders	0.3472	0.030	6.3->6.3	
	Epitopes	0.6436	0.015	3.0->3.0	
POL (54)					
	Precursors	0.1073	0.254	254.1->257.0	
	MHC-binders	0.0852	0.033	33.0->33.3	
	Epitopes	0.0242	0.012	12.2->12.5	
REV (58)					
	Precursors	0.4442	0.183	21.6->20.3	195 y
	MHC-binders	0.3422	0.024	2.9->2.9	2656 y
	Epitopes	0.8804	0.005	0.6->0.7	
TAT (55)					
	Precursors	0.9132	0.191	16.6->15.7	212 y
	MHC-binders	0.9172	0.018	1.6->1.6	440 y
	Epitopes	**0.0062**	0.003	0.3->-0.0	11 y
VIF (62)					
	Precursors	0.0213	0.265	50.8->47.8	202 y
	MHC-binders	0.1021	0.035	6.7->6.9	
	Epitopes	0.7905	0.013	2.4->2.5	
VPR (59)					
	Precursors	0.1106	0.294	28.2->29.2	
	MHC-binders	0.1236	0.033	3.1->3.0	257 y
	Epitopes	0.8780	0.015	1.5->1.5	
VPU (66)					
	Precursors	0.1275	0.342	28.0->28.4	
	MHC-binders	**0.0046**	0.050	4.1->4.8	
	Epitopes	**0.0011**	0.025	2.0->2.5	

See HIV-1 Clade B population data set table for an explanation of the columns. Removed proteins with less than 50 samples. Statistical test: Kendall Tau rank correlation test, with p-values < 0.001 in bold face.

The distribution of CTL epitopes in HIV-1 appears to be random, and similar to that of other proteomes.

3

Boris V. Schmid[1], Can Keşmir[1,2], Rob J. De Boer[1]

1 Theoretical Biology, Utrecht University, The Netherlands
2 Academic Biomedical Centre, Utrecht University, The Netherlands

Manuscript accepted by BMC Evolutionary Biology

3.1 ABSTRACT

Background: HIV-1 viruses are highly capable of mutating their proteins to escape the presentation of CTL epitopes in their current host. Upon transmission to another host, some escape mutations revert, but other remain stable in the virus sequence for at least several years. Depending on the rate of accumulation and reversion of escape mutations, HIV-1 could reach a high level of adaptation to the human population. Yusim et al. (2002) hypothesized that the apparent clustering of CTL epitopes in the conserved regions of HIV-1 proteins could be an evolutionary signature left by large-scale adaptation of HIV-1 to its human/simian host. **Results:** In this chapter we quantified the distribution of CTL epitopes in HIV-1 and found that that in 99% of the HIV-1 protein sequences, the epitope distribution was indistinguishable from random. Similar percentages were found for HCV, Influenza and for three eukaryote proteomes (Human, Drosophila, Yeast). **Conclusions:** We conclude that CTL epitopes in HIV-1 are randomly distributed, and that this distribution is similar to the distribution of CTL epitopes in proteins from other proteomes. Therefore, the visually apparent clustering of CTL epitopes in epitope maps should not be interpreted as a signature of a past large-scale adaptation of HIV-1 to the human cellular immune response.

3.2 BACKGROUND

The human immunodeficiency virus 1 (HIV-1) is a highly adaptive virus, capable of rapidly evolving its proteins to escape cellular immune responses and antiretroviral drugs (reviewed in Walker and Burton (2008) and Chen and Aldrovandi (2008)). This ability of the virus to rapidly adapt to its host has raised the question what level of adaptation to the whole human population the virus will

eventually be able to reach. Currently there is no consensus on this point: on the one hand there are studies that indicate that the current HIV-1 sequences contain signatures of global adaptation (Moore et al., 2002; Leslie et al., 2005; Bhattacharya et al., 2007; Poon et al., 2007; Brumme et al., 2007; Kawashima et al., 2009), while on the other hand the virulence of the virus (Herbeck et al., 2008; Müller et al., 2006) as well as its predicted number of cytotoxic T cell (CTL) epitopes have remained constant over time (Chapter 2).

An alternative way to study viral adaptation would be to look for tell-tale signatures of accumulated escape mutations in the virus. Yusim et al. (2002) suggested that the clustering of CTL epitopes is such a signature. They observed that regions in the virus with a low density of CTL epitopes were more variable than regions with high epitope density. Moreover, these variable regions had a lower level of epitope precursors than the conserved regions, and contained fewer amino acids that were suitable to serve as anchor residues for MHC binding. This led to the hypothesis that HIV-1 had escaped CTL epitopes predominantly in the variable protein regions, and that large-scale adaptation of the ancestral HIV-1 sequence to the human (or prior to that to the chimpanzee) host resulted in the observed clustering of CTL epitopes in current-day HIV-1 sequences.

Another, more proximate hypothesis for the clustering of epitopes was forwarded by Lucchiari-Hartz et al. (2003). Based on the analysis of proteasomal degradation products in HIV-1, they showed that the epitope precursors (and thus epitopes) occur preferentially in the more hydrophobic regions of HIV-1 NEF and RT proteins. They concluded that the clustering of epitopes is a generic feature of proteins, depending on the clustering of hydrophobic amino acids.

In this chapter we tested whether CTL epitopes and hydrophobic amino acids in HIV-1 are significantly clustered, and compared the distribution of epitopes in HIV-1 and other viruses to that of eukaryotes which are not under selection pressure to escape the cellular immune response. We discovered that for all tested protein sequences more than 95% of the epitope distributions, and more than 98% of the hydrophobic amino acid distributions were likely to be random distributions. Secondly, we discovered that there is a large amount of variation in the epitope distribution within HIV-1 proteins, similar to the amount of variation observed in eukaryote proteins of an equal length. Both findings suggest that the distribution of CTL epitopes in HIV-1 is similar to that of other proteins, and that the apparent clustering of CTL epitopes on HIV-1 epitope maps should not be interpreted as an indicator of past HIV-1 adaptation.

3.3 METHODS

3.3.1 *CTL epitope predictions*

There are several algorithms available that can predict the location and binding specificity of CTL epitopes in protein sequences (Hakenberg et al., 2003; Tenzer

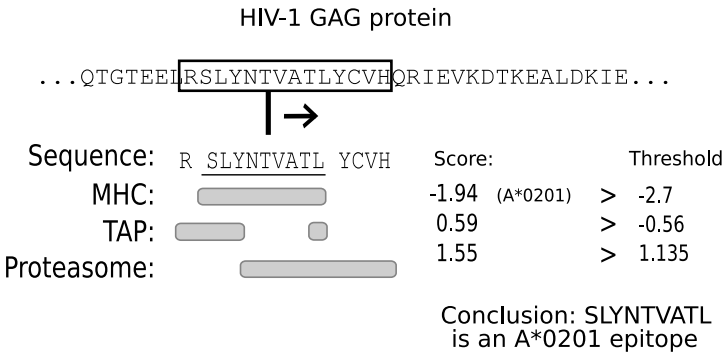

Figure 3.1 Schematic of the MHC-pathway model. A window of 14 amino acids is slided across a protein. Each of these '14mers' consist of a N-terminal flanking region of 1 amino acid, a 9mer epitope candidate and a C-terminal flanking region of 4 amino acids. Beneath the 14mer the parts of the peptide are marked which are used by the MHC, TAP or proteasome predictors. Applying the 14mer to the MHC, TAP and proteasome predictors results in three different scores. Only if each of these scores is higher than a fixed threshold, the epitope candidate is predicted to be a CTL epitope for the MHC allele under consideration.

et al., 2005; Larsen et al., 2005; Doytchinova et al., 2006). In this study we use the MHC-pathway model (Tenzer et al., 2005), which allows us to screen all possible peptide fragments of 14 amino acids within a particular protein for their ability to be correctly processed by the proteasome and transporter associated with antigen processing (TAP), and presented by the MHC class I molecules. Peptide fragments that can be correctly processed by all three steps are subsequently marked as CTL epitopes (Fig. 3.1). An extensive analysis of the quality of these predictors can be found in the methods section of Chapter 2. A brief synopsis is that 81-97% of the predicted CTL epitopes are indeed CTL epitopes Schellens et al. (2008); Fortier et al. (2008).

The threshold values for the proteasome and TAP predictors (Fig. 3.1) were derived by applying the MHC-pathway model to a large bacterial protein data set and selecting threshold values which correspond to the estimated specificity of the proteasome (33%) and TAP (76%) (Burroughs et al., 2004). For the MHC-binding predictions we used the default threshold of -2.7, which corresponds to an IC50 threshold of 500nM (Peters et al., 2006; Assarsson et al., 2007). The predictors used in this paper are available through a web interface (http://www.iedb.org 2006-01-01 version), and consist of an immuno-proteasome cleavage predictor, a TAP transport predictor and 34 different MHC class I alleles binding predictors (18 for human leukocyte antigen (HLA)-A alleles and 14 for HLA-B alleles). Based on our previous work with these predictors in Chapter 2, we excluded the A*3002 and the B*0801 MHC class I binding predictors for being too unspecific and specific, respectively.

3.3.2 *Describing epitope clusters*

CTL epitopes are traditionally defined by their amino acid sequence, the start and end-point of the sequence mapped onto a reference sequence, and the MHC class I allele that they bind to. Based on this definition, a natural way to visually present CTL epitopes is an epitope map (Fig. 3.3). However, for the statistical study of the clustering of CTL epitopes it is more practical to reduce the position of a particular epitope to a single point. In this chapter we have opted to use the C-terminal amino acid of CTL epitopes to position an epitope in a sequence, as the C-terminal amino acid is the most defining property of an epitope: amino acid substitutions at the C-terminal have a large effect on proteasomal cleavage, TAP transportation and MHC binding (Craiu et al., 1997; Cascio et al., 2001) (Fig. 3.1). One effect of this transformation is that epitopes are only defined by their position, and no longer by the MHC allele(s) that they bind. Thus, epitopes that only differ in the MHC that they bind to will be reported as a single epitope. Although this transformation makes it possible to perform a clustering analysis on the distribution of CTL epitopes, it might destroy an evolutionary signature that is contained in the number of MHC alleles that bind to individual epitope precursors. We discuss this in the final section of the paper.

3.3.3 *Clustering methods*

Methods to describe the degree of clustering of sequential events or spatial locations have been developed in a large number of scientific fields, ranging from astronomy to ecology and economics. These methods consider two features of a clustering: the 'intensity' and the 'grain'. The intensity reflects the difference in object density between the rich and poor regions, and the grain describes how frequently rich and poor regions alternate (Pielou, 1977). In this chapter we will use two methods: the cumulative binominal probability (CBP) (Wilk and Gnanadesikan, 1968), and the Hopkins and Skellam index (H&S) (Hopkins and Skellam, 1954; Pielou, 1977).

Regarding the cumulative binominal probability method (CBP) (Wilk and Gnanadesikan, 1968): this method can be used to determine whether a particular amino acid lies in a region that is rich, poor or neutral in epitope density (Eq. 3.1). For example, to determine whether a particular amino acid is located in an epitope-rich region, one counts the number of epitope C-terminals (e) in a window of size w, and based on the average epitope-density in the protein (f) one calculates the chance of finding e or more epitope C-terminals in a window of w amino acids ($i = e, i = e + 1, i = .., i = w$). If this chance P falls below a certain threshold (0.05 in this paper), all amino acids in that window are marked as belonging to an epitope-rich region. The same approach can be used to determine which amino acids belong to epitope-poor regions. The CBP method makes it possible to objectively determine the location of epitope-rich and epitope-poor regions in a protein. These locations can be plotted to generate a CBP profile (see Fig. 3.4).

$$P = \sum_{i=e}^{i=w} \binom{w}{i} * f^i * (1-f)^{w-i} \tag{3.1}$$

Regarding the Hopkins and Skellam index (H&S) (Hopkins and Skellam, 1954; Pielou, 1977): this method is based on the observation that in a fully random distribution (of an infinite size) the distance from a starting point to the nearest object of interest is not influenced by the presence or absence of such an object at the starting point. In an overdispersed distribution, the presence of an object at the starting point will mean that the nearest object is on average further away than when starting at a random location, while in a clustered distribution, the reverse is true. The ratio is calculated as the sum of squared distances from a random point to the nearest object (d_r) to the sum of squared distances from a random object to the nearest object (d_o). When the number of d_r and d_o measurements are not equal, the sum of squared distance of d_r and d_o should be divided by the number of d_r measurements (n) and d_o measurements (m), respectively. The ratio will be a number (R) between 0 for perfectly overdispersed distributions and infinity for fully clustered distributions (Eq. 3.2). In this way the distribution of epitopes within a protein can be characterized by a single ratio. In this chapter we normalized the range of the H&S index in such a way that the index runs from 0 to 2, rather than from 0 to infinity, by translating any score above one to $2 - (1/score)$.

Both the CBP method and the H&S index take into account the 'intensity' and the 'grain' of epitope distributions and correct for the epitope density of the protein. One difference between the two methods is that the latter gives a higher clustering score to coarse grained distributions, whereas the former favors fine-grained distributions (see section 3.4.3).

$$R = \frac{\sum_{i=0}^{i=n} {d_r}^2/n}{\sum_{i=0}^{i=m} {d_o}^2/m} \tag{3.2}$$

3.3.4 *Statistical testing*

The significance of both clustering measures can be tested with permutation tests (Fisher, 1935; Box, 1980; Ludbrook and Dudley, 1998). Permutations are created by randomizing the positions of the epitope C-terminals in the protein that is under scrutiny. The p-value of the test is the fraction of cases in which the randomized sequence has an equal or more extreme outcome than the original sequence. In the case of the CBP method the outcome was measured as the fraction of the protein that is part of an epitope-rich region (or epitope-poor region). In the case of the H&S index, the outcome was measured as the ab-

solute difference of the index score from 1.0 (the expected score for a random distribution).

3.3.5 *Hydrophobicity*

In order to determine whether hydrophobicity is clustered, we calculated the clustering of the top 4 hydrophobic amino acids (Leu, Ile, Phe and Trp according to both the HPLC pH 7.4 scale (Meek, 1980) used by Lucchiari-Hartz et al. (2003), and the consensus scale (Tossi et al., 2002)), with the CBP method and the H&S index. This is somewhat different from the more common approach of calculating the running average hydrophobicity and setting one or two thresholds to determine the hydrophobic and hydrophilic areas of a protein, (as was done in the study of Lucchiari-Hartz et al. (2003)). However, it has the advantage that we can use the same method for determining epitope clustering and hydrophobic amino acid clustering.

3.3.6 *Data sets*

The public data sets used in this paper originate from a variety of sources. Prealigned HIV-1 and HCV data (size: 13093 and 8886 proteins, respectively) were downloaded from the Los Alamos laboratories (www.hiv.lanl.gov, www.hcv.lanl.gov), and the influenza data set (size: 47194 proteins) was downloaded from Biohealthbase (www.biohealthbase.org, under Influenza Virus, Database Search, Sequence), by selecting for all available proteins from human influenza type A, B or C. A Human proteome(Kersey et al., 2004) (IPI.human.prot, size: 72082 unique proteins), a Drosophila (size: 23694 unique proteins) and a Yeast proteome (size: 5863 unique proteins) were downloaded from Integrate (http://www.ebi.ac.uk/integr8/). All data sets were downloaded on 13 Aug 2008.

The public HIV-1 and HCV data sets are already curated, and do not contain multiple clones from one isolate, or multiple sequences from a single person. Furthermore, very similar groups of sequences (based on phylogenetic tree analysis) are also reduced to a single sequence. In all three eukaryote proteomes, only unique protein sequences are used.

The ancestral HIV-1 clade B sequence (Korber et al., 2000) can be downloaded at http://www.hiv.lanl.gov.

3.4 RESULTS AND DISCUSSION

3.4.1 *Imprints of immune evasion in HIV-1*

HIV-1 is capable of maintaining escape mutations to CTL epitopes in the absence of immune selection pressure of a MHC-matched host (Goulder et al., 2001; Kearney et al., 2009), and thus escape variants of HIV-1 can become the consensus HIV-1 sequence (Leslie et al., 2005; Moore et al., 2002). Escape vari-

ants with a low fitness cost, or having compensatory mutations revert slowly or not at all (Schneidewind et al., 2009; Kawashima et al., 2009), and could quickly accumulate in the virus (Leslie et al., 2005; Kearney et al., 2009). Fig. 3.2A sketches the fast spread of a non-reverting escape variant in a hypothetical transmission network. Even though the hosts which carry an MHC allele that can bind to the CTL epitope are not optimally positioned in the transmission network, it only takes a few transmissions before the majority (54%) of the hosts carries the escape variant of the virus.

Yusim et al. (2002) studied the apparent clustering of CTL epitopes in HIV-1 epitope maps, and found a negative correlation between CTL epitope density and sequence variability in HIV-1. Based on the paucity of epitope precursors and suitable MHC anchor residues in these variable protein regions, Yusim et al. (2002) concluded that the lack of epitopes in the variable regions was a signature of immune evasion of the virus. The conserved protein regions were assumed have more constraints related to protein function, and the virus would have fewer viable options to generate escape variants in these regions(Yusim et al., 2002), because the escapes made in these conserved regions would carry a higher fitness cost (Wagner et al., 1999; Walker and Korber, 2001). As a result, Yusim et al. (2002) argued that the accumulation of escape mutations would be slower, and reversion of escape mutations faster in conserved protein regions than in variable regions. These ideas are depicted in Fig. 3.2B. This difference in the rate of accumulation of escape mutations between the variable and conserved protein regions is expected to result in a clustering of CTL epitopes once the virus has accumulated a substantial number of escape mutations. Taken together, Yusim et al. (2002) concluded that the apparent clustering of CTL epitopes in epitope maps was a signature of a large-scale adaptation of HIV-1 to the human population.

3.4.2 *Clustering in epitope maps*

The first reports that the CTL epitopes of HIV-1 occurred in clusters (Culmann et al., 1991; Culmann-Penciolelli et al., 1994; Walker and Korber, 2001) were published a few years after the discovery of HIV-1 CTL epitopes themselves (Walker et al., 1987; Plata et al., 1987). However, the degree of clustering of CTL epitopes has never been tested rigorously, perhaps because the method by which epitope positions are visualized in epitope maps strongly suggests that a clustering exists (Fig. 3.3). Here we list a number of reasons why epitope maps may give an unjust impression of clustering:

1. The epitope map is a compilation of the CTL epitopes found in a large number of sequences. Amino acid variants of the same epitope are all depicted at the same position of the reference sequence, but never occur simultaneously in a single HIV-1 sequence.

2. CTL epitopes that have not been mapped precisely to their minimal length

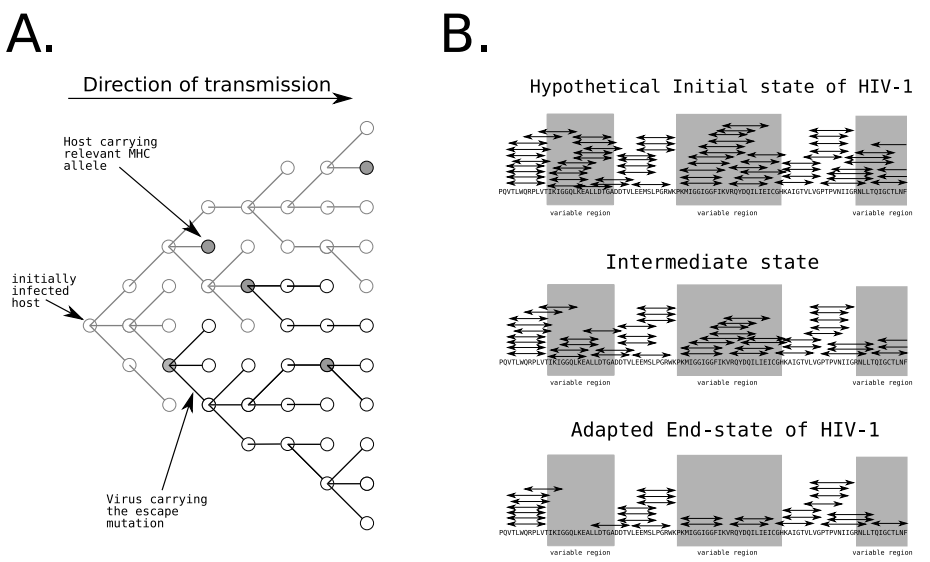

Figure 3.2 A. Schematic of a transmission network of HIV-1 through the human population. An escape variant of a particular CTL epitope can rapidly become the consensus sequence if reversion of the escape happens little or not at all. Of the 50 hosts (circles), only 5 hosts (filled circles) carry an MHC allele that can bind the epitope. Grey lines represent transmission of the wildtype virus, whereas black lines represent transmission of the escape variant. **B. Schematic of the accumulation of escapes in the variable protein regions** When escape mutations occur more often, or reversion happen more slowly in variable protein regions (gray shaded areas), and the number of accumulated escape mutations is large enough, a clustering of CTL epitopes (plotted as arrow-delimited lines) is to be expected. One underlying assumption is that the variable and conserved protein regions are larger than a few amino acids in size.

can end up occurring more than once on the epitope maps as N- or C-terminal extended versions of an epitope.

3. Epitope precursors are expected to be generated at roughly 25% of the positions in a protein (Burroughs et al., 2004). The large polymorphism in MHC class I alleles makes it likely that a single epitope precursor binds to multiple MHC alleles (Frahm et al., 2007). Therefore, the absence or presence of an epitope precursor at a certain position results in either zero or many epitopes reported at that position.

4. CTL epitopes on the maps are vertically ordered to be non-overlapping. This representation results in empty corridors between large slanted towers of epitopes. The corridors need not correspond to epitope-poor regions, but are a visual effect of the representation.

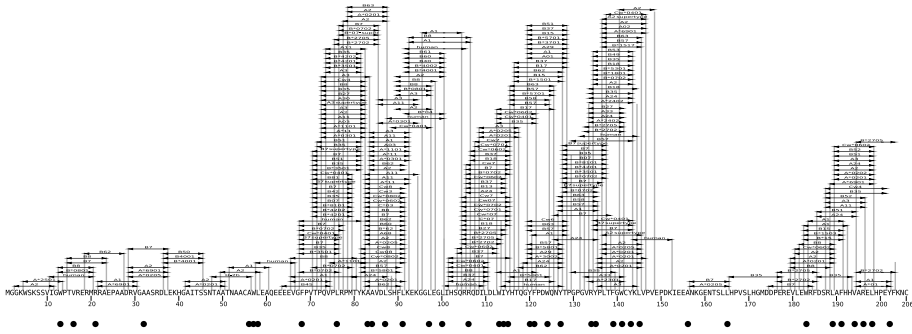

Figure 3.3 Example of an epitope map of HIV-1 NEF and its transformed version (black dots). The epitope map was retrieved from http://www.hiv.lanl.gov (Jul 31 2008 version). This specific epitope sequence (an unspecified HXB2 variant) carries 52 predicted epitope precursors, of which 38 are predicted to be epitopes (binding at least one of the 32 MHC class I allele binding predictors).

All four reasons listed amplify the difference between epitope-rich and epitope-poor regions with the result that a strong clustering of CTL epitopes appears to exist. Most of these reasons have already been listed by the scientists maintaining the epitope maps at www.hiv.lanl.gov, but the suggestive effect on epitope clustering is not mentioned explicitly. We removed these amplifications by using CTL epitope predictors (removing 2.) on a per-sequence basis (removing 1.), and by only plotting their C-terminal position (remove 3. and 4.). This transformation results in a binary pattern of epitope C-terminal positions: each position being either an epitope (of single or many MHC alleles) or a non-epitope (see Fig. 3.3, Methods). Although neccesary for a meaningful analysis of the clustering of CTL epitopes, the transformation could potentially destroy an evolutionary signal if HIV-1 has evolved to have fewer MHC alleles binding per epitope precursors in certain protein regions. We will consider this option in the discussion.

We use epitope predictors rather than lists of known CTL epitopes, as the predicted epitopes are less influenced by an 'attention bias' than experimentally defined CTL epitopes. The bias is caused by of researchers focusing on hot topics or building on previous work. Assarsson et al. (2008) showed that as a result of such a bias, certain protein regions in Influenza are mistakenly classified as CTL epitope-rich or poor.

Epitope predictors, such as the MHC-pathway algorithm, predict proteasomal cleavage, transporter associated with antigen processing (TAP) and MHC class I binding (Tenzer et al., 2005) for all peptide fragments within a protein. Those fragments that can be processed by each of these steps are predicted to be CTL epitopes (see Fig. 3.1). As the MHC-pathway algorithm has been tested extensively (Peters et al., 2006) and has proven to have a high reliability (Schellens

et al., 2008; Fortier et al., 2008) (see Methods), it allowed us to avoid any possible 'attention bias' (Assarsson et al., 2008) for certain HIV-1 protein regions or strains.

3.4.3 *No clustering of epitopes*

We applied two distinct methods of measuring distributions to the epitope distribution of HIV-1 proteins. The first method divides proteins into epitope-rich, epitope-poor, and neutral regions based on the cumulative binominal probability (CBP) (Wilk and Gnanadesikan, 1968) of having e or more amino acids predicted as an epitope C-terminal in a window of size w (Eq. 3.1). The second method is the Hopkins and Skellam (H&S) index (Hopkins and Skellam, 1954; Pielou, 1977), which compares the average distance from an epitope to its nearest epitope with the average distance from a random amino acid to the nearest epitope within proteins (Eq. 3.2). Both methods are subjected to permutation tests in order to establish per protein the significance of its distribution of CTL epitopes. A more extensive discussion of these methods and the permutation testing is available in the Methods section.

Using both the CBP method and the H&S index, we find protein sequences in HIV-1 with CTL epitope distributions that are likely to be random, as well as distributions that are likely to be clustered. We visualized a few of these protein sequences using CBP profiles (Fig. 3.4), as well as a sequence in which the positions of the CTL epitopes were randomized. Note that each of the four visualized sequences, including the randomized one, contain epitope-rich and/or epitope-poor regions. Thus, the mere presence of epitope-poor regions in a protein does not indicate that some active process created it.

We analyzed the predicted CTL epitope distribution in a data set of 11017 HIV-1 proteins from the Los Alamos HIV-1 Sequence compendium with the CBP method, and found that in 99% of these sequences the fraction of epitope-rich regions was not significantly different from random ($p < 0.001$, permutation test). Only 158 sequences had a larger fraction of epitope-rich regions than likely to arise in a random distribution of CTL epitopes. These 158 sequences predominantly occurred in two specific HIV-1 proteins: HIV-1 VPU (79x) and HIV-1 ENV (74x). Changing the window size w from 15 to 9 or 23 shifted the number of significant sequences towards VPU or ENV, respectively, but did not affect the overall lack of significant clustering of CTL epitopes. Similar to what we found for the epitope-rich regions, only 153 sequences in HIV-1 had a larger fraction of epitope-poor regions than expected, most of which occurred in VPU (135x).

The H&S index gives a similar result as the CBP method: only 68 HIV-1 protein sequences (0.6%) had a predicted epitope distribution that is significantly more clustered than expected from a random distribution, and most of these occurred in the VPU protein (55x). The distribution of CTL epitopes in the predicted ancestral HIV-1 clade B sequences (Korber et al., 2000) (green dots, Fig. 3.5B) is also not significantly different from random. The fact that

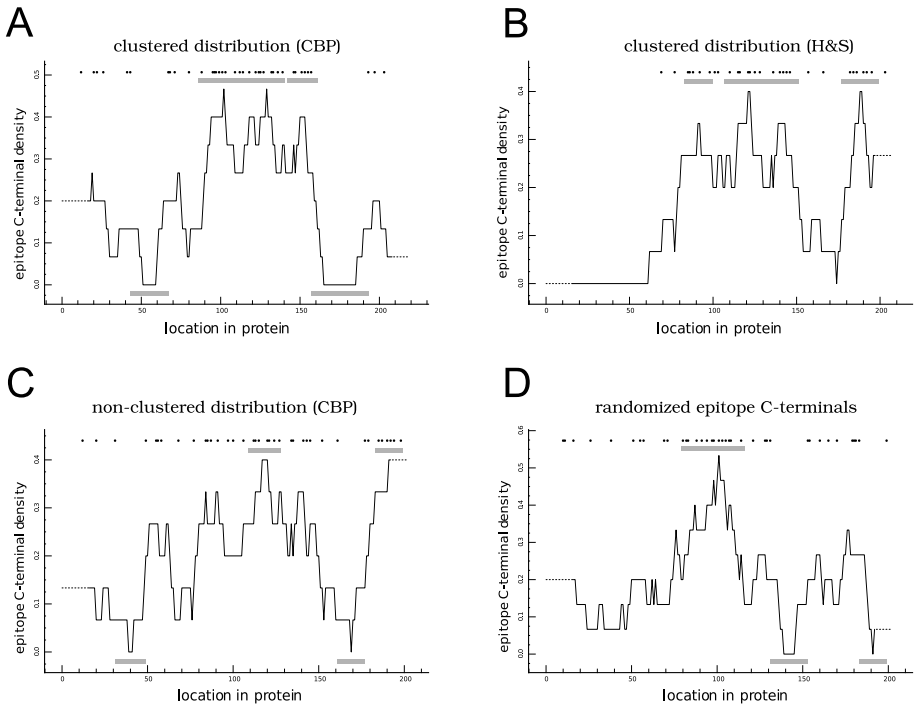

Figure 3.4 Example of the cumulative binominal probability (CBP) profiles of 4 HIV-1 NEF protein sequences. Each profile features the position of predicted epitope C-terminals in the protein sequence (black dots), a running average of the C-terminal density (black line, window size 15), and the epitope-rich and/or epitope-poor regions (grey blocks) (see Eq. 3.1). **(A + B)** The distribution of CTL epitopes in the protein sequence in panel A (accession number:DQ351225), and the protein sequence in panel B (accession number:AJ233029) have a low probability to arise from a random distribution of CTL epitopes (Panel A, CBP: p_{rich} = 0.0056, p_{poor} = 0.0062; H&S: p = 0.015. Panel B, CBP: p_{rich} = 0.06, p_{poor} not computable for a window size of 15 or smaller (see Methods); H&S: $p < 0.001$.). **(C)** The distribution of CTL epitopes in the protein sequence in panel C (accession number:AY905390) has a high probability to arise from a random distribution. CBP: p_{rich} = 0.7026, p_{poor} = 0.5328; H&S: p = 0.195. **(D)** The same sequence as in Panel C, but with the epitope C-terminals randomized.

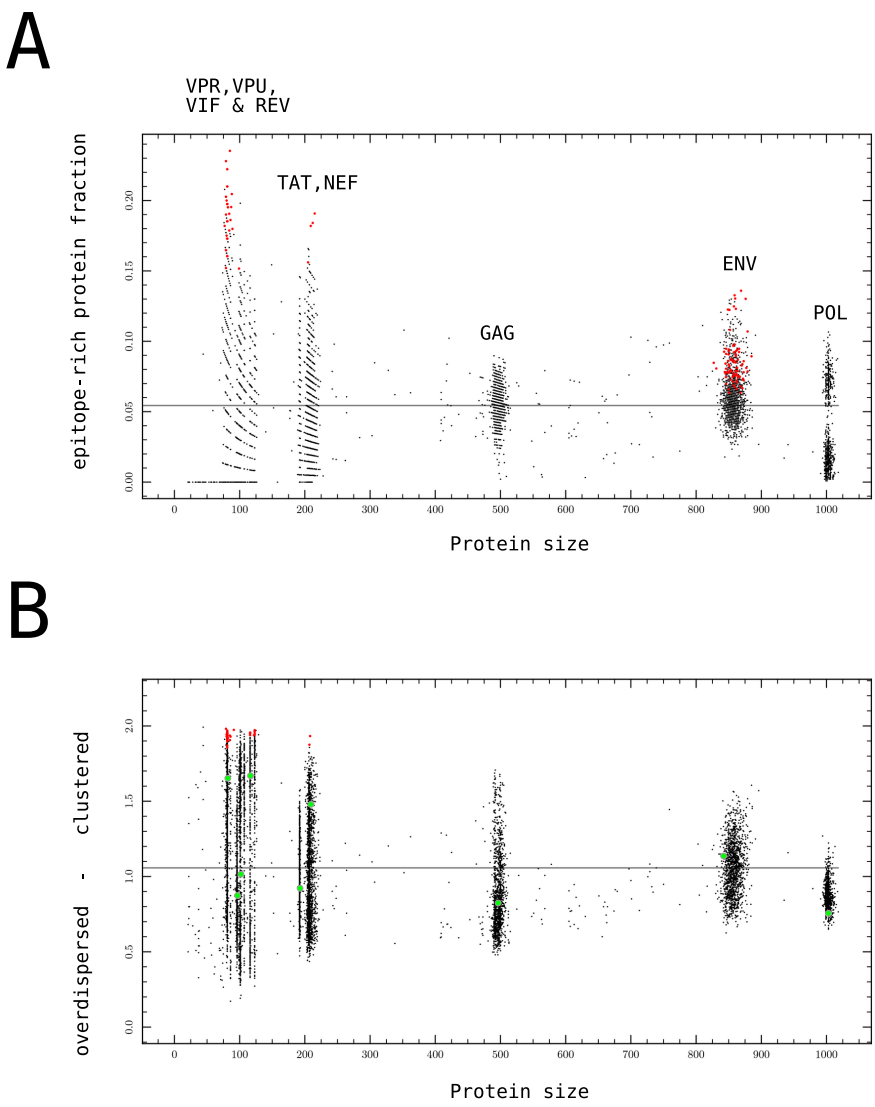

Figure 3.5 Analysis of HIV proteomes. Each protein sequence is plotted on the horizontal axis according to its size (dots). Red dots are significant ($p < 0.001$). Above each large cloud of sequences, the corresponding protein name is denoted. Sequences in between clouds are likely to be truncated version of larger proteins that were not pruned from the data set. The grey line denotes the average score. **(A)** The CBP method yielded only 153 out of 11017 sequences whose epitope distribution was unlikely to be random (less than 2%). **(B)** The H&S index yielded 68 sequences unlikely to be random (less than 1%). Marked with green dots are the predicted ancestral sequence of HIV-1 clade B proteins.

we found HIV-1 ENV and VPU to be the proteins in which some sign of clustering occurred would not be what one expects if the clustering would be due to adaptation. The GAG protein would have been a more obvious candidate, both for its early presentation on the cell surface (Sacha et al., 2007a) and its immunodominant CTL epitopes (Kiepiela et al., 2007).

There is a remarkable large variation between sequences of the same protein in both the CBP and the H&S methods. Sequences range from being devoid of epitope-rich protein regions to having 20% of their amino acids belonging to an epitope-rich region in the CBP method (Fig. 3.5A). The same holds for epitope-poor regions (data not shown). Sequences range from a highly clustered to a moderately overdispersed H&S index score (Fig. 3.5B), even when reducing the data set to specific HIV-1 clades (data not shown). Apparently, even within the closely related sequences, relatively small amino acid differences can cause large variations in the degree of CTL epitope clustering, up to a degree that their H&S index score variation for most of HIV-1's proteins is similar to that of randomized proteins of the same size. The POL and TAT proteins displayed less variation than expected for their protein size.

When comparing the significantly clustered sequences predicted by both methods, we find an overlap of 21%. While this is significantly higher than the expected overlap of 1.4% (Fisher's exact test, $p = 0.0008$), the difference in outcome between the methods utilized is substantial. This could be due to the difference in how both methods valuate the 'grain' of a pattern (i.e. how frequently rich and poor regions alternate). The H&S index valuates coarse-grained clusters (Fig. 3.5) above fine grained clusters, whereas the CBP method does the opposite.

3.4.4 *Comparison between species*

Although CTL epitopes in HIV-1 are typically randomly distributed (Fig. 3.5), a direct comparison of the CTL epitope distributions between virus and eukaryote proteomes might reveal a difference between both groups that is due to immune selection pressure. We included two additional virus sequence data sets in the analysis, namely the Hepatitis C Virus (HCV) and Influenza, and picked three eukaryote proteomes: the human *Homo sapiens*, fruitfly *Drosophila melanogaster* (Leulier et al., 2003) and yeast *Saccharomyces cerevisiae* proteome. The latter two are proteomes that normally do not come into contact with the human antigen presentation pathway, and should therefore not be adapted to it.

The distribution of predicted CTL epitopes in HCV and Influenza was similar to that of HIV-1. The vast majority of sequences featured a random distribution of CTL epitopes ($> 99\%$), and the equally large amount of variation in H&S clustering score as seen in HIV-1 (Fig. 3.6A). Although in all three viruses some proteins tended towards clustering, and others towards overdispersion of epitopes, we have not been able to detect a pattern in these tendencies. One difference between the viruses was that the small fraction of significantly clustered sequences was somewhat higher in HIV-1 (0.6%) than in HCV (0.01%) and In-

fluenza (0.00%), but as we are comparing many related copies of only a small number of proteins, this difference could well be due to chance.

A comparison of the three eukaryote proteomes revealed that their CTL epitope distributions are remarkably similar to each other. In all three proteomes there is a steady trend towards clustered epitope distributions with increasing protein size (Fig. 3.6B, grey line). It could be that these significant proteins contain more structural motifs and repeating elements that the other proteins, and that these motifs influence the epitope distribution (Irbäck et al., 1996; Landolt-Marticorena et al., 1993). The percentage of significantly clustered sequences (H&S) is a few percentage-points higher than in the viruses (Human: 1.9%, Drosophila: 3.4%, Yeast: 1.9%, at a $p < 0.001$, permutation test), but is still only a small percentage of all sequences.

An overlay of HIV-1 on the human proteome shows that the H&S clustering scores for HIV-1 proteins fall within the range of scores for human proteins (Fig. 3.7). The variation within HIV-1 proteins spans about the same range as proteins of comparable size in the human proteome. This is surprising, as the sequences within HIV-1 proteins are closely related to each other, and would therefore be expected to have a smaller range of clustering scores (for the POL and TAT protein this seems to be partially true).

Summarizing, the CTL epitopes of $> 99\%$ of HIV-1, HCV and Influenza sequences were found to be randomly distributed. A comparison between viral and eukaryote proteomes showed no qualitative differences in the epitope distribution between the two groups that would point towards the adaptation of viruses to the human host.

3.4.5 *No clustering of hydrophobicity*

An alternative hypothesis on epitope clustering that was forwarded by Lucchiari-Hartz et al. (2003), challenged the idea that the distribution reflected the adaptation of the HIV-1 to its new host (Yusim et al., 2002), Lucchiari-Hartz et al. (2003) suggested that the clustering of CTL epitopes merely mirrored the clustering of hydrophobic amino acids. As the proteasome, TAP, and many of the MHC alleles favor hydrophobic amino acids at or near their C-terminal end (Peters et al., 2003; Burroughs et al., 2004; Uebel and Tamp, 1999), a clustering of hydrophobic amino acids would result in a clustering of epitope precursors, and subsequently result in a clustering of CTL epitopes.

Our results thus far dispute the idea that CTL epitopes are clustered, as we found the epitope distribution in the vast majority ($> 99\%$) of protein sequences to be not different from a random distribution. Therefore we wondered if hydrophobic amino acids are truly clustered in proteins, and repeated our clustering analysis for hydrophobic amino acids. By taking the four most hydrophobic amino acids (Leu, Ile, Phe and Trp (Meek, 1980; Tossi et al., 2002)), we could construct binary maps similar to the transformed epitope maps of Fig. 3.3.

We found that nearly 100% of the protein sequences in HIV-1 had no significant clustering of hydrophobic amino acids in their primary structure (Fig.

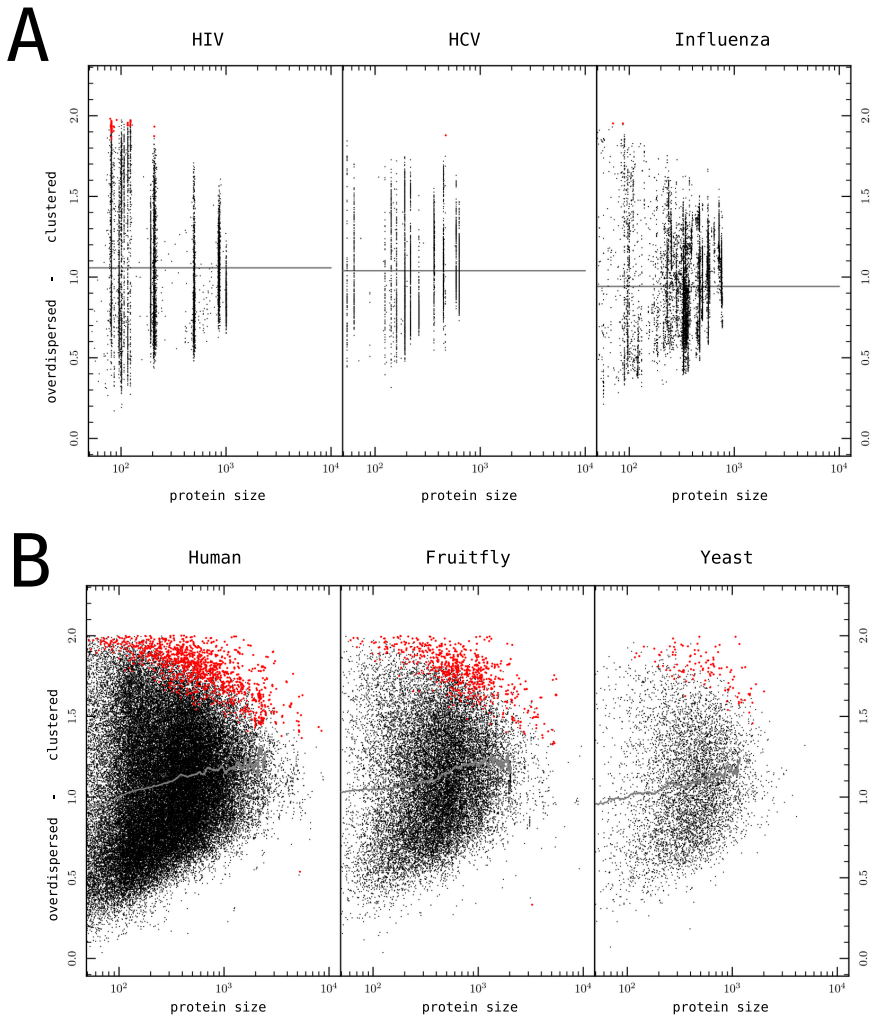

Figure 3.6 The H&S index for six proteomes, plotted against protein length. The top three panels display virus proteome data sets, and contain multiple sequences per protein. Each vertical cloud corresponds to another protein. The bottom three panels display the proteomes of Human, Drosophila and Yeast. **(A)** Grey line: average index score taken over all sequences. **(B)** Grey line: running average of the index score (window size of 70). **(A + B)** Red dots: protein sequences whose epitope distribution is significantly unlikely to be random ($p < 0.001$, permutation test).

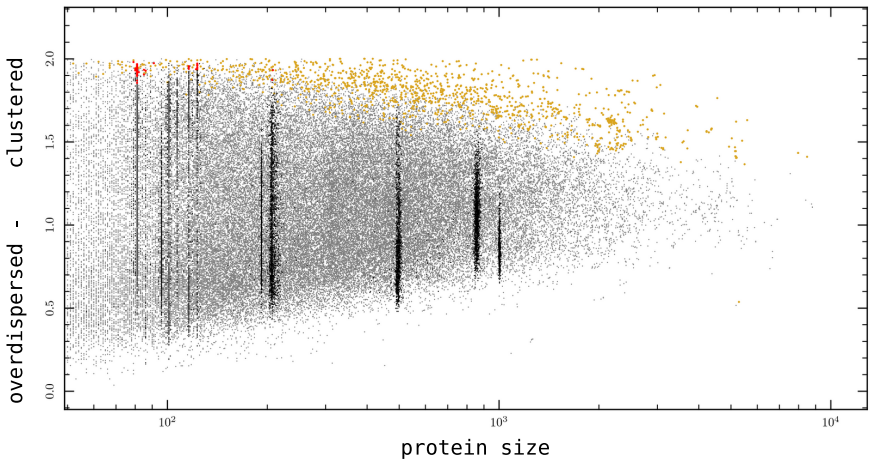

Figure 3.7 An overlay of HIV-1 proteins on top of the human proteome. The degree of clustering of CTL epitopes in proteins is determined by the H&S index and plotted against protein length. Significantly clustered sequences are denoted in yellow for the human proteome and red dots for HIV-1 ($p < 0.001$, permutation test). The index scores of HIV-1 fall within the range found in human proteomes, suggesting that the epitope distribution pattern of HIV-1 is not extraordinary.

3.8C). However, the biophysical community is somewhat divided on this point: depending on the method used and the subset of proteins studied, both random (White and Jacobs, 1990) and non-random distributions (Pande et al., 1994; Irbäck et al., 1996) are reported. We agree that some signs of non-randomness is to be expected in the distribution of amino acids in proteins, as common protein structures like α helices, β sheets have a certain periodicity in their use of hydrophobic amino acids (Irbäck et al., 1996). However, because we find so few proteins in which hydrophobic amino acids are significantly clustered, it seems safe to conclude that the effect of protein structure on the distribution of hydrophobic amino acids is rather subtle.

As was shown previously by Lucchiari-Hartz et al. (2003), hydrophobic amino acids and the location of epitope C-terminals in HIV-1 correlate. This is visible in CBP profiles (Fig. 3.8A, Fig. 3.8B, black and grey lines), and statistically confirmed in the overlap between sequences with significantly clustered epitope distributions and hydrophobic amino acid distributions in the human proteome (Fischer's exact test, $p < 0.0001$, $n = 69685$). Summarizing, we find that epitope-poor regions correlate with hydrophilic regions, but neither has a distribution that is significantly different from random.

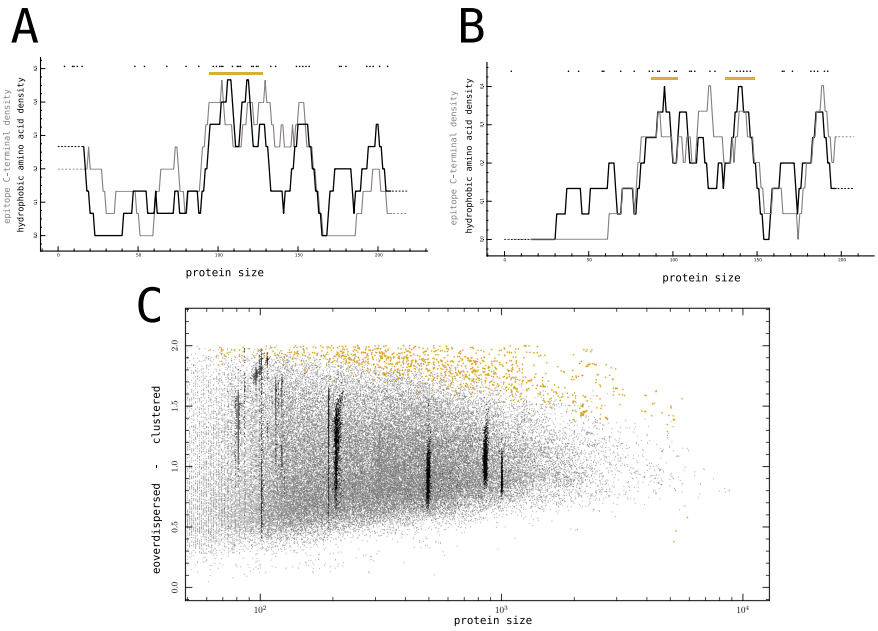

Figure 3.8 (A + B) CBP profiles of the hydrophobic amino acids (Leu, Ile, Phe and Trp) of the same two HIV-1 NEF sequences as profiled in Fig. 3.4A and Fig. 3.4B. Hydrophobic amino acids (black dots) and hydrophobic areas (orange blocks) are depicted above the running average (window size 15) of hydrophobic amino acid density (black line) and of the epitope C-terminal density (grey line). **(A)** Both the fraction of the sequence that is part of a hydrophobic region (14%), and the H&S index score (1.48) are likely to occur at random (CBP: $p_{rich} = 0.099$, H&S: $p = 0.095$). **(B)** Both the fraction of the sequence that is part of a hydrophobic region (15%), and the H&S index score (0.758) are likely to occur at random (CBP: $p_{rich} = 0.047$, H&S: $p = 0.576$). **(C)** An overlay of HIV-1 proteins on top of the human proteome. The degree of clustering of hydrophobic amino acids is determined by the H&S index and plotted against protein length. Significant sequences (i.e. $p < 0.001$, permutation test) are plotted as red dots for HIV-1 (only 5 out of 11039), and yellow dots for the human proteome (1195 out of 70269).

3.5 CONCLUSION

We showed that the vast majority (>99%) of HIV-1, HCV and Influenza proteins has a predicted CTL epitope distribution that is indistinguishable from a random distribution (Fig. 3.5). Additionally, the distribution of hydrophobic amino acids in these proteins is also likely to be random (Fig. 3.8C). These findings cast doubt on two recent hypothesis in which it was argued that the clustering of CTL epitopes in HIV-1 proteins is the product of virus adaptation (Yusim et al., 2002), or the result of clustered hydrophobic amino acids (Lucchiari-Hartz et al., 2003), respectively.

To further investigate if there was any sign of evolution in the distribution of CTL epitopes in viruses, we compared three virus proteomes to those of Human, Drosophila and Yeast (Fig. 3.6). The epitope distribution in HIV-1 proteins, as measured by the Hopkins & Skellam index score, is not extraordinary and falls within the range of proteins of comparable size in the human proteome (Fig. 3.7). Remarkably, the variation in epitope distribution that exists for any HIV-1 protein when sampling the virus from many hosts is as broad as the whole range of distributions found between all eukaryotic proteins of a comparable size as the sampled HIV-1 protein. Such a large amount of variation in epitope distributions is not what one would expect if HIV-1 has been undergoing large-scale adaptation to the human population. If HIV-1 had been globally accumulating the same CTL epitope escapes in its variable protein regions(Yusim et al., 2002), the distribution of CTL epitopes within HIV-1 viruses should be converging towards the same distribution.

The transformation that we applied to analyse the spatial distribution of CTL epitopes in proteins (discussed in section 4.2 and 5.2) could have destroyed one possible fingerprint of HIV-1 adaptation, namely that in the variable regions HIV-1 has adapted to select for epitope precursor to which only a limited number of MHC alleles can bind (Yusim et al., 2002). A study of the distribution of CTL epitopes over epitope precursors in HIV-1 revealed that a larger fraction of epitope precursors is predicted not to bind to any of the 32 studied MHC alleles (33%), than expected for a random distribution. Furthermore, the number of epitope precursors that bind to 1, 2 or 3 MHC molecules is underrepresented, whereas the number of epitope precursors that bind to 4 or more MHC molecules is overrepresented in HIV-1. This pattern of under- and overrepresentation strongly suggests that the number of MHC alleles that can bind to a particular amino acid sequence is clustered. However, this pattern is not only observed for HIV-1 proteins, but also for HCV, Influenza, and the Human proteome (see Fig. 3.9), which suggests that the clustering of MHC alleles over epitope precursors reflects patterns in the binding preferences of MHC alleles, and not as much a fingerprint of HIV-1 adaptation to its human host.

Whether or not HIV-1 is currently adapting to the human population is debated in the literature, and investigated with the help of a variety of methods (Moore et al., 2002; Yusim et al., 2002; Leslie et al., 2005; Müller et al., 2006; Brumme et al., 2007; Bhattacharya et al., 2007; Poon et al., 2007; Herbeck et al.,

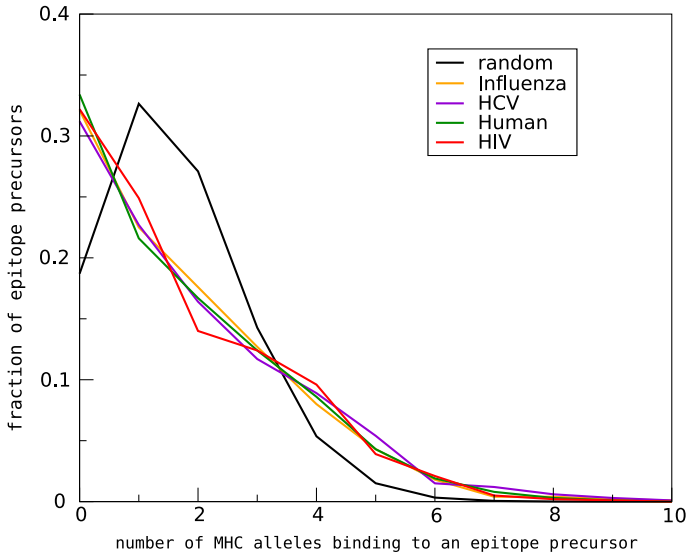

Figure 3.9 MHC alleles are clustered over epitope precursors. Multiple MHC alleles can bind to the same epitope precursor. Here we studied the distribution of MHC alleles over epitope precursors. With 32 different MHC matrices available for the MHC-pathway predictor, a single epitope precursors will be able to bind between 0 and 32 HLA molecules. In a situation where MHC alleles are randomly distributed over epitope precursors, each MHC allele has a chance to bind to an epitope precursor with the same chance as the specificity of that MHC allele (which for a threshold of IC50 500nM ranged from 0.5% to 13% of the epitope precursors). Given a random distribution, 18% of the epitope precursors are expected to bind to none of the available MHC alleles, 33% to bind a single MHC allele, and 49% to bind 2 or more MHC alleles (black line). In contrast with the random distribution, the predicted distribution for HIV-1 shows a higher percentage of epitope precursors that bind no epitope precursors, fewer than expected epitope precursors that bind between 1 and 3 MHC alleles, and more than expected epitope precursors that bind to 4 or more MHC alleles (red line). Not only HIV-1, but also HCV, Influenza, and the Human proteome follow this pattern.

2008; Kawashima et al., 2009) (Chapter 2). We reported in Chapter 2 that HIV-1 did not show any large-scale adaptation to the cellular immune response over the last three decades, using HIV-1 population sequence data sets and CTL epitope predictors. In this paper we show that the distribution of predicted CTL epitopes in HIV-1 appears to be random, and is similar to the distribution of CTL epitopes in organisms that are not under selection pressure to escape the human antigen presentation pathway. Therefore we conclude that the visually apparent clustering of CTL epitopes in epitope maps should not be interpreted as a signature of a past large-scale adaptation of HIV-1 to the human cellular immune response.

3.6 AUTHORS' CONTRIBUTIONS

BVS, CK and RJB all contributed to conception and design of the study. BVS analysed the data, and performed the statistical testing. BVS, CK and RJB all contributed to the interpretation of the results. BVS drafted the manuscript and created the figures. CK and RJB extensively commented on the drafts and figures. All authors read and approved the final manuscript.

3.7 ACKNOWLEDGEMENTS

The authors would like to thank the Netherlands Organisation for Scientific Research (NWO) for their financial support (VICI grant 016.048.603, Open Program 812.07.003).

Quantifying how MHC polymorphism prevents pathogens from adapting to the antigen presentation pathway

4

Boris V. Schmid, Rob J. De Boer

Theoretical Biology, Utrecht University, The Netherlands

manuscript submitted

4.1 ABSTRACT

Background: The classical antigen presentation pathway consists of two monomorphic (proteasome and TAP) and one polymorphic component (MHC Class I). Viruses are capable of escaping CTL responses by mutating an epitope such that it is no longer correctly processed by proteasome, TAP and/or the MHC. Whereas escape mutations that affect MHC binding are typically no longer under selection pressure in the next host of the virus (as hosts differ in their MHC alleles), escape mutations that affect the processing of an epitope precursor would make this epitope precursor inaccessible for MHC-binding for all MHC alleles in the population.

Results: We designed an agent-based model in which an HIV-1 like virus adapts to the antigen presentation pathway of individual hosts, and spreads through the host population. We tracked whether the virus would adapt to the monomorphic proteasome and TAP, and what the consequences were for the level of adaptation to the host population that the virus could reach. We found that in a host population with a high degree of MHC polymorphism, viruses are under selection pressure to accumulate escape mutations that prevent antigen processing, rather than MHC-binding. As expected, an increase in the degree of MHC polymorphism also increased the number of epitope precursors in the virus that were intermittently under host immune selection pressure. In an unadapted virus, the typical number of epitope precursors under selection pressure within a host is independent of the degree of MHC polymorphism in the population. Thus an increase in the total number of viral epitope precursors utilized by the host population resulted in an average increase in the time between exposures to the immune system for each epitope and epitope precursor. In the model, the increased time between exposures led to an increased reversion of CTL epitope escape mutations, and limited the level of CTL adaptation that the virus could reach.

Conclusion: Even though an HIV-1 like virus will adapt to the monomorphic components of the antigen presentation pathway when encountering a high degree of MHC polymorphism, the virus does not achieve a higher level of adaptation to the host population by doing so. The increased time between exposures to the immune system for individual epitopes and epitope precursors in a population with a high degree of MHC polymorphism results in an increased reversion of escape mutations, which negates the benefit the virus has by adapting to the monomorphic proteasome and TAP.

4.2 INTRODUCTION

The antigen presentation pathway provides the immune system with a way to detect intracellular pathogens, by displaying intracellular protein fragments on the cell surface. One of the most remarkable features of this pathway is the high degree of polymorphism in one of its components. With over 2300 alleles known (Robinson et al., 2006; Sayers et al., 2009), the major histocompatibility (MHC) class I molecules are the most polymorphic genes in the human genome. This polymorphism is thought to have developed in response to the selection pressure exerted by pathogens in at least two ways: by means of the heterozygote advantage (Doherty and Zinkernagel, 1975; Carrington et al., 1999) (i.e. the ability of heterozygote hosts to present a wider range of epitopes) and the rare allele advantage (Slade and McCallum, 1992; Langefors et al., 2001; Borghans et al., 2004) (i.e. pathogens typically carry the least escape mutations for the rarest MHC alleles).

The antigen presentation pathway involves two other molecules besides the polymorphic MHC alleles: the proteasome that cleaves proteins, and the transporter associated with antigen processing (TAP) that transports epitope precursors into the endoplasmic reticulum, where they bind to the MHC. Surprisingly, it is only the antigen presentation step of the pathway (MHC class I) that has developed a large degree of polymorphism, even though viruses can escape CTL responses by adapting to any of the steps in the antigen presentation pathway (Yokomaku et al., 2004; Kwun et al., 2007) (Chapter 2). It would seem that there is a fitness advantage for viruses to escape epitope processing by the monomorphic proteasome and TAP.

In Chapter 2 we studied 30 years of available HIV-1 sequence data, but found that HIV-1 did not appear to accumulate epitope precursor escape mutations. We postulated a mechanism why HIV-1 would not be able to do so, based on the specificity of the MHC alleles within a host, and the polymorphism of the MHC in the population. These two characteristics of the MHC result in an intermittent exposure of epitope precursors to immune selection pressure, as HIV-1 is transmitted from one host to another. Without a constant selection pressure to maintain epitope precursor escapes, these escape mutations can frequently revert back into the wildtype sequence.

In this Chapter we explored the ability of an HIV-1 like virus to adapt its genome to a host population, using an agent-based simulation model. We report

that even though the proteasome and TAP in these host populations are monomorphic, the virus cannot exploit this similarity between hosts to completely escape the cellular immune response. In all host populations that have an MHC polymorphism, the virus reaches a quasi-steady state in which the accumulation of new escape mutations is balanced by the reversion of escape mutations that are not under immune selection pressure in its current host.

In a host population with a high degree of MHC polymorphism, the virus is selected to predominantly generate epitope precursor escape mutations. However, the benefit of adapting to escape proteasome and TAP is negated by an increase in reversion of escape mutations, as the total time that each individual epitope precursor or CTL epitope is under selection pressure in such a host population is lower than in a host population with a low degree of MHC polymorphism.

4.3 MATERIALS & METHODS

4.3.1 *Agent-based model: actors and events*

The agent-based model consists of two types of actors (*hosts* and *viruses*) and four types of events (*procreation*, *death*, and *infection / sexual contact* for the hosts, and *adaptation* for the virus). For each time-step of 0.1 years in the model, the events that will take place are determined based on their predefined frequency per year, and are subsequently applied in a random order to all hosts in the population. On average, each host participates in 0.5 procreation events, 1 death event, and 2 infection events per year, and each virus in 10 adaptation events. Following is a detailed description of each of these 4 events.

- **Host procreation:** The selected host passes on its proteasome, TAP and half of its MHC alleles to a child. The other half of the MHC alleles is drawn from a constant pool of MHC alleles in order to keep the degree of MHC polymorphism in the population stable. The chance of successful childbirth decreases linearly with the population size (i.e. we have logistic growth). The newborn children are given the age of 15, and are added to the host population.

- **Host death:** The host is removed from the population if it fails to pass an age-dependent and viral-load-dependent death chance. The chance of dying is a mathematical approximation of an age-specific mortality curve (Gompertz, 1825; Carnes et al., 2006; Hallén, 2007). If the host is infected by a virus, the chance of dying is further increased, based on the number of years that the host has been infected, and the viral load of the virus (Lavreys et al., 2006) (Eq. 5.3).

- **Host infection / sexual contact** events are short-term relationships between two hosts. Transmission of the virus can happen in both directions between the host and a randomly selected partner if one, but not both of them are

infected. The chance to transmit the virus depends on the viral load in the infected host (Eq. 5.2) (Chakraborty et al., 2001; Wawer et al., 2005). In the model every host is the initiator of an infection event once per year, and thus on average every host takes part in an infection event twice per year. The chance of infection per event (Eq. 5.2) is based on the chance of infection per sexual contact and the number of contacts per year (Wawer et al., 2005).

- **Viral adaptation** can only happen in hosts that are infected with a virus. A mutant of the virus is created by randomly changing one of its amino acids, and this mutant replaces the original virus if its fitness in the host is equal to or greater than that of the original virus. The fitness of a virus is calculated by its number of epitopes and its distance from the wild-type, and is expressed as a viral load (Eq. 4.1). This adaptation scheme is a simplistic representation of selective sweeps that occur during the within-host competition between viruses (Asquith et al., 2006), and captures the essence of within-host epitope escape mutations, reversion of escape mutations and neutral drift in a computationally non-intensive way.

4.3.2 *Antigen presentation pathway*

The classical antigen presentation pathway can be described as three filters (proteasome, TAP, MHC) that are applied to intracellular proteins. The pathway tests which peptides in a protein can successfully pass through all three filters, and thus be presented as CTL epitopes on the cell surface (Tenzer et al., 2005; Groothuis et al., 2005) (Chapter 2). Although current algorithms can accurately model this pathway for a limited number of MHC alleles (Larsen et al., 2005; Tenzer et al., 2005), we have opted for a simpler and computationally faster approach in this model, and let three regular expressions represent the proteasome, TAP and MHC specificity. Regular expressions are commonly used to search for complicated text patterns, and can efficiently locate certain letter combinations in a string of text (Fig. 4.1). The model does not use 20 amino acids, but just 8. The effect is twofold: it increases the chance of reverting escape mutations from 1 over 19 to 1 over 7, and increases the fraction of possible escape mutations per position from ± 4 out of 19 to $\pm$ 3.5 out of 7, taking into account alphabetic distance (see below section on Model equations). Because a larger fraction of the mutations has an effect on escape or reversion, the computation speed of the model increases.

Hosts are diploid, but always homozygous for proteasome and TAP. Hosts have two MHC loci, and thus can carry up to four different MHC class I alleles. The total number of *unique* CTL epitopes presented by these four MHC class I alleles (rather than just the total) is used to determine the fitness of the host (thus allowing for a heterozygote advantage (Doherty and Zinkernagel, 1975; Carrington et al., 1999)).

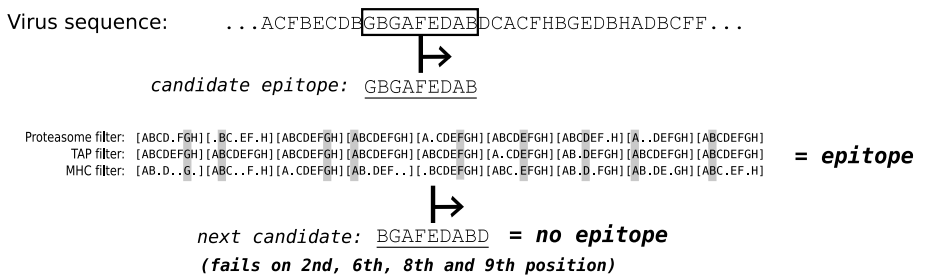

Figure 4.1 Three pattern filters act as the proteasome, TAP and MHC steps of the Ag presentation pathway. In this example, proteasome can only recognize 9-mer peptides that do not start with the amino acid E, and do not have an A, D or G amino acid in the second position, etc. TAP, being less specific than the proteasome, is only sensitive in its 6th and 7th position and cannot recognize any 9-mer peptides that have a B and a C, respectively at those positions. Per position highlighted region indicate a match to all three filters. Peptides like GBGAFEDAB that can be recognized by all three filters are counted as CTL epitopes.

4.3.3 *Model equations*

The fitness of the HIV-1 like virus in the model is determined by its viral load, and expressed as the logarithm of the viral load V. A virus has a base viral load V_b of 9 (i.e. 10^9 copies per ml), and is penalized for every unique CTL epitope e that a host can recognize, and for the number of mutations m that the virus carries. Each epitope decreases the viral load by a 0.2 log change (Kiepiela et al., 2007), and each amino acid that is different from the wildtype sequence decreases the viral load by a 0.1 log change multiplied by the alphabetic distance between the wildtype amino acid, and the mutant amino acid. e.g, mutating a 'D' amino acid into an 'A' would cost the virus a $0.1 \times 3 = 0.3$ log change in fitness, whereas mutating the 'D' amino acid into an 'E' would only cost 0.1 log change in fitness. The log viral load V can not fall below 0. Eq. 4.1 describes the above explained relationship:

$$V = \max\{\ 0\ ,\ V_b - 0.20e - 0.10d\ \}. \tag{4.1}$$

The infectiousness I of an infected host per infection event is determined by the viral load of the virus in the infected host (Chakraborty et al., 2001) (Eq. 5.2). During the acute phase of an infection (first three months), the infectiousness I is temporarily inflated by increasing the viral load by two a hundredfold. One infection event represents a short-term relationship of six months and 60 sexual acts in which the infected host can transmit the virus to susceptible host (adapted from sexual activity data in Wawer et al. (2005)). Eq. 5.2 mathematically approximates the dependency of the probability of transmission on viral load, adapted from Chakraborty et al. (2001):

$$I = 1 - \left(1 - 0.0004 \times \left(0.25 \times V\right)^8\right)^{60}. \tag{4.2}$$

By means of these infection events, hosts are connected to each other in a simple dynamical sexual contact network. Mother-to-child transfer of the virus is not included in this model.

The death rate of hosts D consists of two components. Firstly, hosts can die of of old age, which is a function of age a, based on the estimated intrinsic death rate of North Americans (Carnes et al., 2006). Secondly, hosts can die of the viral infection. This infection-related chance of dying is a function of the time since infection y and the viral load V (Chakraborty et al., 2001; Wawer et al., 2005). In infected hosts, these two chances are summed (Eq. 5.3). In uninfected hosts, the chance of dying to infection is set to zero.

$$D = \mathrm{e}^{(0.1a)-10.5} + \mathrm{e}^{(-0.4a)-8} \; + \; \mathrm{e}^{0.1yV-5} \tag{4.3}$$

4.3.4 *Model initialization*

The model is initialized with a host population at its maximum population size of a 5000 hosts, with a random age between 0 and 100. An MHC polymorphism for both MHC loci of 20 alleles is in place, modeled after the frequency distribution and polymorphism of MHC alleles with a frequency of $\geq$ 1% in the European population (dbMHC-Anthropology (Sayers et al., 2009)). The proteasome and TAP genes are monomorphic. 5% of the population is inoculated with a randomly generated wildtype virus sequence at the start of the model. Simulations typically run for a 1000 years.

4.4 RESULTS

4.4.1 *Model*

To study the potential of viruses like HIV-1 to adapt to the human population, we constructed an agent-based model of a small host population infected with a chronic virus. We kept track of the level of adaptation that this virus reached to the whole population, while it was undergoing within-host adaptation.

The model itself is simple in design: a host population was created, from which members were randomly selected and subjected to one of four events: procreation, death, within-host adaptation of the virus, and infection of another host. The selected host had to meet certain natural criteria for an event to successfully take place (e.g. infection of another host could only occur if the selected host was infected with a virus in the first place). This cycle of selecting hosts and applying events to them is simply repeated as long as the simulation runs.

Hosts are defined by their age, time since infection, and their antigen presentation pathway molecules. The latter consist of a single proteasome and TAP al-

Table 4.1 Model parameters

Parameter	Value	Notes
Host population parameters		
Maximum population size	5000	
MHC diversity	2 loci, each with 20 alleles	a sample of 3500 Europeans contained 16 and 20 different MHC alleles (HLA-A, HLA-B) in frequencies $\geq$ 2%, dbMHC Anthropology (Robinson et al., 2006).
Procreation events	0.5 per host per year	logistic growth
Age-dependent death events	1 per host per year	(Carnes et al., 2006)
Virus parameters		
Length	500aa	single protein is the main determinant of viral fitness in HIV-1 (Kiepiela et al., 2007)
Evolution events	10 per pathogen per year	leads to an average of 3.5 escape mutations, and an equal amount of reversions per infection.
Infection events	1 per host per year	
Effect of a single CTL response on the viral load	-0.20 log_{10}	(Kiepiela et al., 2007)
Effect of an escape mutation of alphabetic distance 1 on the viral load	-0.10 log_{10}	assumed to be smaller than the positive effect of escaping an epitope. Estimation is based on the natural amount of variation found in HIV-1 GAG sequences.
Effect of acute phase of the disease	+2 log_{10} viral load, for 3 months.	(Piatak et al., 1993; Costin, 2007)
Virus-dependent death rate	dependent on time since infection, and viral load	(Lavreys et al., 2006)
Chemical parameters		
Number of amino acids	8	see 4.3.2
Epitope size	9	Most common size for CTL epitopes
Specificity of proteasome + TAP combined	0.25	(Burroughs et al., 2004)
Specificity of MHC alleles	0.05	(Burroughs et al., 2004; Assarsson et al., 2007; Tenzer et al., 2005)

lele and four MHC alleles. Each of these alleles is implemented as a pattern filter which only allows a subset of peptides to 'pass through' (see Fig. 4.1, Methods). The specificity of each of these filters matches the estimated specificity of its corresponding step in human antigen presentation.

Viruses are represented by a string of letters, and tested against the antigen presentation pathway of the host to determine the number of epitopes that the virus carries (i.e. the number of substrings that can pass through the filters of the current host). The number of unique epitopes that the virus carries, as well as its distance (alphabetic distance, see Methods) to the wildtype determine the viral load. The viral load in turn influences the chance that a host dies during a death event (Eq. 5.3), and the chance that transmission of the virus occurs during infection events (Eq. 5.2). Within hosts the virus produces mutants. A mutant virus replaces the resident virus within the host if it has an equal or higher fitness. In this way both selection and neutral sequence drift can occur within a host. The model is described in more detail in the Methods section, and its parameters are given in Table 4.1.

With this model, we can track the adaptation of a virus to its current host (e.g, first panel, Fig. 4.2), and observe the rate and fraction at which the virus accumulates antigen presentation escape mutations (black lines), and antigen processing escape mutations (red lines).

4.4.2 *Intermittent Exposure*

In Chapter 2, we estimated that in a typical host only 18% of the epitope precursors of a virus will bind to any of the hosts' 4 MHC class I alleles, due to the high specificity of the MHC. The other epitope precursors are not under immune selection pressure for as long as the virus remains in that host. When the virus is transmitted to a new host, the immune selection pressure shifts to a new set of epitope precursors, due to the MHC polymorphism in humans, and any epitope precursor escape mutation that is no longer under selection pressure can revert to the wildtype sequence, just as MHC binding escapes do (Friedrich et al., 2004; Barouch et al., 2005; Herbeck et al., 2006). We postulated that this intermittent exposure of epitope precursors due to the selectivity and polymorphism of the MHC is what prevents HIV-1 from efficiently exploiting the monomorphic property of the proteasome and TAP to escape the CTL responses against the virus (Chapter 2).

The effect that this intermittent exposure of epitope precursors has on the adaptation of a virus to the antigen processing machinery can be visualized by tracking a single virus as it passes from one host to the other (Fig. 4.2). During these passages, three properties of the virus are monitored: 1) its current number of CTL epitopes in the host, divided by the number of CTL epitopes that the original wildtype of the virus would have had in the host (black line). 2) its loss of epitope precursors compared to the number of epitope precursors in the wildtype, divided by the total number of epitope precursors in the wildtype that could bind to any of the MHC alleles in the population (red line). 3) its

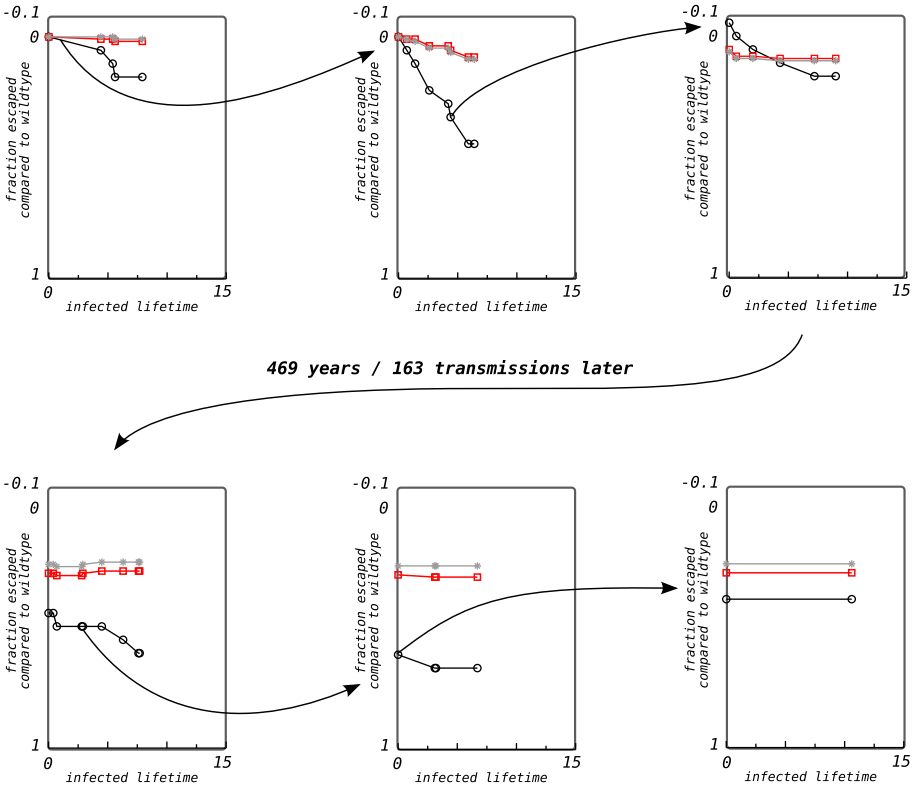

Figure 4.2 The adaptations of a single virus variant as it passes through multiple hosts. Each panel represents a host. The time since infection is plotted on the x-axis, and the within-host level of adaptation to CTL epitopes (black lines), the population-wide adaptation to epitope precursors (red lines), and the population-wide level of adaptation to CTL epitopes (grey lines) is shown on the y-axis. The y-axis ranges from -0.1 (less adapted than the wildtype) to 0 (as adapted as the wildtype to the host population) to 1 (fully adapted to the host population). Arrows indicate the moment in the infection where the host transmitted the virus to the next host.

total number of epitope precursors that could bind to at least one of the MHC alleles in the population, divided by the same measurement in the wildtype (grey line).

Because we scale to wildtype expectation, and correct for the fraction of epitope precursors that is expected not to bind to any of the MHC alleles, we can directly compare the levels of within-host CTL epitope adaptation (1), population-wide epitope precursor adaptation (2), and population-wide CTL epitope adaptation (3), as all three range from 0 to 1 after scaling.

On average, the wildtype virus carried 16.5 CTL epitopes in each host, in a short sequence of 500 amino acid that we designated as the source of all immunodominant epitopes in the virus, and escaped 3.5 CTL epitopes during

its stay in a host. Within a host, the number of CTL epitopes in the virus is generally decreasing (black lines), as the virus adapts to its host. When the virus is transmitted to a new host with different MHC alleles, the number of within-host CTL epitopes changes to a new starting point, from which it starts decreasing again (Fig. 4.2.)

Both the population-wide level of adaptation to epitope precursors (dark grey lines) and the population-wide level of adaptation to CTL epitopes (light grey lines) drop during within-host evolution, but at a slower pace than the loss of within-host CTL epitopes (black line), as each host represents only a small part of the host population. When a virus is transmitted to a new host, the population-wide level of adaptation of epitope precursors and CTL epitopes remain at exactly the same level as in the old host, as these two measurements are not directly affected by the differences in genetic composition between the old and the new host (Fig. 4.2).

At the beginning of the epidemic, the virus predominantly accumulates adaptations to its current host and to the host population (Fig. 4.2, top 3 panels), as it has not yet accumulated many escape mutations that could be reverted. As the virus passes through more and more hosts, and accumulates CTL epitope escape mutations, the virus carries more escape mutations, and reversion of escape mutations start to occur more frequently (Fig. 4.2, bottom 3 panels). Eventually, reversions of escape mutations that are no longer under selection pressure in the current host, and escape mutations in the current host happen at the same frequency, and the population-wide adaptation approaches a quasi-steady state.

4.4.3 *Viral adaptation approaches a quasi-steady state*

In the previous section, we followed a single virus variant through its subsequent hosts, and described how it approached a quasi-steady state level of adaptation to the population. Now we study the adaptation of all virus variants in the population from the start of the epidemic onwards. As the virus adapts to the population, the average population-wide level of adaptation to CTL epitopes in the virus decreases rapidly at first, and then slowly settles into a state in which it has about 28% less population CTL epitopes than the wildtype virus had (black line Fig. 4.3), i.e. is pre-adapted for 28% of the CTL epitopes. Over the course of the epidemic, the population size dropped to 82% of the original population size, and the average age of individuals drops from 46 to 24 years. In the quasi-steady state, $\pm$ 65% of the population is infected with the virus. This is a higher prevalence than the maximum prevalence that has been observed thus far in Swaziland (42% (Mathunjwa and Gary, 2006)), and close to an earlier estimate for HIV-1 of a prevalence of 70% once the epidemic approaches a quasi-steady state (van Ballegooijen et al., 2003). It could be that the HIV-1 epidemic has not yet approached a quasi-steady state, and that the prevalence will rise even higher. However, our model does not account for possible changes in the social behaviour of the afflicted host population, or the effect of

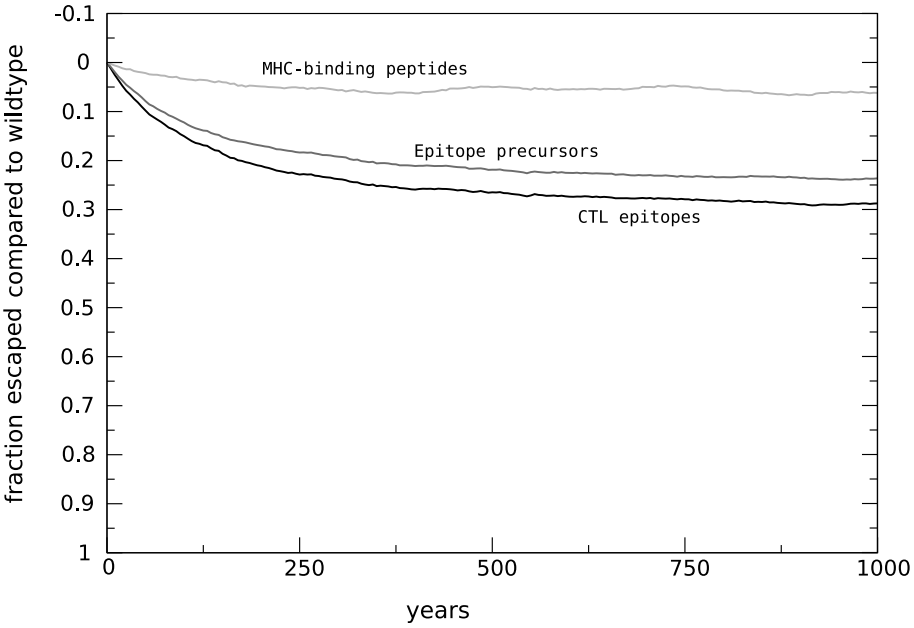

Figure 4.3 The quasi-steady state that the HIV-1 like virus approaches in a population with a polymorphism of 40 MHC alleles is characterised by a population-wide loss of ± 28% of its CTL epitopes, 24% of the epitope precursors that could bind to at least one of the MHC alleles in the population, and 6% of the MHC-binding peptides, compared to their respective densities in the wildtype sequence. 81% of the escape mutations affected the processing of the epitope precursors, and 21% affected MHC binding. The graph shows the averaged values of 20 simulations.

drug treatment. Nor does it allow for large changes in its initial MHC allele frequenties, such that MHC alleles that confer resistance to the virus could rise in frequency (Carrington and O'Brien, 2003). These differences between the simulation model and reality could account for the high prevalence in the model, compared to the prevalence observed for HIV-1 epidemics in southern africa. Remarkably, 81% of the adaptations that the virus had accumulated were mutations that prevented the processing of an epitope precursor, while only 21% of the mutations affect the MHC-binding of the peptide (Fig. 4.3), even though the latter type of mutation is more frequently generated due to the higher specificity of the MHC (Brander et al., 1999; Yokomaku et al., 2004) (Chapter 2).

4.4.4 *Effect of MHC polymorphism on adaptation*

In a population with a total of 40 MHC alleles, the virus predominantly accumulated epitope precursor escapes (Fig. 4.3). We found that the ratio of accumulated epitope precursor to MHC-binding escapes in the virus gradually shifted

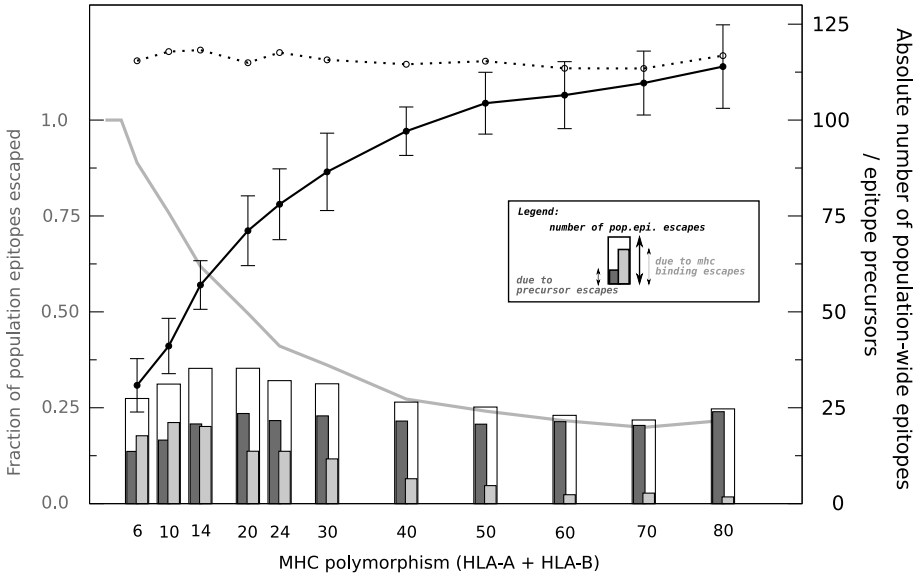

Figure 4.4 The effect of the degree of MHC polymorphism on the quasi-steady level of adaptation that the virus approaches. Increasing the MHC polymorphism (x-axis) increases the fraction of the absolute number of escaped CTL epitopes (white bars, righthand y-axis) that are due to precursor escapes (dark grey bars) and decreases the fraction that is due to MHC binding escape mutations (light grey bars). At the same time, increasing the MHC polymorphism increases the number of epitope precursors that bind to one or more of the MHC alleles in the population (black line, with std. dev, righthand y-axis) from 31 up to 115 epitope precursors, which approaches the total number of epitope precursors in the virus (dotted line). The grey line (lefthand y-axis) is the fraction of all CTL epitopes in the host population (black line) that the virus carries escape variants for (white bar), and thus represents the level of adaptation that the virus can evolve to, for different degrees of MHC polymorphism in the host population. At an MHC polymorphism lower than 6 MHC alleles the host population would not survive the virus.

towards epitope precursor escapes with an increase in the degree of MHC polymorphism in the host population (bars in Fig. 4.4). At an MHC polymorphism of 6 alleles, the majority of escape mutations are MHC binding escape mutations, but at high degrees of MHC polymorphism (e.g. $\geq$ 20 alleles), the virus variants that acquired epitope precursor escape mutations outcompete the viruses that acquired MHC binding escape mutations.

Not only does a high degree of MHC polymorphism select for pathogens that have accumulated epitope precursor escape mutations, it also has an obvious effect on the number of epitopes in the population that can be bound by at least one of the MHC alleles in the host population (black line, Fig. 4.4). Raising the MHC polymorphism from 6 MHC alleles to 80 MHC alleles increases the absolute number of epitope precursors that can bind to at least one of the

MHC alleles from ± 31 to a 115 epitope precursors. As the number of epitope precursors that is expected to bind to the MHC alleles of a single host remains constant, an increase in the absolute number of epitope precursors in the population implies that on average each epitope precursor in the virus is under immune selection pressure a shorter fraction of the time. As a result, even in a situation where the virus was predominantly adapting to the monomorphic proteasome and TAP, the virus fails to achieve a high level of adaptation to the host population (grey line, Fig. 4.4). The level of adaptation to the host population that pathogens can reach decreases with an increase in MHC polymorphism, and levels off at ± 25%.

4.5 DISCUSSION

Viruses that escape a CTL response against a particular epitope by abrogating the MHC-binding of that epitope would typically have no advantage of this adaptation in another host, due to the extensive MHC polymorphism. A major concern of ours when we studied the antigen presentation pathway was that in response to the MHC polymorphism, viruses would evolve to prevent the processing of epitope precursors by the monomorphic proteasome and TAP, rather than to prevent MHC binding. This concern was based on reports of antigen processing escapes (Brander et al., 1999; Yokomaku et al., 2004), and was confirmed in our simulations: in human populations with a high degree of MHC polymorphism, viruses were selected to adapt almost exclusively to the monomorphic components of the antigen presentation pathway (Fig. 4.3).

Surprisingly, even when adaptation occurred primarily to the proteasome and TAP, viruses could not escape all of the CTL responses against their CTL epitopes, but approached a quasi-steady state in which escape mutations for CTL epitopes were created and reverted at an equal rate. The long time between exposure to immune selection pressure for individual epitope precursors in a population with a high degree of MHC polymorphism balances the accumulation of epitope precursor escapes with the reversion of escape mutations. The exact amount of time between two subsequent exposures of an epitope or epitope precursor to the immune system is determined by the host history of a particular virus variant, and depends on the distribution of MHC alleles, their promiscuity (Frahm et al., 2007) and the degree of MHC polymorphism in the contact network through which the virus moves. This direct relation between escape variant frequency and MHC allele frequency is clearly visible in a recent study by Kawashima et al. (2009).

Recently, Kawashima et al. (2009) and Schellens (2009) reported that HIV-1 is still accumulating adaptations to the human immune system, because for several CTL epitopes[1], the fraction of HIV-1 viruses that carried escape variants of these epitopes had increased between the 1980's and now (or in the case of

[1] A discussion on how to interpret the apparent lack of large-scale adaptation of HIV-1, and the accumulation of escape variants for certain epitopes can be found in Chapter 2.5

Schellens (2009), the number of CTL epitopes that were predicted to bind to certain MHC alleles had decreased since the 1980's). In both cases, this apparent accumulation of escape mutations in the virus could also reflect the adaptation of HIV-1 to the different MHC frequencies in new locale. For example, in the Kawashima et al. (2009) study, the HLA-B*51-negative Japanese haemophiliacs from 1983 were infected with HIV-1 through blood transfusion from blood plasma imported from the USA (Shimizu et al., 1992). Thus, the low escape variant frequency in 1983 (21% contained the RT I135X escape mutation) could just reflect the level of adaptation that HIV-1 virus variants in the USA had reached for this particular MHC allele (prevalence in the USA: 12% (Kawashima et al., 2009; Robinson et al., 2006)). By 1997 and 2008, the HIV-1 virus that circulated in Japan would have encountered HLA-B*51 positive hosts (prevalence: 22% (Kawashima et al., 2009)), and the high frequency of the I135X escape variant in HLA-B51-negative Japanese patients ($>$ 50%) would by now reflect the Japanese HLA-B*51 allele frequency. Based on this alternative explanation, we predict that epitope escape variants for MHC alleles that are common in the USA but rare in Japan, would have decreased in frequency in the Japanese HIV-1 infected patients. A recent paper by Vider-Shalit et al. (2009) describes a decrease in the 'size of the immune repertoire (SIR)'-score between the transition from SIV to HIV, which suggests a loss of multiple epitopes in HIV-1 since the virus jumped species. Oddly enough, the GAG protein appears to be increasing in SIR score over time. What causes the differences in predictions of Vider-Shalit et al. (2009) and the predictions based on the MHC-pathway is currently under investigation.

In our model, the HIV-1 like virus started completely unadapted to the host population, and evolved within 200-500 years towards a stable state where it carries escape variants for $\pm$ 25-30% of its CTL epitopes (depending on the degree of MHC polymorphism, see Fig. 4.4). However, HIV-1 is related to and likely originated from the SIVcpz virus in the chimpanzee subspecies *Pan troglodytes troglodytes* (Keele et al., 2006). Furthermore, humans and chimpanzee are closely related, and share common features and binding patterns in their antigen presentation pathway (Hoof et al., 2008; Anzai et al., 2003). Therefore, it seems possible that part of the adaptation of the virus to the proteasome, TAP and MHC alleles of the human population had already occurred in chimpanzee. Such a partial adaptation to the human population could be the reason why in Chapter 2 a decrease in the total number of (predicted) HIV-1 CTL epitopes in the period between 1980 and 2005 was absent: the quasi-steady state of Fig. 4.3 would have already been approached prior to the 1980's.

Concluding, we find that a high degree of MHC polymorphism in a host population selects for viruses that escape the monomorphic proteasome and TAP. However, a high degree of MHC polymorphism also increases the time between exposures to CTL responses for any epitope precursor. As a result of this, the selection pressure on a virus to maintain a particular escape mutation decreases, and the level of adaptation to the antigen presentation pathway that a virus can evolve is reduced. By increasing the number of epitope precursors that

are intermittently under selection pressure, the MHC polymorphism indirectly protects the monomorphic proteasome and TAP from large-scale adaptation by viruses.

4.6 ACKNOWLEDGEMENTS

The authors would like to thank Can Keşmir for her suggestions and for critically reading the manuscript. This study was financially supported by the Netherlands Organisation for Scientific Research (NWO) (VICI grant 016.048.603, Open Program 812.07.003), and a High Potential grant (2007) from the Utrecht University.

The emergence of polymorphism in the antigen presentation pathway

5

Boris V. Schmid[1], Can Keşmir[1,2], Rob J. de Boer[1]

1 Theoretical Biology, Utrecht University, The Netherlands.
2 Academic Biomedical Centre, Utrecht University, The Netherlands.

Manuscript in preparation

5.1 ABSTRACT

The classical antigen presentation pathway has three main components; the proteasome, TAP and MHC class I molecules. Of these three, only the MHC evolved a functional polymorphism, whereas the former two have remained monomorphic. As pathogens can escape antigen processing and presentation by adapting to any of these molecules, the heterozygote advantage, and rare allele advantage selection pressures -which are thought to be responsible for the extensive MHC polymorphism- should apply to all three steps in the pathway. It is not understood why proteasome and TAP have not evolved a polymorphism. Previously we predicted that it would be sufficient for the host population if at least one of the steps of the pathway would be polymorphic, and that it would be most beneficial for the host population if that step was the most specific step of the antigen presentation pathway (i.e. the MHC). Here we studied the evolution of polymorphism in the antigen presentation pathway in an agent-based model, in which pathogens can adapt to any of the three steps of the pathway, and all three steps of the pathway are capable of evolving a polymorphism in response to the pathogen selection pressures. Under the condition that the host population maintains a high degree of coevolvedness between the steps of the Ag presentation pathway, we found that it is always, and only, the most specific step that becomes polymorphic.

5.2 INTRODUCTION

One of the most remarkable features of the classical antigen (Ag) presentation pathway is the high degree of polymorphism of one of its molecules. The major histocompatibility complex (MHC) class I molecules are the most polymorphic genes in the human genome, with over two-thousand alleles described (Robinson et al., 2006). This polymorphism is thought to have developed as a

response to pathogens, by means of the heterozygote advantage (HA) (Doherty and Zinkernagel, 1975; Carrington et al., 1999) and the rare allele advantage (RAA) (Slade and McCallum, 1992; Langefors et al., 2001; Borghans et al., 2004; de Boer et al., 2004).

The Ag presentation pathway involves two other molecules besides the polymorphic MHC alleles: the (immuno-) proteasome which cleaves proteins, and the transporter associated with antigen processing (TAP) which transports epitope precursors into the endoplasmic reticulum, where they bind to the MHC. Surprisingly, only the MHC class I molecules have developed a large degree of polymorphism, even though viruses can escape CTL responses by adapting to any of the steps in the Ag presentation pathway (Yokomaku et al., 2004; Kwun et al., 2007) (see Chapter 2). It would seem that there is a fitness advantage for viruses to escape epitope processing by the monomorphic proteasome and TAP (Yusim et al., 2002). Indeed, in an agent-based simulation model of an MHC-polymorphic host population, pathogens were strongly selected to accumulate escape mutations that affect proteasomal and TAP processing of epitope precursors (see Chapter 4).

Because pathogens adapt not only to the MHC, but also to the proteasome and TAP, one would expect that all steps of the Ag presentation pathway would have evolved a polymorphism in response to pathogens. However, in the human species, proteasome and TAP are functionally monomorphic (Gomez et al., 2006; Alvarado-Guerri et al., 2005; Faucz et al., 2000). To study the factors that shaped the current diversity of molecules in the human Ag presentation pathway, we build an agent-based model of a host population that is infected with several endemic pathogens. Hosts have a small chance of passing a mutated proteasome, TAP or MHC allele to their offspring, and thus contribute to the polymorphism of these molecules in the population. Pathogens infect individual hosts and adapt their amino-acid sequence to the Ag presentation pathway of that host. We found that all three steps of the Ag presentation pathway are individually capable of becoming polymorphic in response to the pathogens.

In Chapter 2 we suggested that the evolution of one polymorphic step in the Ag presentation pathway would be sufficient to prevent pathogens from adapting to any step in the pathway. We further suggested that a polymorphism would be expected to evolve in the most specific step of the Ag presentation pathway, as that would provide the best protection against adaptation to the Ag presentation pathway. Our results confirm both of these predictions. Under the condition that new alleles are limited in their mutational freedom such that the simulated host population maintains a high level of coevolution between the individual steps of the Ag presentation pathway, there is indeed a strong selection pressure for the MHC to become polymorphic. We can further refine our hypothesis from Chapter 2 by stating that under the aforementioned condition, the expectation should not be that *at least* one step of the Ag presentation pathway will become polymorphic, but that *only the most specific* step will become polymorphic. In simulations where the alleles were evolving fully unconstrained, additional quasi-steady states existed, but with lower host fitness.

In these simulations, the initial conditions of the Ag presentation pathway put the host population in different basins of attraction.

5.3 METHODS

We use an agent-based model that consists of two types of actors (*hosts* and *viruses*) and five types of events (*procreation*, *death*, *infection* of the hosts, and *escape* and *reversion* of the pathogens within hosts). The timestep of the model is 1 year, and each year every host participates in 0.5 birth events, 1 death event, 2 infection events, 1 escape and 1 reversion event. The order of these events, and the order of hosts is randomized each year. The occurrence of an escape or a reversion event is applied on the host level, i.e. in a single escape event, all pathogens in a particular host have the opportunity to escape one of their epitopes. The following is a detailed description of both actors, and each of the 5 events.

5.3.1 *Pathogens*

The pathogens in the simulations are modeled as chronic pathogens that increase the natural age-related death rate of the host. Once infected with a particular pathogen species, the host remains a lifelong carrier of this pathogen and cannot be superinfected by the same pathogen. Typical simulations contain twenty of such pathogen species, which is sufficient to ensure that the hosts evolve a generic Ag presentation system, and not one that is specifically selected to deal only with the simulated pathogens.

Each pathogen species has a unique randomly chosen wildtype sequence, and has the opportunity to accumulate escape mutations during escape events, and can revert escape mutations that are no longer under selection pressure back to the wildtype sequence during reversion events. The fitness of a pathogen is expressed in viral load. Pathogens start with a base viral load V_b of 9 (i.e. 10^9 copies per ml), and are penalized for every unique CTL epitope e that a host can recognize, and for the number of mutations m that the pathogen carries. Each epitope decreases the viral load by a 0.1 log change (Kiepiela et al., 2007). The cost of a single mutation starts at 0.035 log change, and becomes increasingly more costly the more mutations the pathogen carries:

$$V = V_b - 0.1e - 0.035 \times m^{1.25}. \tag{5.1}$$

The exponent of 1.25 for the cost of additional escape mutations limits the sequence variability in a pathogen to $\pm$ 28 mutations, which translates to $\pm$ 60% of the epitopes in a single host. In general, the pathogens in the model quickly accumulate 28 escape mutations, and thus their rate of escape is limited by the rate of reversion of escape mutations that are no longer under selection pressure in the new host.

Not the whole pathogen is modeled, as not all genes of a pathogen are likely to be expressed early on in the pathogens life cycle, or are expressed in a large enough quantity to play a dominant role in the selection pressure on the pathogen (Kiepiela et al., 2007). The pathogens in the simulations consist of 500 amino acids that will generate the immunodominant immune responses.

The infectiousness of a pathogen I per infection event is determined by its viral load V, and increases non-linearly from 0 to 1 as the viral load V increases from 0 to its maximum V_b (Fig. 5.1A, black line). In the model we use a power of 3, and therefore the relationship between infectiousness of a pathogen and its viral load is described by

$$I = (V/V_b)^3. \tag{5.2}$$

5.3.2 *Host actors*

Host actors in the model are simplified humans, and carry a diploid genome that encodes for an Ag presentation pathway consisting of one proteasome, one TAP and one MHC locus. Furthermore, any combination of two hosts can reproduce meiotically from the moment they are born.

Antigen presentation pathway

The classical Ag presentation pathway can be described as three pattern filters (proteasome, TAP, MHC) that are applied to intracellular proteins. As pattern filters, the proteasome creates overlapping 9 amino-acid long peptides (9-mers) from the virus protein, based on the amino acid patterns that the proteasome allele can recognize, TAP transport those 9-mers into the endoplasmic reticulum, if they match the pattern for which TAP is specific, and the MHC alleles of the host bind to those 9-mer peptides that it can recognize. In this implementation, the three main steps of the Ag presentation pathway differ only in their specificities. Current algorithms such as the MHC-pathway (Tenzer et al., 2005) and NetCTL (Larsen et al., 2005) can accurately model this pathway for the proteasome, TAP and a large number of MHC alleles. However, we have opted for a simpler and computationally faster approach in this paper, and use regular expressions to represent the proteasome, TAP and MHC pattern filters (Fig. 5.2). In a regular expression filter, we can define for each position in a sliding window which amino-acids can be recognized by each allele in the Ag presentation pathway of the host. In contrast to the more sophisticated MHC-pathway predictors, a regular expression either accepts or does not accept a certain amino acid at a particular position. As all of the three steps are diploid, epitopes can be generated by up to 8 different combinations of a proteasome, TAP and MHC allele. The total number of *unique* CTL epitopes presented by these pathways is used to determine the viral load of the pathogens in the host (Eq. 5.1), which increases the death-rate of the host (Eq. 5.3, Fig. 5.1).

Generation of new alleles

There are several ways to generate the pattern filters that serve as Ag presentation pathway genes in the model. Unfortunately, randomly drawing the sets of amino acids that can be recognized at each of the nine positions that make up a gene (Fig. 5.2) is inefficient for pattern filters with a very low or very high specificity. The initial alleles in Fig. 5.3, Fig. 5.4, Fig. 5.5B & C were generated by starting with a fully aspecific pattern filter, and then to remove or add random amino acids from the filter until the desired specificity is reached (e.g. between 31-35%, 71-79%, and 4-6% for proteasome, TAP, and MHC, respectively).

New alleles in Fig. 5.3, and Fig. 5.4 were generated in a way that artificially assisted in maintaining a coevolved Ag presentation pathway. Rather than starting from a fully aspecific pattern filter, a mutated allele in a child for one particular step would be started from a pattern that was the intersection of the filters in the other two steps of the pathway (e.g. a new TAP allele would be constructed from the intersection of the two proteasome and two MHC alleles that the parents contributed to the child). This approach made it more likely that the new allele remains coevolved with the rest of the Ag presentation pathway.

New alleles in Fig. 5.5 were generated by taking the original allele, randomizing the number and identity of the amino acids in one of the nine positions, and then correcting for the change in specificity by adding or removing amino acids from any position in the allele until the right specificity was reached. The biological interpretation of this algorithm is that mutations change one of the key positions in proteasome, TAP or MHC alleles radically, and the structure of the whole molecule changes slightly to accomodate for this change.

5.3.3 *Model events*

- **Procreation:** The selected host and a randomly drawn second host both pass on half of their proteasome, TAP and MHC alleles to a child. During this event, there is a small chance to mutate (10^{-4} per allele) into a novel allele. The chance of succesful childbirth decreases linearly on the density of the host population (i.e., the model implements logistic growth).

- **Death:** The host is removed from the population if it fails to pass an age-dependent and viral-load-dependent death chance. The chance of dying is a function of age a, mathematically approximated from the age-specific intrinsic death rate of North Americans (Carnes et al., 2006) (Fig. 5.1B). This death rate D is multiplied by a factor based on the total viral load of the pathogens within the host (Fig. 5.1A, grey line), and on the fraction of the maximum disease burden that a host could have:

$$D = \left(e^{(0.1a)-10.5} + e^{(-0.4a)-8}\right) \times \max\{\ 1\ ,\ 1000 \times \left(\frac{\sum_i^n 10^{V_i}}{N \times 10^{V_b}}\right)^5\}, \quad (5.3)$$

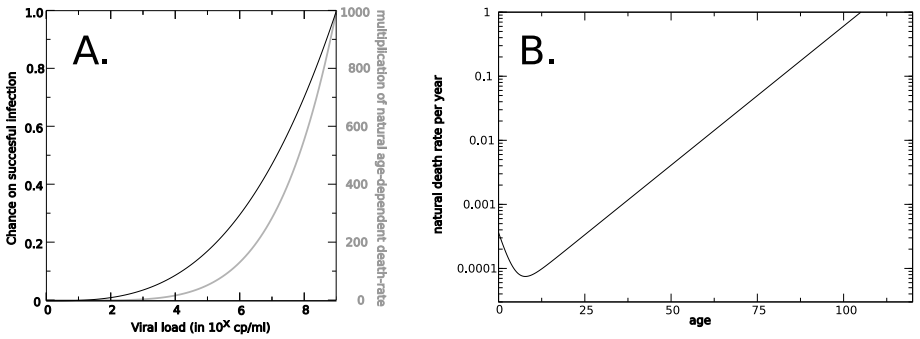

Figure 5.1 **A.** The relation between log viral load and infectiousness (black line, lefthand axis), and log viral load and increased deathrate (grey line, righthand axis). **B.** The intrinsic age-dependent death rate of the host population.

in which N is the total number of pathogen species in the population, V_i is the log viral load of pathogen i, and V_b is the maximum log viral load. The impact of disease burden increases non-linearly, with a power of 5.

- **Infection** of a host occurs by randomly selecting a contact partner, whose different pathogen species each have a chance to infect the host, based on their viral load (Eq. 5.2, Fig. 5.1A, black line). Infection is a one-directional event, and hosts can only be infected by a particular species of pathogen once.

- **Escape** events are applied to all pathogens in a host. Differences in escape mutation rate between pathogens are accounted for by giving each pathogen species a fixed chance (range 0.75 – 1.0) to participate in an escape event. When a pathogen participates in an escape event, we randomly select one of the possible escape mutations that result in a higher viral load (see Eq. 5.1). Escape events do not model the mutation rate of pathogens, but the selective sweep resulting from the competition between escape variants in the within-host quasispecies of a pathogen (Althaus and de Boer, 2008).

- **Reversion** events are applied to all pathogens in a host. Similar to escape events, every pathogen species has a fixed chance of participating in reversion events (range 0.5 – 0.75). During a reversion event, we randomly select one of the mutations in the pathogen that when reverted, results in a higher viral load (see Eq. 5.1). Reversion events do not model the chance that a random mutation is a reversion, but model the outcome of the competition between reversion variants in the within-host quasispecies of a pathogen (Asquith et al., 2006).

5.3.4 *Model initialization*

The model is initialized with a monomorphic host population at its maximum population size of 5000 hosts, with a random age between 0 and 100. A new virus species is introduced into a single host every 100 years until the maximum number of 20 virus species in the population is reached. If a virus species goes extinct, a new virus species with a new wildtype and mutation rate is introduced. Simulations typically run for 80,000 years.

5.3.5 *Simpsons Reciprocal Index (SRI)*

The level of polymorphism in a particular locus can be expressed as the Simpsons Reciprocal Index score (Simpson, 1949). The Simpsons Index is a measurement of diversity that can be interpreted as the probability that two randomly chosen alleles from two random hosts in the population are identical. The lower the Simpsons Index, the higher is the diversity of alleles in the population. The reciprocal of the Simpsons Index puts a number to this diversity, and has the advantage over the total number of unique alleles as a measurement of diversity, that it is less sensitive to fluctuations in allele numbers caused by random neutral drift. For example, a polymorphism where all alleles are equally frequent has an SRI score equal to the number of alleles in the population, whereas a population that is dominated by a single allele will have an SRI score close to 1.

The Simpsons Reciprocal Index R can be expressed as

$$R = \frac{1}{\sum_i^N f_i^2} \tag{5.4}$$

in which f_i is the fraction of allele i over all alleles of that locus in the population, and N is the total number of unique alleles.

5.4 RESULTS

5.4.1 *Agent-based model*

To study the evolution of the Ag presentation pathway, we constructed an agent-based model of a small host population infected with several endemic chronic viruses. Hosts in this model are diploid, and their Ag presentation pathway is implemented as sequential pattern filters, each of which can recognize a predefined fraction of all possible 9 amino-acid long peptides (9-mers). Any peptide from a virus protein that can pass through a proteasome, TAP and an MHC pattern filter is marked as an epitope (Fig. 5.2). As hosts are diploid there are 8 different combinations of proteasome, TAP and MHC alleles through which a peptide might be recognized. Pathogens in the model are implemented as a string of letters that represents their protein sequence. The pathogens can mutate their protein sequence to escape the presentation of epitopes. Pathogen fitness depends on the number of epitopes that they present, and the number of

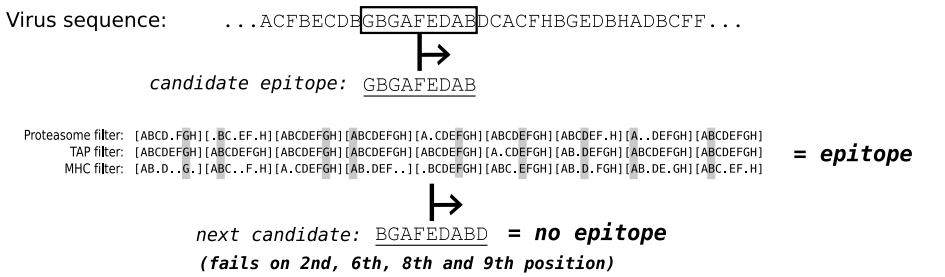

Figure 5.2 Three pattern filters act as the proteasome, TAP and MHC steps of the Ag presentation pathway. In this example, proteasome can only recognize 9-mer peptides that do not start with the amino acid E, and do not have an A, D or G amino acid in the second position, etc. TAP, being less specific than the proteasome, is only sensitive in its 6th and 7th position and cannot recognize any 9-mer peptides that have a B and a C, respectively at those positions. Per position highlighted regions indicate a match to all three filters. Peptides like GBGAFEDAB that can be recognized by all three filters are counted as CTL epitopes. The example regular expressions here can only detect 8 amino acids, but in the simulations the alphabet has been increased to 20 amino acids.

mutations they carry compared to the wildtype sequence of the pathogen (Eq. 5.1).

Pathogens evolve within hosts. The formalism for within-host evolution that is used in the model is that all single-point escape mutation variants of the pathogen are assumed to exist in low numbers in the quasispecies, and that these variants compete until one of them becomes the dominant within-host variant of that pathogen (Althaus and de Boer, 2008). At that point this new dominant variant forms the basis of a new quasispecies, and the process repeats itself. Reversions of escape mutations are implemented in the same way (Davenport et al., 2008). By escaping the presentation of epitopes and reverting obsolete escape mutations, pathogens can increase their infectiousness during host-to-host contacts, but also impair themselves by increasing the death-rate of their host (see Eq. 1-3, and Fig. 5.1).

During the simulation, each host in the population is subjected to one of four events: reproduction, death, adaptation of its pathogens, and infection with new pathogens from another host. The model is described in greater detail in the Methods section.

5.4.2 *Low specificity also allows for polymorphism*

One factor that directly influences the potential degree of polymorphism is the specificity of the mutating molecule (Borghans et al., 2004; Brander et al., 1999). Proteasome and TAP have a low specificity (33% (Burroughs et al., 2004) and 62–84% (Peters et al., 2003; Burroughs et al., 2004), respectively) compared to the MHC class I molecules (1–8% (Burroughs et al., 2004; Tenzer et al., 2005)). As a result of their low specificity, new proteasome and TAP alleles typically have

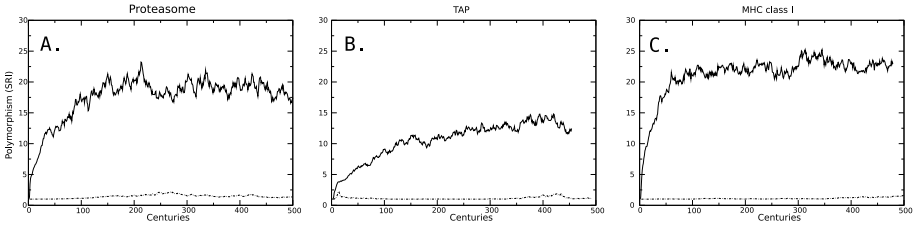

Figure 5.3 All three steps in the Ag presentation pathway are specific enough to evolve a polymorphism. Each panel shows the degree of polymorphism (solid line, expressed in Simpsons Reciprocal Index on the y-axis) that an Ag presentation pathway consisting only of a single diploid step with a specificity of 33%, 75% or 5% respectively, would evolve to. The expected number of epitopes in random pathogens was scaled down to the same number ($\pm$ 48) for the proteasome and TAP simulations by reducing the size of the pathogens in those simulations. Dotted lines indicate the SRI caused by neutral drift for a population with a comparable host turnover rate as the simulations with pathogens.

a large overlap in the 9-mers they can recognize with other alleles in the host population, which reduces the RAA and HA fitness advantage for host carrying these new alleles, compared to hosts carrying a new MHC class I allele. Thus, in a setting where only one of the steps in the pathway is allowed to evolve a polymorphism, the degree of polymorphism is expected to be inversely related to the specificity of that step, because the selective advantage of a new allele depends on how different it is from the existing alleles.

The effect of specificity on polymorphism was tested in a simplified version of the model, in which the Ag presentation pathway of the hosts was modeled as containing only a single step. The specificity of this step was set to 33%, 75% or 5%, to simulate the proteasome, TAP, and the MHC, respectively. The length of the virus protein was adjusted such that the number of epitopes in a random virus was comparable between the three specificity settings. The degree of polymorphism is expressed in terms of the Simpsons Reciprocal Index (Simpson, 1949) (see Methods), which is a measurement of diversity that is less sensitive to fluctuations in allele numbers than counting the total number of unique alleles.

As expected, the host population whose single-step Ag presentation pathway had the specificity of the MHC (5%, panel C), aquired the highest degree of polymorphism with a Simpsons Reciprocal Index (SRI) diversity score of $\pm$ 22.6 alleles (Fig. 5.3C, solid line). The polymorphism in the other host populations followed the order of their specificity, with the specificity of 33% resulting in a SRI polymorphism of $\pm$ 18.2 alleles, and the specificity of 75% resulting in a SRI polymorphism of $\pm$ 13.5 alleles in the last 100 centuries. (solid lines, Fig. 5.3A and Fig. 5.3B). In all simulations, the degree of polymorphism was higher than expected from random neutral drift (dotted lines, Fig. 5.3). The host populations with a specificity of 5% and 33% had a similar increase in death rate (15x and 17x), and expected lifespan (45.1y and 44.65y), whereas for the host population

with a single-step TAP Ag presentation pathway, the increase in death rate was twice as high (33x) and the expected lifespan three years shorter (41.5y).

These results suggest that all three specificities of the Ag presentation pathway allow for a polymorphism. Thus, the low degree of specificity of the proteasome and TAP are not the direct cause of why these molecules are monomorphic in the human host population.

5.4.3 *Emergence of polymorphism in a coevolved Ag presentation pathway*

The above simulations demonstrate that in host populations facing endemic pathogens, all three steps of the Ag presentation pathway are individually capable of forming a polymorphism, in response to the combined selection pressures of the RAA and HA. When all three steps of the pathway are simultaneously allowed to evolve, there is an additional selection pressure on the hosts: the 9-mers that the proteasome pattern filter can recognize should also be recognizable to the TAP and the MHC alleles of a host, for a host to maximize the number of epitopes. In the rest of the paper we will refer to this selection pressure as the "coevolvedness of the Ag presentation pathway", and it will be calculated for each possible combination of the proteasome, TAP and MHC alleles within a host. For naturally occurring 9-mers, it was shown that MHC class I alleles were $\pm$2.5 times more likely to present a 9-mer peptide that was processed by proteasome and TAP than the average 9-mer (Burroughs et al., 2004). We hypothesized that the selection pressure on hosts to keep a coevolved Ag presentation pathway would affect the emergence of polymorphisms in the different steps of the Ag presentation pathway. Therefore, to provide a proof of principle, we first perform simulations with a host population in which the mutation process of the hosts is adapted such that novel alleles are fairly well coevolved to the other steps in the Ag presentation pathway of an individual (see Methods for more details on this algorithm).

In this proof of principle simulation, where the coevolution of new alleles is enforced upon the hosts, the population develops an Ag presentation pathway that consists of a monomorphic proteasome and TAP, and a highly polymorphic MHC with a SRI score of $\pm$ 20-25 alleles (Fig. 5.4). The overlap in 9-mer peptide repertoire between MHC alleles is 25-26%, which is higher than the average of 10% overlap observed between MHC alleles in Fig. 5.3C. The difference between the two cases is that in the single-step Ag presentation pathway (Fig. 5.3C), the MHC alleles could distribute their peptide repertoire over all possible 9-mer peptides, whereas in the latter case the MHC alleles restrict their own peptide repertoires to the smaller set of peptides that both proteasome and TAP can recognize. As a result, the MHC alleles are more similar to each other in the latter case, which leads to a lower functional MHC polymorphism, and a higher increase in death rate for the host population (Table 5.1).

In duplicate runs of Fig. 5.4, it is always the MHC that becomes polymorphic. Even in simulations where initially only TAP or proteasome could evolve, the MHC would eventually become the polymorphic step of the pathway, and push

Table 5.1 Characteristics of the simulation presented in Fig. 5.4.

Host population:	
average age of the hosts:	20 – 21y[a] (down from 38y)[b]
expected lifespan of the hosts:	38 – 39y (down from 81y)
average increase in death rate/year:	38x – 41x (up from 1x)
Antigen presentation pathway:	
average pathway coevolvedness[c]:	0.94 – 0.99
weighted functional polymorphism of proteasome[d]:	0.96 – 1.0
weighted functional polymorphism of tap:	0.97 – 1.0
weighted functional polymorphism of mhc:	0.25 – 0.26[e]
Pathogens:	
average number of epitopes per pathogen:	31
average number of epitopes of a pathogen at infection:	39
average number of epitopes in a random pathogen:	48
average pre-adaptation of pathogens[f]:	24%
average viral load of pathogens:	3.7

[a] ranges and averages are based on the minimum and maximum value over the last 300 centuries of the simulation.

[b] in a host population without pathogens.

[c] average pathway coevolvedness: the average number of 9-mers that all the pathways that are present in the host population can present, divided by the maximum number of 9-mers that a proteasome-TAP-mhc pathway can be expected to present, based on the specificities of the steps.

[d] weighted functional polymorphism: the frequency-weighted functional polymorphism indicates the average overlap in the peptide repertoire between two alleles that were randomly drawn from the population. It therefore takes into account both the functional diversity between different alleles and the frequency distribution of the alleles.

[e] The mutation algorithm used in the simulations is not fully random, both in Fig. 5.4 and in Fig. 5.3C. Therefore the overlap in both simulations is higher than the expected overlap of 5% for two completely random MHC alleles with a specificity of 5% - see Methods.

[f] the pre-adaptation of a pathogen compares the average number of epitopes that a pathogen has at the moment of infection, to the average number of epitopes that the wildtype sequence of that pathogen would have had.

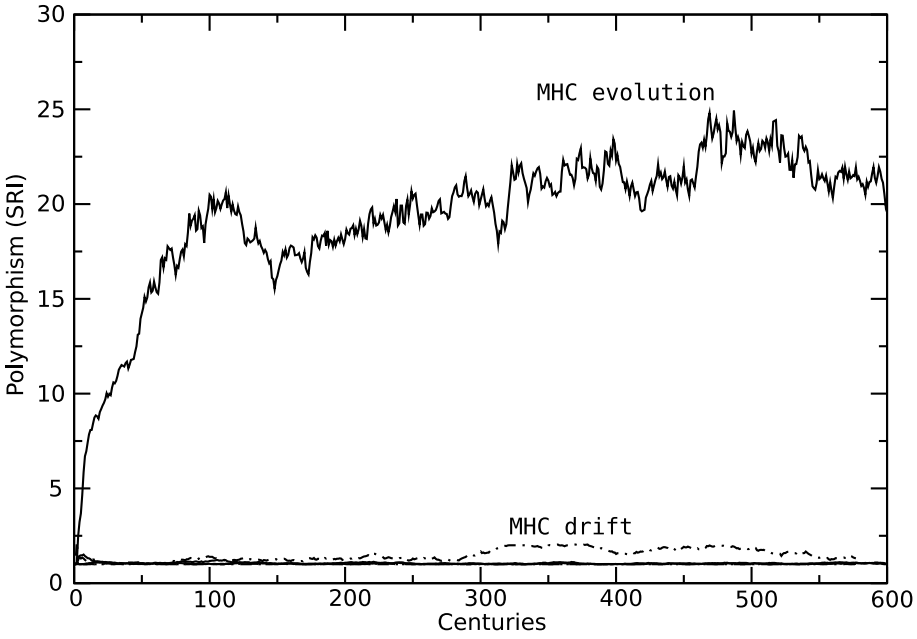

Figure 5.4 When all three steps in the Ag presentation pathway evolve simultaneously, only the MHC alleles become polymorphic (upper solid line), and reach a SRI score of 20-25 alleles, and a total number of unique MHC alleles that varies between 29 and 39 alleles. The SRI score of proteasome and TAP is close to 1 (bottom solid lines), and is not distinguishable from drift (the three largely overlapping dotted lines at the bottom).

the other steps of the pathway towards monomorphism (results not shown). Taken together, these results suggest that, under the condition that the host population maintains a coevolved pathway, one obtains a polymorphism only in the most specific step of the pathway. This confirms our prediction of Chapter 2, where we expected that at least one of the steps of the Ag presentation pathway needed to evolve a polymorphism to protect the pathway from adapting pathogens, and that it would be most beneficial if the polymorphic step was the MHC.

5.4.4 *Factors influencing the shape of the Ag presentation pathway*

The previous section demonstrates that the current state of the human Ag presentation pathway (two monomorphic steps, and one polymorphic step) can evolve from a monomorphic Ag presentation pathway, if new alleles remain coevolved. However, when we replace the algorithm with other mutation algorithms, the monomorphism of proteasome and TAP becomes conditional. In Section 5.4.2 we created new alleles by randomly 'breaking down' a completely generic allele until it reached the required specificity (see Methods),

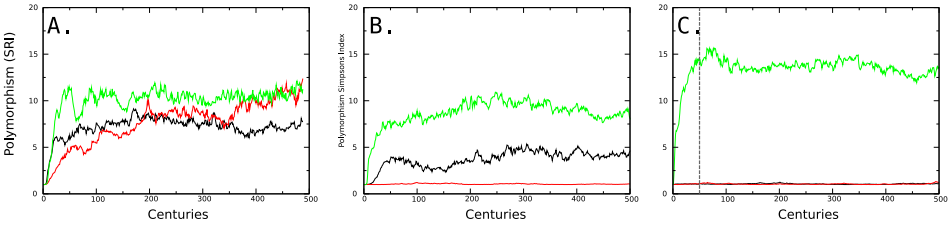

Figure 5.5 Starting conditions influence the evolution of polymorphism in proteasome (black lines), TAP (red lines) or the MHC (green lines). Three different initial conditions; (A) non-coevolved monomorphic start, (B) coevolved monomorphic start, and (C) coevolved, polymorphic MHC start, consistently result in three different quasi-steady states for the polymorphism of the Ag presentation pathway. The degree of polymorphism is expressed in the Simpsons Reciprocal Index score (y-axis). The coevolved, polymorphic MHC start (C) was initialized in the same way as Fig. 5.4, but after 50 centuries (vertical striped line) the mutation algorithm was switched to the one used in (A) and (B).

which resulted in large amount of variation in the recognized peptide pools between the alleles. Using this mutation algorithm in a host population where the whole Ag presentation pathway could evolve would consistently result in a weak polymorphism in all three steps of the pathway. For both the enforced-coevolvedness, and the breaking-down algorithm, the results remained consistent for different host mutation rates, pathogen mutation and reversion rates, the initial conditions of the Ag presentation pathway.

Next we use moderate mutation algorithm, in which new alleles are generated by randomizing the number and identity of the recognized amino acids in one of the nine positions of the parent allele (see Fig. 5.2), and then corrected for a change in pre-defined specificity by repeatedly adding or removing amino acids from any of the positions until the required specificity was approached. The biological interpretation of this algorithm is that a mutation changes one of the key positions in proteasome, TAP or MHC alleles radically, and the structure of the whole molecule changes slightly to accomodate for this change. The algorithm does not artificially force co-evolvedness between the different steps in the Ag presentation pathway, but neither does it generate mutated alleles that are so radically different from the original allele that the host population cannot maintain a coevolved pathway (i.e. the system crosses the information threshold (Nowak and Schuster, 1989)). With the moderate mutation algorithm, the monomorphism of proteasome and TAP depends on the starting conditions of the Ag presentation pathway.

When the host population starts with a non-coevolved (10–20% coevolvedness), monomorphic Ag presentation pathway, all three steps of the Ag presentation pathway evolved a similar degree of polymorphism (Fig. 5.5A). The host population reached an Ag presentation pathway coevolvedness of 32%, and a degree of polymorphism in all three steps that is typically enough to ensure that all steps of the pathway are heterozygous (Borghans et al., 2004; de Boer

Table 5.2 Characteristics of the simulations presented in Fig. 5.5.

starting conditions:	Fig. 5.5A: non-coevolved	Fig. 5.5B: coevolved	Fig. 5.5C: coevo. & polym. MHC
Host Population:			
average age of the hosts:	15y	18y	19y
expected lifespan of the hosts:	27y	33y	36y
average increase in death rate/year:	113x	80x	54x
Antigen presentation pathway:			
average pathway coevolvedness[c]:	0.32	0.69	0.97
weighted fun. polym. of proteasome[d]:	0.48	0.67	0.98
weighted fun. polym. of tap:	0.77	0.99	0.99
weighted fun. polym. of mhc:	0.21[e]	0.30	0.27
Pathogens:			
average number of epitopes per pathogen:	17	22	26
ave. num. epi. of a pathogen at infection:	24	28	33
ave. num. epi. in a random pathogen:	28	40	46
average pre-adaptation of pathogens[f]:	17%	32%	28%
average viral load of pathogens:	5.2	4.7	4.2

a–f: see Table 5.1.

et al., 2004). The low level of overlap between proteasome alleles, between TAP alleles and between MHC alleles (Table 5.2 first column: 41%, 77% and 21%, respectively) suggests that there is a selection pressure in this population for each step of the pathway to be as functionally heterozygous as possible. Because for proteasome, TAP and MHC both alleles are expressed (Gimelbrant et al., 2007), heterozygous hosts typically have 8 different combinations of a proteasome, TAP and MHC allele to recognize pathogen epitopes with (see Methods), which compensate for the low level of coevolvedness. Pathogens in the host population of Fig. 5.5A only reach a pre-adaptation state of 17% (Table 5.2), i.e. at the time of infection carry escape mutations for 15% of the epitopes that the host could have presented in the wildtype virus. This is lower than in any of the other quasi-steady state solutions. However, host populations with a polymorphism still have the lowest expected lifespan (Table 5.2).

When the host population starts with a coevolved ($\geq$ 90% coevolvedness), monomorphic Ag presentation pathway, a polymorphism evolves in the proteasome (Fig. 5.5B). In contrast with the previous quasi-steady state, the host population maintains a relatively high degree of coevolvedness of 69%. The degree of polymorphism in the proteasome is low, with a SRI score of 4.3 alleles. The functional polymorphism between these proteasomes is less pronounced than in that of the previous quasi-steady state: the overlap in 9-mer repertoire that two proteasomes that are randomly selected from the host population can

recognize, is ± 67% (Table 5.2B, second column). Pathogens reach the highest level of preadaptation to host populations in the quasi-steady state approached in Fig. 5.5B, but the hosts can cope better with these pathogens than the hosts in the earlier described quasi-steady state. The high level of preadaptation is compensated by hosts by having a more coevolved Ag presentation pathway than in the situation of Fig. 5.5A. Duplicate runs with the same starting conditions suggest that in some cases TAP evolves a very limited polymorphism with a SRI score of < 3. Whether this behaviour indicates another quasi-steady state in which the polymorphism of each step of the pathway is inversely related to the specificity of the step, or falls within the variation of this quasi-steady state, is unclear.

The third quasi-steady state (Fig. 5.5C) is similar to the one reached in the simulations where coevolution of the Ag presentation pathway was enforced by the mutation algorithm (Fig. 5.4). The host population starts with a coevolved, and MHC-polymorphic Ag presentation pathway, and is capable of maintaining the MHC polymorphism at a higher level then the host populations in the other two quasi-steady states (SRI score of 13.6 alleles versus 10.6 and 8.8 MHC alleles in Fig. 5.5B and C). The host population also maintains its high level of coevolvedness (97%). Pathogens in these host populations reach an intermediate level of preadaptation, but can be well recognized by the Ag presentation pathways of the hosts. The host populations in this third quasi-steady state have the highest expected lifespan (Table 5.2, third column).

Summarizing, changing the way hosts evolve new alleles affects the quasi-steady states that are approached in the model, and the initial starting conditions of the Ag presentation pathway determine in which basin of attraction the simulation starts.

5.5 DISCUSSION

Because pathogens do not only adapt to the MHC alleles of the classical antigen presentation (Ag) pathway, but also to the proteasome and TAP (see Chapter 2, Chapter 4, (Brander et al., 1999; Yokomaku et al., 2004; Kwun et al., 2007)), all of the steps in the Ag presentation pathway are expected to evolve a polymorphism, under the selection pressures of the heterozygote advantage (HA) (Doherty and Zinkernagel, 1975; Carrington et al., 1999) and the rare allele advantage (RAA) (Slade and McCallum, 1992; Langefors et al., 2001; Borghans et al., 2004; de Boer et al., 2004). The fact that no functional polymorphism has been reported for either proteasome or TAP suggests that there are selection pressures on these molecules to remain monomorphic. To understand what this selection pressure could be, we build an agent-based model in which a host population faces several endemic chronic pathogens.

In a setting where the Ag presentation pathway was reduced to a single step, and in which coevolvedness therefore played no role (Fig. 5.3), the proteasome, TAP, and MHC each individually responsed to the endemic pathogens by evolving a polymorphism. However, in settings where all three steps could

simultaneously evolve, and therefore affect the coevolvedness of the Ag presentation pathway, only the MHC alleles became polymorphic (Fig. 5.4, Fig. 5.5C). These results indicate that, under the condition that the host population are capable of maintaining a high degree of coevolvedness, it becomes a strong selection pressure to only let the MHC become polymorphic, counterbalancing the selection pressure for polymorphism by the HA and RAA on the proteasome and TAP.

A much simpler explanation of why the proteasome and TAP molecules are monomorphic would be that they limited by biochemical constraints, and cannot form a polymorphism. However, existing variation in proteasome and TAP molecules makes this explanation unlikely. Within every human host, there are several different proteasome variants; aside from the constituitive proteasome, there is the immunoproteasome, the thymoproteasome (Murata et al., 2008), as well as a hypothesized testis-specific proteasome (Tanaka, 2009). All of these are confirmed, or speculated, to have different cleavage patterns, and thus present different CTL epitopes (Craiu et al., 1997; Rock et al., 2002). Furthermore, in several frog species of the genus *Xenopus*, a functional proteasome polymorphism has been reported (Nonaka et al., 2000). The TAP dimer is an ABC transporter, i.e. is part of a broad class of membrane bound peptide transporters with different specificities within the human genome (Hollenstein et al., 2007; Procko and Gaudet, 2009), and has been reported to be polymorphic in several animal species (Heemels et al., 1993; Gubler et al., 1998; Sironi et al., 2008; Jensen et al., 2008). Therefore there appear to be no biochemical constrains on proteasome and TAP that prohibit a functional polymorphism.

The evolution of polymorphism in the Ag presentation pathway in our model simulations proved to be sensitive to the way in which mutations in proteasome, TAP or MHC alleles affected their phenotype (see section 5.4.4). In our simulations, all the three steps of the pathway were implemented as pattern filters that were subjected to the same algorithm mutating the phenotype of these alleles. In reality, the three steps of the pathway are radically different molecules that are either cleaving, transporting or binding peptides. Therefore, it could be that the effect of mutations on the phenotype of mutant alleles differs between the three molecules, which would in turn affect the degree of polymorphism of each molecule, and possibly the evolution of polymorphism in the human Ag presentation pathway.

Interestingly, with the last mutation algorithm used in the results of Fig. 5.5, the quasi-steady state that the Ag presentation pathway approaches is heavily dependent on the initial Ag presentation pathway. The degree of coevolvedness of the Ag presentation pathway at the start of the simulation (Fig. 5.5A, B) determined whether all three steps of the pathway would acquire a similar polymorphism, or the level of polymorphism would inversely correlate to the specificity of the molecule. Furthermore, only an Ag presentation pathway that started with a limited MHC polymorphism could keep both the proteasome and TAP monomorphic. In a review on the evolution of MHC class I proteins, Lawlor et. al. (Lawlor et al., 1990) suggests that the class I pathway evolved

from class II molecules, but also comments that the evidence is scarce and that little is known about the original circumstances under which the MHC class I pathway evolved. Nevertheless, it opens up the possibility that the classical Ag presentation pathway might have started its evolution with a MHC polymorphism originating from polymorphic MHC class II alleles.

In conclusion, in a host population where pathogens can adapt to the proteasome, TAP and MHC class I molecules, all three steps of the Ag presentation pathway are capable of evolving a polymorphism in response to the pathogen selection pressure. However, under the condition that the host population maintains a high degree of coevolvedness between the steps of the Ag presentation pathway, we found that it is always, and only, the most specific step (i.e. MHC) that becomes polymorphic.

5.6 ACKNOWLEDGEMENTS

The authors would like to thank Henk-Jan van den Ham for his comments on the manuscript, and Rich Hickey, Chris Houser, and Timothy Pratley from the Clojure group for their assistance in designing the parallel core of the simulation model. This study was financially supported by the Netherlands Organisation for Scientific Research (NWO) (VICI grant 016.048.603, Open Program 812.07.003), and a High Potential grant (2007) from the Utrecht University.

General Discussion

6

6.1 FINDINGS PRESENTED IN THIS THESIS

In this thesis we address the evolution of the classical antigen (Ag) presentation pathway, both from a proximate and an ultimate point of view. In order to do so, we made thankful use of the public Los Alamos HIV-1 sequence data base and of prediction algorithms for the individual steps of the Ag presentation pathway. By studying how a virus like HIV-1 changes and adapts to the Ag presentation pathway, we learned of the selection pressures that the virus exerts on the pathway, and visa versa. The newfound understandings that flowed from these initial studies allowed us to build simulation models of an evolving human population, in which we could study the evolution of the Ag presentation pathway.

6.1.1 *Proximate findings*

Summarizing, in this thesis we have studied the recent evolution of HIV-1 (Chapter 2), the possibility that HIV-1 had already adapted to CTL epitopes in the more distant past (Chapter 3), and the effect of an MHC polymorphism on the adaptation of a HIV-1 like virus to the Ag presentation pathway (Chapter 4). These three lines of research provided us with an answer to the first half of the double-sided question that we posed in the introduction:

"Why do pathogens not adapt to the monomorphic proteasome and TAP?"

In Chapter 2 we showed that the total number of both CTL epitopes and epitope precursors in HIV-1 has remained more or less constant in the last thirty years. Due to the high specificity and the high degree of polymorphism of the MHC class I molecules, the epitope precursors of HIV-1 are not constantly under selection pressure. In an average host, only 18% of the epitope precursors can actually bind to the MHC molecules of that host, leaving 82% of the epitope precursors free to revert possibly costly escape mutations that were acquired in previous hosts. This *intermittent exposure* of epitope precursors to immune selection pressure limits the level of adaptation that viruses can reach to the monomorphic components of the Ag presentation pathway.

But how much of the adaptation of HIV-1 to its new human host had already happened prior to the 1980's? Yusim et al. (2002) proposed that the past adaptation of HIV-1 had left a 'footprint' in the distribution of CTL epitopes in current-day HIV-1 sequences, and hypothesized that large-scale CTL epitope adaptations had predominantly occurred in the variable regions of the ances-

tral HIV-1 sequence. The epitope-poor and epitope-rich regions in HIV-1 were hypothesized to be the result of this process of localized adaptation. Although this hypothesis was compelling, a detailed study of the distribution of CTL epitopes in HIV-1 sequences revealed that these distributions were not discriminable from random (Chapter 3), and that the variation in the epitope distributions in HIV-1 was comparable to that of other organisms like humans, yeast or fruitflies. Based on these findings, we discarded the hypothesis that there is evidence for large-scale adaptation of HIV-1 to the human host in the apparent clustering of its CTL epitopes. The question how much the ancestral HIV-1 had adapted to the human host prior to the 1980's remains open, but it might be less than previously appreciated.

To understand how the MHC polymorphism affects the potential for pathogens like HIV-1 to adapt to a new host population, we constructed a host-pathogen interaction model (Chapter 4). Varying the MHC polymorphism of the host population, we discovered how an increase in MHC polymorphism of the population steered pathogens towards adapting to the monomorphic components of the pathway. However, at the same time an increase in MHC polymorphism strengthened the intermittent exposure effect by increasing the number of epitope precursors that were intermittently under selection pressure at the population level, and thus the average time between exposures for individual epitope precursors. Our prediction from Chapter 2 turned out to be correct: a single polymorphic step in the pathway could indeed prevent pathogens from exploiting the monomorphic properties of the other steps. Taken together, the answer to the above question is that pathogens *do* adapt to the monomorphic proteasome and TAP, but that due to the MHC polymorphism, there are few, if any, negative consequences of this adaptation for the host population.

6.1.2 *Ultimate findings*

In the fifth chapter of this thesis we explored how the Ag presentation pathway of a host population could have evolved into its current shape. In Chapter 2 we had already postulated that the intermittent exposure effect would 'work best' if it was the most specific step of the pathway (i.e. MHC) that was polymorphic. However, we had no answer as to why the other two steps of the pathway would remain monomorphic. Just like MHC class I molecules, proteasome and TAP were under the selection pressures of the heterozygote advantage (HA) and rare allele advantage (RAA) to become polymorphic.

"Why did only the MHC become polymorphic?"

We constructed a simulation model in which all three steps of the Ag presentation pathway were capable of evolving a polymorphism in response to the selection pressure exerted by pathogens. We found that, under the condition that the host population maintains a coevolved Ag presentation pathway, that it is

always and *only*, the most specific step that becomes polymorphic. When we removed the condition of coevolvedness, and the alleles of the pathway were evolving fully unconstrained, additional quasi-steady states existed and the initial conditions of the Ag presentation pathway would determine which of these quasi-steady states the host population approached. Interestingly enough, multiple quasi-steady states appear to exist in the animal kingdom as well.

We therefore have two possible answers to our question. Based on certain restrictions, an Ag presentation pathway with only a polymorphic MHC is strongly selected for, and gives a higher fitness to the host population than that other configurations of the Ag presentation pathway do. However, the existence of alternative quasi-steady states in the animal kingdom suggests that our current Ag presentation pathway might just be the coincidental end-result of past conditions in our evolutionary history, and not necessarily the result of selection pressure from our current environment.

6.2 OUTLOOK

"At the end of my paper, there is always some research left."[1]

The process of researching seems to generate not only publications, but also a battlefield of interesting results and promising sidetracks that have not been further pursued. Here we take the opportunity to discuss two of them: the role of specificity in the antigen presentation pathway, and the effect that superinfection might have on the adaptation of HIV-1 to the human host.

6.2.1 *Specificity of the proteasome, TAP and MHC class I*

We addressed the evolution of the polymorphism of the antigen presentation pathway in the 4th and 5th chapter of this thesis, but did not address the evolution of its specificity. As mentioned in the introduction, each of the steps has its own specificity, and the complete pathway can only present a small percentage of the peptide fragments in a protein. Why is our antigen presentation pathway not far more generic? A greater variety of presented CTL epitopes would increase the chance of generating CTL responses that can control, or even clear an infection.

The hosts in our simulation models (chapter 4 and 5) would, if given the opportunity, rapidly evolve an antigen presentation pathway that can present close to a 100% of the peptides in a protein. That the human antigen presentation pathway has not evolved to present every possible peptide fragment, suggests that there is a negative effect of presenting too many different CTL epitopes.

Intuitively, one possible explanation would be that a too generic antigen presentation pathway would limit the T cell repertoire. As mentioned in the introduction, developing T cells undergo positive and negative selection in the

[1] rephrased from a famous Loesje

thymus. A more generic antigen presentation pathway would also increase the repertoire of presented self-peptides, and as a consequence the number of T cells lost by negative selection. This explanation is similar to the one forwarded by Nowak et al. (1992) for the limited number of MHC loci in human cells. An increase in the number of MHC loci (e.g. HLA-A, HLA-B etc) also increases the number of self-peptides presented in the thymus, and thus increase the impact of negative selection. However, Borghans et al. (2003), showed that positive selection is the largest bottleneck during T cell development, and that increasing the number of MHC loci would actually increase the T cell repertoire.

Whilst increasing the number of MHC loci increases the number of developing T cells that survive positive selection, it is not known what the effect of a generic MHC allele would be on positive selection. If a generic MHC molecule only affects negative selection by increasing the repertoire of presented self-peptides, then increased deletion of developing T cells could prohibit a too generic antigen presentation pathway.

Alternatively, hosts that present many different CTL epitopes might have a larger chance of generating an auto-immune response. With every CTL response against a foreign peptide, there is a small chance that the responding T-cell population is cross-reactive with a self-peptide. Normally, these T cells would have been negatively selected during their development in the thymus, but this process is not perfect (Gallegos and Bevan, 2006), and some self-reactive naive T cell might escape negative selection (Huseby et al., 2001). Increasing the number of self and foreign epitopes might therefore also increase the chance on auto-immune diseases.

A third option is that with a more generic antigen presentation pathway, the frequency of each particular MHC-epitope complex on the cell surface will be diluted. Human cells present an estimated 50,000 - 100,000 MHC class I molecules on their cell surface (Yewdell et al., 2003), and increasing the repertoire of presented CTL epitopes could lower the frequency of a particular epitope on the cell surface, and with it the strength of the T cell response (van den Berg and Rand, 2004; Goldwich et al., 2008)

All three options could well be implemented in models such as the ones used in this thesis, and explored for their ability to limit the specificity of the antigen presentation pathway.

6.2.2 *HIV-1 superinfection*

The simulation models in this thesis are simplified representations of reality, and help us understand how host and pathogens influence the selection pressures that act on them. One of the challenges in constructing such models lies in the choices we make on which processes to include, and which to neglect (e.g. do we give pathogens a wildtype, or let them evolve with little or no restrictions, how do we implement the Ag presentation pathway, is it neccesary to include a sexual contact network).

One of the choises we made was to not include the ability of HIV-1 variants to superinfect HIV-1 patients. The ability of HIV-1 to superinfect has been described since the discovery of recombinant strains (Burke, 1997), but was initially considered a rare event. With the increased availability of longitudinal HIV-1 sequence data, more cases of superinfection have been reported. However, the reported incidence vary widely between different studies, from virtually non-existent to as common as primary infections (van der Kuyl and Cornelissen, 2007; Piantadosi et al., 2008; Willberg et al., 2008; Sidat et al., 2008)

The risk associated with HIV-1 superinfection is that of a more rapid progression to AIDS for the superinfected patient (Sidat et al., 2008; Streeck et al., 2008). However, if superinfection is relatively common, it will also have a large effect on the level of adaptation to the human population that HIV-1 can reach. The practice of 'sero-sorting', in which two HIV-1 infected patients have unprotected sex, might therefore put more people at an increased risk than just themselves. Superinfection generates an additional selection pressure on the virus. In a population with a high incidence of superinfection, the virus is not only under selection to adapt to the CTL responses of its current host, but also selected to be able to invade other HIV-1 infected patients, and outcompete the local virus. HIV-1 variants that can do so, have a larger pool of susceptible hosts, and will therefore spread through the population.

Without superinfection, escape mutations in epitopes that are not targeted in the current host have no selection pressure on them to be maintained. With superinfection, there is a higher selection pressure to maintain escape mutations, because the superinfecting virus has to compete with a locally adapted virus. This would decrease the intermittent exposure effect, and thus allow for an increased adaptation of the virus to the human Ag presentation pathway.

Bibliography

Allen, T. M., Altfeld, M., Yu, X. G., O'Sullivan, K. M., Lichterfeld, M., Gall, S. L., John, M., Mothe, B. R., Lee, P. K., Kalife, E. T., Cohen, D. E., Freedberg, K. A., Strick, D. A., Johnston, M. N., Sette, A., Rosenberg, E. S., Mallal, S. A., Goulder, P. J. R., Brander, C., and Walker, B. D. (2004). Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *J Virol*, 78(13):7069–7078. (Cited on page 10.)

Allen, T. M., Yu, X. G., Kalife, E. T., Reyor, L. L., Lichterfeld, M., John, M., Cheng, M., Allgaier, R. L., Mui, S., Frahm, N., Alter, G., Brown, N. V., Johnston, M. N., Rosenberg, E. S., Mallal, S. A., Brander, C., Walker, B. D., and Altfeld, M. (2005). De novo generation of escape variant-specific cd8+ t-cell responses following cytotoxic t-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. *J Virol*, 79(20):12952–12960. (Cited on page 20.)

Althaus, C. L. and de Boer, R. J. (2008). Dynamics of immune escape during HIV/SIV infection. *PLoS Comput Biol*, 4(7):e1000103. (Cited on pages 72 and 74.)

Alvarado-Guerri, R., Cabrera, C. M., Garrido, F., and López-Nevot, M. A. (2005). TAP1 and TAP2 polymorphisms and their linkage disequilibrium with HLA-DR, -DP, and -DQ in an eastern Andalusian population. *Hum Immunol*, 66(8):921–930. (Cited on pages 6 and 68.)

Andersen-Nissen, E., Smith, K. D., Strobe, K. L., Barrett, S. L. R., Cookson, B. T., Logan, S. M., and Aderem, A. (2005). Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc Natl Acad Sci U S A*, 102(26):9247–9252. (Cited on page 2.)

Anzai, T., Shiina, T., Kimura, N., Yanagiya, K., Kohara, S., Shigenari, A., Yamagata, T., Kulski, J. K., Naruse, T. K., Fujimori, Y., Fukuzumi, Y., Yamazaki, M., Tashiro, H., Iwamoto, C., Umehara, Y., Imanishi, T., Meyer, A., Ikeo, K., Gojobori, T., Bahram, S., and Inoko, H. (2003). Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc Natl Acad Sci U S A*, 100(13):7708–7713. (Cited on page 64.)

Asquith, B., Edwards, C. T. T., Lipsitch, M., and McLean, A. R. (2006). Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. *PLoS Biol*, 4(4):e90. (Cited on pages 16, 54, and 72.)

Assarsson, E., Bui, H.-H., Sidney, J., Zhang, Q., Glenn, J., Oseroff, C., Mbawuike, I. N., Alexander, J., Newman, M. J., Grey, H., and Sette, A. (2008).

Immunomic analysis of the repertoire of T-cell specificities for influenza A virus in humans. *J Virol*, 82(24):12241–12251. (Cited on pages 39 and 40.)

Assarsson, E., Sidney, J., Oseroff, C., Pasquetto, V., Bui, H.-H., Frahm, N., Brander, C., Peters, B., Grey, H., and Sette, A. (2007). A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. *J Immunol*, 178(12):7890–7901. (Cited on pages 10, 11, 22, 33, and 57.)

Barouch, D. H., Powers, J., Truitt, D. M., Kishko, M. G., Arthur, J. C., Peyerl, F. W., Kuroda, M. J., Gorgone, D. A., Lifton, M. A., Lord, C. I., Hirsch, V. M., Montefiori, D. C., Carville, A., Mansfield, K. G., Kunstman, K. J., Wolinsky, S. M., and Letvin, N. L. (2005). Dynamic immune responses maintain cytotoxic T lymphocyte epitope mutations in transmitted simian immunodeficiency virus variants. *Nat Immunol*, 6(3):247–252. (Cited on pages 20 and 58.)

Baugh, L. L., Garcia, J. V., and Foster, J. L. (2008). Functional characterization of the human immunodeficiency virus type 1 nef acidic domain. *J Virol*, 82(19):9657–9667. (Cited on page 24.)

Beatson, S. A., Minamino, T., and Pallen, M. J. (2006). Variation in bacterial flagellins: from sequence to structure. *Trends Microbiol*, 14(4):151–155. (Cited on page 2.)

Beekman, N. J., van Veelen, P. A., van Hall, T., Neisig, A., Sijts, A., Camps, M., Kloetzel, P. M., Neefjes, J. J., Melief, C. J., and Ossendorp, F. (2000). Abrogation of CTL epitope processing by single amino acid substitution flanking the C-terminal proteasome cleavage site. *J Immunol*, 164(4):1898–1905. (Cited on page 10.)

Bergmann, C. C., Tong, L., Cua, R., Sensintaffar, J., and Stohlman, S. (1994). Differential effects of flanking residues on presentation of epitopes from chimeric peptides. *J Virol*, 68(8):5306–5310. (Cited on page 10.)

Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., Mallal, S., Mullins, J. I., Nickle, D. C., Herbeck, J., Rousseau, C., Learn, G. H., Miura, T., Brander, C., Walker, B., and Korber, B. (2007). Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, 315(5818):1583–1586. (Cited on pages 15, 32, and 48.)

Bodmer, W. F. (1972). Evolutionary significance of the HL-A system. *Nature*, 237(5351):139–45 passim. (Cited on page 10.)

Boehmer, H. (2008). Positive and negative selection in Basel. essay. *Nature Immunology*, 9(2). (Cited on page 3.)

Borghans, J. A. M., Beltman, J. B., and de Boer, R. J. (2004). MHC polymorphism under host-pathogen coevolution. *Immunogenetics*, 55(11):732–739. (Cited on pages 6, 10, 15, 52, 68, 74, 79, and 81.)

Borghans, J. A. M., Noest, A. J., and de Boer, R. J. (2003). Thymic selection does not limit the individual MHC diversity. *Eur J Immunol*, 33(12):3353–3358. (Cited on pages 3 and 88.)

Box, J. F. (1980). R.A. Fisher and the Design of Experiments, 1922-1926. *The American Statistician*, 1(1):1–7. (Cited on page 35.)

Brander, C., Frahm, N., and Walker, B. D. (2006). The challenges of host and viral diversity in HIV vaccine design. *Curr Opin Immunol*, 18(4):430–437. (Cited on page 5.)

Brander, C., Yang, O. O., Jones, N. G., Lee, Y., Goulder, P., Johnson, R. P., Trocha, A., Colbert, D., Hay, C., Buchbinder, S., Bergmann, C. C., Zweerink, H. J., Wolinsky, S., Blattner, W. A., Kalams, S. A., and Walker, B. D. (1999). Efficient processing of the immunodominant, HLA-A*0201-restricted human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitope despite multiple variations in the epitope flanking sequences. *J Virol*, 73(12):10191–10198. (Cited on pages 7, 18, 61, 63, 74, and 81.)

Brumme, Z. L., Brumme, C. J., Heckerman, D., Korber, B. T., Daniels, M., Carlson, J., Kadie, C., Bhattacharya, T., Chui, C., Szinger, J., Mo, T., Hogg, R. S., Montaner, J. S. G., Frahm, N., Brander, C., Walker, B. D., and Harrigan, P. R. (2007). Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of hiv-1. *PLoS Pathog*, 3(7):e94. (Cited on pages 15, 24, 32, and 48.)

Burke, D. S. (1997). Recombination in HIV: an important viral evolutionary strategy. *Emerg Infect Dis*, 3(3):253–259. (Cited on page 89.)

Burroughs, N. J., de Boer, R. J., and Keşmir, C. (2004). Discriminating self from nonself with short peptides from large proteomes. *Immunogenetics*, 56(5):311–320. (Cited on pages 4, 11, 33, 38, 44, 57, 74, and 76.)

Carnes, B. A., Holden, L. R., Olshansky, S. J., Witten, M. T., and Siegel, J. S. (2006). Mortality partitions and their relevance to research on senescence. *Biogerontology*, 7(4):183–198. (Cited on pages 53, 56, 57, and 71.)

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., and O'Brien, S. J. (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science*, 283(5408):1748–1752. (Cited on pages 6, 16, 52, 54, 68, and 81.)

Carrington, M. and O'Brien, S. J. (2003). The influence of hla genotype on aids. *Annu Rev Med*, 54:535–551. (Cited on page 61.)

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655. (Cited on page 2.)

Cascio, P., Hilton, C., Kisselev, A. F., Rock, K. L., and Goldberg, A. L. (2001). 26S proteasomes and immunoproteasomes produce mainly N-extended versions of an antigenic peptide. *EMBO J*, 20(10):2357–2366. (Cited on page 34.)

Chakraborty, H., Sen, P. K., Helms, R. W., Vernazza, P. L., Fiscus, S. A., Eron, J. J., Patterson, B. K., Coombs, R. W., Krieger, J. N., and Cohen, M. S. (2001). Viral burden in genital secretions determines male-to-female sexual transmission of HIV-1: a probabilistic empiric model. *AIDS*, 15(5):621–627. (Cited on pages 54, 55, and 56.)

Chen, T. K. and Aldrovandi, G. M. (2008). Review of HIV antiretroviral drug resistance. *Pediatr Infect Dis J*, 27(8):749–752. (Cited on page 31.)

Chopera, D. R., Woodman, Z., Mlisana, K., Mlotshwa, M., Martin, D. P., Seoighe, C., Treurnicht, F., de Rosa, D. A., Hide, W., Karim, S. A., Gray, C. M., Williamson, C., and Team, C. A. P. R. I. S. A. . S. (2008). Transmission of HIV-1 CTL escape variants provides HLA-mismatched recipients with a survival advantage. *PLoS Pathog*, 4(3):e1000033. (Cited on page 10.)

Costin, J. M. (2007). Cytopathic mechanisms of HIV-1. *Virol J*, 4:100. (Cited on page 57.)

Craiu, A., Akopian, T., Goldberg, A., and Rock, K. L. (1997). Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proc Natl Acad Sci U S A*, 94(20):10850–10855. (Cited on pages 4, 6, 10, 34, and 82.)

Culmann, B., Gomard, E., Kiny, M. P., Guy, B., Dreyfus, F., Saimot, A. G., Sereni, D., Sicard, D., and Lvy, J. P. (1991). Six epitopes reacting with human cytotoxic cd8+ t cells in the central region of the hiv-1 nef protein. *J Immunol*, 146(5):1560–1565. (Cited on page 37.)

Culmann-Penciolelli, B., Lamhamedi-Cherradi, S., Couillin, I., Guegan, N., Levy, J. P., Guillet, J. G., and Gomard, E. (1994). Identification of multirestricted immunodominant regions recognized by cytolytic t lymphocytes in the human immunodeficiency virus type 1 nef protein. *J Virol*, 68(11):7336–7343. (Cited on page 37.)

Danchin, E., Vitiello, V., Vienne, A., Richard, O., Gouret, P., McDermott, M. F., and Pontarotti, P. (2004). The major histocompatibility complex origin. *Immunol Rev*, 198:216–232. (Cited on page 7.)

Davenport, M. P., Loh, L., Petravic, J., and Kent, S. J. (2008). Rates of HIV immune escape and reversion: implications for vaccination. *Trends Microbiol*, 16(12):561–566. (Cited on page 74.)

de Boer, R. J., Borghans, J. A. M., van Boven, M., Keşmir, C., and Weissing, F. J. (2004). Heterozygote advantage fails to explain the high degree of polymorphism of the MHC. *Immunogenetics*, 55(11):725–731. (Cited on pages 6, 68, 79, and 81.)

Doherty, P. C. and Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 256(5512):50–52. (Cited on pages 6, 52, 54, 68, and 81.)

Doytchinova, I. A., Guan, P., and Flower, D. R. (2006). EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics*, 7:131. (Cited on pages 10, 11, and 33.)

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9):755–763. (Cited on page 14.)

Elias, P. M. (2007). The skin barrier as an innate immune element. *Semin Immunopathol*, 29(1):3–14. (Cited on page 1.)

Faucz, F. R., Probst, C. M., and Petzl-Erler, M. L. (2000). Polymorphism of LMP2, TAP1, LMP7 and TAP2 in Brazilian Amerindians and Caucasoids: implications for the evolution of allelic and haplotypic diversity. *Eur J Immunogenet*, 27(1):5–16. (Cited on pages 6 and 68.)

Fenner, J. N. (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol*, 128(2):415–423. (Cited on page 1.)

Finlay, B. B. and McFadden, G. (2006). Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*, 124(4):767–782. (Cited on page 2.)

Fisher, R. (1935). The design of experiments. *New York: Hafner Publishing*, (1). (Cited on page 35.)

Fortier, M.-H., Caron, E., Hardy, M.-P., Voisin, G., Lemieux, S., Perreault, C., and Thibault, P. (2008). The MHC class I peptide repertoire is molded by the transcriptome. *J Exp Med*, 205(3):595–610. (Cited on pages 13, 33, and 40.)

Frahm, N., Yusim, K., Suscovich, T. J., Adams, S., Sidney, J., Hraber, P., Hewitt, H. S., Linde, C. H., Kavanagh, D. G., Woodberry, T., Henry, L. M., Faircloth, K., Listgarten, J., Kadie, C., Jojic, N., Sango, K., Brown, N. V., Pae, E., Zaman, M. T., Bihl, F., Khatri, A., John, M., Mallal, S., Marincola, F. M., Walker, B. D., Sette, A., Heckerman, D., Korber, B. T., and Brander, C. (2007). Extensive HLA class I allele promiscuity among viral CTL epitopes. *Eur J Immunol*, 37(9):2419–2433. (Cited on pages 38 and 63.)

Frankild, S., de Boer, R. J., Lund, O., Nielsen, M., and Kesmir, C. (2008). Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. *PLoS ONE*, 3(3):e1831. (Cited on page 17.)

Frater, A. J., Brown, H., Oxenius, A., Gnthard, H. F., Hirschel, B., Robinson, N., Leslie, A. J., Payne, R., Crawford, H., Prendergast, A., Brander, C., Kiepiela, P., Walker, B. D., Goulder, P. J. R., McLean, A., and Phillips, R. E. (2007). Effective T-cell responses select human immunodeficiency virus mutants and slow disease progression. *J Virol*, 81(12):6742–6751. (Cited on pages 16 and 17.)

Friedrich, T. C., Dodds, E. J., Yant, L. J., Vojnov, L., Rudersdorf, R., Cullen, C., Evans, D. T., Desrosiers, R. C., Moth, B. R., Sidney, J., Sette, A., Kunstman, K., Wolinsky, S., Piatak, M., Lifson, J., Hughes, A. L., Wilson, N., O'Connor, D. H., and Watkins, D. I. (2004). Reversion of CTL escape-variant immunodeficiency viruses in vivo. *Nat Med*, 10(3):275–281. (Cited on page 58.)

Furutsuki, T., Hosoya, N., Kawana-Tachikawa, A., Tomizawa, M., Odawara, T., Goto, M., Kitamura, Y., Nakamura, T., Kelleher, A. D., Cooper, D. A., and Iwamoto, A. (2004). Frequent transmission of cytotoxic-T-lymphocyte escape mutants of human immunodeficiency virus type 1 in the highly HLA-A24-positive japanese population. *J Virol*, 78(16):8437–8445. (Cited on page 16.)

Gaillard, J.-M., Yoccoz, N. G., Lebreton, J.-D., Bonenfant, C., Devillard, S., Loison, A., Pontier, D., and Allaine, D. (2005). Generation time: a reliable metric to measure life-history variation among mammalian populations. *Am Nat*, 166(1):119–23; discussion 124–8. (Cited on page 1.)

Gallegos, A. M. and Bevan, M. J. (2006). Central tolerance: good but imperfect. *Immunol Rev*, 209:290–296. (Cited on page 88.)

Gimelbrant, A., Hutchinson, J. N., Thompson, B. R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140. (Cited on page 80.)

Goldwich, A., Hahn, S. S. C., Schreiber, S., Meier, S., Kmpgen, E., Wagner, R., Lutz, M. B., and Schubert, U. (2008). Targeting HIV-1 Gag into the defective ribosomal product pathway enhances MHC class I antigen presentation and CD8+ T cell activation. *J Immunol*, 180(1):372–382. (Cited on page 88.)

Gomez, L. M., Camargo, J. F., Castiblanco, J., Ruiz-Narvez, E. A., Cadena, J., and Anaya, J. M. (2006). Analysis of IL1B, TAP1, TAP2 and IKBL polymorphisms on susceptibility to tuberculosis. *Tissue Antigens*, 67(4):290–296. (Cited on pages 6 and 68.)

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. *Philosophical Transactions of the Royal Society of London*. (Cited on page 53.)

Goulder, P. J., Brander, C., Tang, Y., Tremblay, C., Colbert, R. A., Addo, M. M., Rosenberg, E. S., Nguyen, T., Allen, R., Trocha, A., Altfeld, M., He, S., Bunce, M., Funkhouser, R., Pelton, S. I., Burchett, S. K., McIntosh, K., Korber, B. T., and Walker, B. D. (2001). Evolution and transmission of stable CTL escape

mutations in HIV infection. *Nature*, 412(6844):334–338. (Cited on pages 10 and 36.)

Groothuis, T. A. M., Griekspoor, A. C., Neijssen, J. J., Herberts, C. A., and Neefjes, J. J. (2005). MHC class I alleles and their exploration of the antigen-processing machinery. *Immunol Rev*, 207:60–76. (Cited on pages 10 and 54.)

Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S., and van Endert, P. M. (1998). Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol Immunol*, 35(8):427–433. (Cited on pages 6, 23, and 82.)

Hakenberg, J., Nussbaum, A. K., Schild, H., Rammensee, H.-G., Kuttler, C., Holzhtter, H.-G., Kloetzel, P.-M., Kaufmann, S. H. E., and Mollenkopf, H.-J. (2003). MAPPP: MHC class I antigenic peptide processing prediction. *Appl Bioinformatics*, 2(3):155–158. (Cited on pages 11 and 32.)

Hallén, A. (2007). Gompertz law and aging as exclusion effects. *Biogerontology*, 8(5):605–612. (Cited on page 53.)

Heemels, M. T., Schumacher, T. N., Wonigeit, K., and Ploegh, H. L. (1993). Peptide translocation by variants of the transporter associated with antigen processing. *Science*, 262(5142):2059–2063. (Cited on pages 6 and 82.)

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends Microbiol*, 8(11):504–508. (Cited on page 1.)

Herbeck, J. T., Gottlieb, G. S., Li, X., Hu, Z., Detels, R., Phair, J., Rinaldo, C., Jacobson, L. P., Margolick, J. B., and Mullins, J. I. (2008). Lack of evidence for changing virulence of HIV-1 in North America. *PLoS ONE*, 3(2):e1525. (Cited on pages 15, 32, and 48.)

Herbeck, J. T., Nickle, D. C., Learn, G. H., Gottlieb, G. S., Curlin, M. E., Heath, L., and Mullins, J. I. (2006). Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. *J Virol*, 80(4):1637–1644. (Cited on page 58.)

Holland, J. and Domingo, E. (1998). Origin and evolution of viruses. *Virus Genes*, 16(1):13–21. (Cited on page 1.)

Hollenstein, K., Dawson, R. J. P., and Locher, K. P. (2007). Structure and mechanism of ABC transporter proteins. *Curr Opin Struct Biol*, 17(4):412–418. (Cited on page 82.)

Hoof, I., Keşmir, C., Lund, O., and Nielsen, M. (2008). Humans with chimpanzee-like major histocompatibility complex-specificities control HIV-1 infection. *AIDS*, 22(11):1299–1303. (Cited on page 64.)

Hopkins, B. and Skellam, J. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany (London)*, 18(1):213–227. (Cited on pages 34, 35, and 40.)

Huseby, E. S., Sather, B., Huseby, P. G., and Goverman, J. (2001). Age-dependent T cell tolerance and autoimmunity to myelin basic protein. *Immunity*, 14(4):471–481. (Cited on page 88.)

Huttner, K. M. and Bevins, C. L. (1999). Antimicrobial peptides as mediators of epithelial host defense. *Pediatr Res*, 45(6):785–794. (Cited on page 2.)

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314. (Cited on page 14.)

Irbäck, A., Peterson, C., and Potthast, F. (1996). Evidence for nonrandom hydrophobicity structures in protein chains. *Proc Natl Acad Sci U S A*, 93(18):9533–9538. (Cited on pages 44 and 46.)

Jensen, L. F., Hansen, M. M., Mensberg, K.-L. D., and Loeschcke, V. (2008). Spatially and temporally fluctuating selection at non-MHC immune genes: evidence from TAP polymorphism in populations of brown trout (salmo trutta, l.). *Heredity*, 100(1):79–91. (Cited on pages 6 and 82.)

Jin, X., Bauer, D. E., Tuttleton, S. E., Lewin, S., Gettie, A., Blanchard, J., Irwin, C. E., Safrit, J. T., Mittler, J., Weinberger, L., Kostrikis, L. G., Zhang, L., Perelson, A. S., and Ho, D. D. (1999). Dramatic rise in plasma viremia after CD8(+) T cell depletion in simian immunodeficiency virus-infected macaques. *J Exp Med*, 189(6):991–998. (Cited on page 16.)

Karlsson, A. C., Iversen, A. K. N., Chapman, J. M., de Oliviera, T., Spotts, G., McMichael, A. J., Davenport, M. P., Hecht, F. M., and Nixon, D. F. (2007). Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. *PLoS ONE*, 2(2):e225. (Cited on pages 20 and 24.)

Kaslow, R. A., Rivers, C., Tang, J., Bender, T. J., Goepfert, P. A., Habib, R. E., Weinhold, K., Mulligan, M. J., and vaccine evaluation group, N. I. A. I. D. A. (2001). Polymorphisms in HLA class I genes associated with both favorable prognosis of human immunodeficiency virus (HIV) type 1 infection and positive cytotoxic T-lymphocyte responses to ALVAC-HIV recombinant canarypox vaccines. *J Virol*, 75(18):8681–8689. (Cited on pages 13 and 16.)

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J. G., Fidler, S., Weber, J., Pybus, O. G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman,

D., Harrigan, P. R., Walker, B. D., Takiguchi, M., and Goulder, P. (2009). Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*. (Cited on pages 24, 32, 37, 50, 63, and 64.)

Kay, A. and Zoulim, F. (2007). Hepatitis B virus genetic variability and evolution. *Virus Res*, 127(2):164–176. (Cited on page 5.)

Kearney, M., Maldarelli, F., Shao, W., Margolick, J. B., Daar, E. S., Mellors, J. W., Rao, V., Coffin, J. M., and Palmer, S. (2009). Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J Virol*, 83(6):2715–2727. (Cited on pages 36 and 37.)

Keele, B. F., Heuverswyn, F. V., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313(5786):523–526. (Cited on page 64.)

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(Pt. 1-2):81–93. (Cited on page 14.)

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988. (Cited on page 36.)

Kesmir, C., van Noort, V., de Boer, R. J., and Hogeweg, P. (2003). Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. *Immunogenetics*, 55(7):437–449. (Cited on pages 6 and 7.)

Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G. H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J. I., Heckerman, D., Walker, B. D., and Goulder, P. (2007). CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med*, 13(1):46–53. (Cited on pages 43, 55, 57, 69, and 70.)

Kimura, M. (1991). Recent development of the neutral theory viewed from the wrightian tradition of theoretical population genetics. *Proc Natl Acad Sci U S A*, 88(14):5969–5973. (Cited on page 17.)

Klein, M. R., van der Burg, S. H., Hovenkamp, E., Holwerda, A. M., Drijfhout, J. W., Melief, C. J., and Miedema, F. (1998). Characterization of HLA-B57-restricted human immunodeficiency virus type 1 Gag- and RT-specific cytotoxic T lymphocyte responses. *J Gen Virol*, 79 (Pt 9):2191–2201. (Cited on pages 13 and 16.)

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288(5472):1789–1796. (Cited on pages 14, 17, 18, 36, and 40.)

Kuiken, C., Korber, B., and Shafer, R. W. (2003). HIV sequence databases. *AIDS Rev*, 5(1):52–61. (Cited on pages 13 and 14.)

Kutsch, O., Vey, T., Kerkau, T., Hünig, T., and Schimpl, A. (2002). HIV type 1 abrogates TAP-mediated transport of antigenic peptides presented by MHC class I. transporter associated with antigen presentation. *AIDS Res Hum Retroviruses*, 18(17):1319–1325. (Cited on page 24.)

Kwun, H. J., da Silva, S. R., Shah, I. M., Blake, N., Moore, P. S., and Chang, Y. (2007). Kaposi's sarcoma-associated herpesvirus latency-associated nuclear antigen 1 mimics Epstein-Barr virus EBNA1 immune evasion through central repeat domain effects on protein processing. *J Virol*, 81(15):8225–8235. (Cited on pages 52, 68, and 81.)

Landolt-Marticorena, C., Williams, K. A., Deber, C. M., and Reithmeier, R. A. (1993). Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J Mol Biol*, 229(3):602–608. (Cited on page 44.)

Langefors, A., Lohm, J., Grahn, M., Andersen, O., and von Schantz, T. (2001). Association between major histocompatibility complex class IIB alleles and resistance to Aeromonas salmonicida in Atlantic salmon. *Proc Biol Sci*, 268(1466):479–485. (Cited on pages 6, 52, 68, and 81.)

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. (2005). 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. (Cited on pages 11, 33, 54, and 70.)

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, 8:424. (Cited on pages 10, 11, 12, and 14.)

Lavreys, L., Baeten, J. M., Chohan, V., McClelland, R. S., Hassan, W. M., Richardson, B. A., Mandaliya, K., Ndinya-Achola, J. O., and Overbaugh, J. (2006). Higher set point plasma viral load and more-severe acute HIV type 1 (HIV-1) illness predict mortality among high-risk HIV-1-infected African women. *Clin Infect Dis*, 42(9):1333–1339. (Cited on pages 53 and 57.)

Lawlor, D. A., Zemmour, J., Ennis, P. D., and Parham, P. (1990). Evolution of class-i MHC genes and proteins: from natural selection to thymic selection. *Annu Rev Immunol*, 8:23–63. (Cited on pages 7 and 82.)

Lawton, A. P. and Kronenberg, M. (2004). The Third Way: Progress on pathways of antigen processing and presentation by CD1. *Immunol Cell Biol*, 82(3):295–306. (Cited on page 3.)

LeBien, T. W. and Tedder, T. F. (2008). B lymphocytes: how they develop and function. *Blood*, 112(5):1570–1580. (Cited on page 2.)

Leslie, A., Kavanagh, D., Honeyborne, I., Pfafferott, K., Edwards, C., Pillay, T., Hilton, L., Thobakgale, C., Ramduth, D., Draenert, R., Gall, S. L., Luzzi, G., Edwards, A., Brander, C., Sewell, A. K., Moore, S., Mullins, J., Moore, C., Mallal, S., Bhardwaj, N., Yusim, K., Phillips, R., Klenerman, P., Korber, B., Kiepiela, P., Walker, B., and Goulder, P. (2005). Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J Exp Med*, 201(6):891–902. (Cited on pages 15, 24, 32, 36, 37, and 48.)

Leslie, A. J., Pfafferott, K. J., Chetty, P., Draenert, R., Addo, M. M., Feeney, M., Tang, Y., Holmes, E. C., Allen, T., Prado, J. G., Altfeld, M., Brander, C., Dixon, C., Ramduth, D., Jeena, P., Thomas, S. A., John, A. S., Roach, T. A., Kupfer, B., Luzzi, G., Edwards, A., Taylor, G., Lyall, H., Tudor-Williams, G., Novelli, V., Martinez-Picado, J., Kiepiela, P., Walker, B. D., and Goulder, P. J. R. (2004). HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med*, 10(3):282–289. (Cited on pages 5, 10, and 16.)

Leulier, F., Parquet, C., Pili-Floury, S., Ryu, J.-H., Caroff, M., Lee, W.-J., Mengin-Lecreulx, D., and Lemaitre, B. (2003). The drosophila immune system detects bacteria through specific peptidoglycan recognition. *Nat Immunol*, 4(5):478–484. (Cited on page 43.)

Li, B., Gladden, A. D., Altfeld, M., Kaldor, J. M., Cooper, D. A., Kelleher, A. D., and Allen, T. M. (2007). Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J Virol*, 81(1):193–201. (Cited on page 16.)

Lucchiari-Hartz, M., Lindo, V., Hitziger, N., Gaedicke, S., Saveanu, L., van Endert, P. M., Greer, F., Eichmann, K., and Niedermann, G. (2003). Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic t lymphocyte epitope clustering in hiv-1-nef. *Proc Natl Acad Sci U S A*, 100(13):7755–7760. (Cited on pages 32, 36, 44, 46, and 48.)

Ludbrook, J. and Dudley, H. (20061998). Why permutation tests are superior to t and f tests in biomedical research. *The American Statistician*, 52(2):127–132. (Cited on page 35.)

Lundegaard, C., Nielsen, M., and Lund, O. (2006). The validity of predicted T-cell epitopes. *Trends Biotechnol*, 24(12):537–538. (Cited on page 11.)

Matano, T., Shibata, R., Siemon, C., Connors, M., Lane, H. C., and Martin, M. A. (1998). Administration of an anti-CD8 monoclonal antibody interferes

with the clearance of chimeric simian/human immunodeficiency virus during primary infections of rhesus macaques. *J Virol*, 72(1):164–169. (Cited on page 16.)

Mathunjwa, T. R. and Gary, F. A. (2006). Women and HIV/AIDS in the kingdom of swaziland: culture and risks. *J Natl Black Nurses Assoc*, 17(2):39–46. (Cited on page 60.)

Maurer, K., Harrer, E. G., Goldwich, A., Eismann, K., Bergmann, S., Schmitt-Haendle, M., Spriewald, B., Mueller, S. M., Harrer, T., and for HIV/AIDS, G. C. N. (2008). Role of cytotoxic T-lymphocyte-mediated immune selection in a dominant human leukocyte antigen-B8-restricted cytotoxic T-lymphocyte epitope in nef. *J Acquir Immune Defic Syndr*, 48(2):133–141. (Cited on page 24.)

Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., van Seventer, G., and Klein, J. (1988). Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. *EMBO J*, 7(9):2765–2774. (Cited on page 7.)

Mayr, E. (1961). Cause and effect in biology. *Science*, 134:1501–1506. (Cited on page 1.)

Mayr, E. (1997). *This is biology: The science of the living world.* Harvard University Press. (Cited on page 1.)

Meek, J. L. (1980). Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc Natl Acad Sci U S A*, 77(3):1632–1636. (Cited on pages 36 and 44.)

Milicic, A., Price, D. A., Zimbwa, P., Booth, B. L., Brown, H. L., Easterbrook, P. J., Olsen, K., Robinson, N., Gileadi, U., Sewell, A. K., Cerundolo, V., and Phillips, R. E. (2005). CD8+ T cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 Nef. *J Immunol*, 175(7):4618–4626. (Cited on page 10.)

Moore, C. B., John, M., James, I. R., Christiansen, F. T., Witt, C. S., and Mallal, S. A. (2002). Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439–1443. (Cited on pages 15, 32, 36, and 48.)

Müller, V., Ledergerber, B., Perrin, L., Klimkait, T., Furrer, H., Telenti, A., Bernasconi, E., Vernazza, P., Günthard, H. F., Bonhoeffer, S., and Study, S. H. C. (2006). Stable virulence levels in the HIV epidemic of Switzerland over two decades. *AIDS*, 20(6):889–894. (Cited on pages 15, 32, and 48.)

Murata, S., Takahama, Y., and Tanaka, K. (2008). Thymoproteasome: probable role in generating positively selecting peptides. *Curr Opin Immunol*, 20(2):192–196. (Cited on pages 6 and 82.)

Navis, M., Schellens, I., van Baarle, D., Borghans, J., van Swieten, P., Miedema, F., Kootstra, N., and Schuitemaker, H. (2007). Viral replication capacity as a correlate of HLA B57/B5801-associated nonprogressive HIV-1 infection. *J Immunol*, 179(5):3133–3143. (Cited on page 24.)

Neisig, A., Melief, C. J., and Neefjes, J. (1998). Reduced cell surface expression of HLA-C molecules correlates with restricted peptide binding and stable TAP interaction. *J Immunol*, 160(1):171–179. (Cited on page 5.)

Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, 20(9):1388–1397. (Cited on pages 8 and 10.)

Nonaka, M., Yamada-Namikawa, C., Flajnik, M. F., and Pasquier, L. D. (2000). Trans-species polymorphism of the major histocompatibility complex-encoded proteasome subunit lmp7 in an amphibian genus, xenopus. *Immunogenetics*, 51(3):186–192. (Cited on page 82.)

Nowak, M. and Schuster, P. (1989). Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet. *J Theor Biol*, 137(4):375–395. (Cited on page 79.)

Nowak, M. A., Tarczy-Hornoch, K., and Austyn, J. M. (1992). The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci U S A*, 89(22):10896–10899. (Cited on page 88.)

Pamer, E. G., Sijts, A. J., Villanueva, M. S., Busch, D. H., and Vijh, S. (1997). Mhc class i antigen processing of listeria monocytogenes proteins: implications for dominant and subdominant ctl responses. *Immunol Rev*, 158:129–136. (Cited on page 3.)

Pande, V. S., Grosberg, A. Y., and Tanaka, T. (1994). Nonrandomness in protein sequences: evidence for a physically driven stage of evolution? *Proc Natl Acad Sci U S A*, 91(26):12972–12975. (Cited on page 46.)

Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152(1):163–175. (Cited on page 10.)

Paulsson, K. M. (2004). Evolutionary and functional perspectives of the major histocompatibility complex class I antigen-processing machinery. *Cell Mol Life Sci*, 61(19-20):2446–2460. (Cited on page 9.)

Penn, D. J., Damjanovich, K., and Potts, W. K. (2002). Mhc heterozygosity confers a selective advantage against multiple-strain infections. *Proc Natl Acad Sci U S A*, 99(17):11260–11264. (Cited on page 5.)

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586. (Cited on page 1.)

Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., Wilson, S. S., Sidney, J., Lund, O., Buus, S., and Sette, A. (2006). A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol*, 2(6):e65. (Cited on pages 10, 11, 22, 33, and 39.)

Peters, B., Bulik, S., Tampe, R., Endert, P. M. V., and Holzhütter, H.-G. (2003). Identifying mhc class i epitopes by predicting the tap transport efficiency of epitope precursors. *J Immunol*, 171(4):1741–1749. (Cited on pages 44 and 74.)

Peters, B. and Sette, A. (2005). Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6:132. (Cited on pages 8 and 10.)

Piantadosi, A., Ngayo, M. O., Chohan, B., and Overbaugh, J. (2008). Examination of a second region of the hiv type 1 genome reveals additional cases of superinfection. *AIDS Res Hum Retroviruses*, 24(9):1221–1221. (Cited on page 89.)

Piatak, M., Saag, M. S., Yang, L. C., Clark, S. J., Kappes, J. C., Luk, K. C., Hahn, B. H., Shaw, G. M., and Lifson, J. D. (1993). High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science*, 259(5102):1749–1754. (Cited on page 57.)

Pielou, E. (1977). Mathematical ecology. *Wiley-Interscience*, (1). (Cited on pages 34, 35, and 40.)

Plata, F., Autran, B., Martins, L. P., Wain-Hobson, S., Raphal, M., Mayaud, C., Denis, M., Guillon, J. M., and Debr, P. (1987). Aids virus-specific cytotoxic t lymphocytes in lung disorders. *Nature*, 328(6128):348–351. (Cited on page 37.)

Poon, A. F. Y., Pond, S. L. K., Bennett, P., Richman, D. D., Brown, A. J. L., and Frost, S. D. W. (2007). Adaptation to human populations is revealed by within-host polymorphisms in HIV-1 and hepatitis C virus. *PLoS Pathog*, 3(3):e45. (Cited on pages 15, 23, 32, and 48.)

Procko, E. and Gaudet, R. (2009). Antigen processing and presentation: Tapping into abc transporters. *Curr Opin Immunol*, 21(1):84–91. (Cited on page 82.)

Prez, C. L., Larsen, M. V., Gustafsson, R., Norstrm, M. M., Atlas, A., Nixon, D. F., Nielsen, M., Lund, O., and Karlsson, A. C. (2008). Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *J Immunol*, 180(7):5092–5100. (Cited on page 12.)

Reits, E., Neijssen, J., Herberts, C., Benckhuijsen, W., Janssen, L., Drijfhout, J. W., and Neefjes, J. (2004). A major role for TPPII in trimming proteasomal degradation products for MHC class I antigen presentation. *Immunity*, 20(4):495–506. (Cited on page 4.)

Robinson, J., Waller, M. J., Fail, S. C., and Marsh, S. G. E. (2006). The IMGT/HLA and IPD databases. *Hum Mutat*, 27(12):1192–1199. (Cited on pages 52, 57, 64, and 67.)

Rock, K. L., York, I. A., Saric, T., and Goldberg, A. L. (2002). Protein degradation and the generation of MHC class I-presented peptides. *Adv Immunol*, 80:1–70. (Cited on pages 6, 10, and 82.)

Rodrigo, A. G., Shpaer, E. G., Delwart, E. L., Iversen, A. K., Gallo, M. V., Brojatsch, J., Hirsch, M. S., Walker, B. D., and Mullins, J. I. (1999). Coalescent estimates of HIV-1 generation time in vivo. *Proc Natl Acad Sci U S A*, 96(5):2187–2191. (Cited on page 1.)

Rolland, M., Nickle, D. C., and Mullins, J. I. (2007). HIV-1 group M conserved elements vaccine. *PLoS Pathog*, 3(11):e157. (Cited on page 17.)

Romero, V., Azocar, J., Ziga, J., Clavijo, O. P., Terreros, D., Gu, X., Husain, Z., Chung, R. T., Amos, C., and Yunis, E. J. (2008). Interaction of NK inhibitory receptor genes with HLA-C and MHC class II alleles in Hepatitis C virus infection outcome. *Mol Immunol*, 45(9):2429–2436. (Cited on page 5.)

Sacha, J. B., Chung, C., Rakasz, E. G., Spencer, S. P., Jonas, A. K., Bean, A. T., Lee, W., Burwitz, B. J., Stephany, J. J., Loffredo, J. T., Allison, D. B., Adnan, S., Hoji, A., Wilson, N. A., Friedrich, T. C., Lifson, J. D., Yang, O. O., and Watkins, D. I. (2007a). Gag-specific CD8+ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. *J Immunol*, 178(5):2746–2754. (Cited on pages 17 and 43.)

Sacha, J. B., Chung, C., Reed, J., Jonas, A. K., Bean, A. T., Spencer, S. P., Lee, W., Vojnov, L., Rudersdorf, R., Friedrich, T. C., Wilson, N. A., Lifson, J. D., and Watkins, D. I. (2007b). Pol-specific CD8+ T cells recognize simian immunodeficiency virus-infected cells prior to Nef-mediated major histocompatibility complex class I downregulation. *J Virol*, 81(21):11703–11712. (Cited on page 17.)

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 37(Database issue):D5–15. (Cited on pages 52 and 56.)

Schellens, I. (2009). *Adaptation of HIV-1 to the human immune system at the population level is driven by protective HLA-B alleles*. PhD thesis, Utrecht University. (Cited on pages 63 and 64.)

Schellens, I. M. M., Keşmir, C., Miedema, F., van Baarle, D., and Borghans, J. A. M. (2008). An unanticipated lack of consensus cytotoxic T lymphocyte epitopes in HIV-1 databases: the contribution of prediction programs. *AIDS*, 22(1):33–37. (Cited on pages 12, 33, and 39.)

Schmitz, J. E., Kuroda, M. J., Santra, S., Sasseville, V. G., Simon, M. A., Lifton, M. A., Racz, P., Tenner-Racz, K., Dalesandro, M., Scallon, B. J., Ghrayeb, J., Forman, M. A., Montefiori, D. C., Rieber, E. P., Letvin, N. L., and Reimann, K. A. (1999). Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science*, 283(5403):857–860. (Cited on page 16.)

Schneidewind, A., Brumme, Z. L., Brumme, C. J., Power, K. A., Reyor, L. L., O'Sullivan, K., Gladden, A., Hempel, U., Kuntzen, T., Wang, Y. E., Oniangue-Ndza, C., Jessen, H., Markowitz, M., Rosenberg, E. S., Skaly, R.-P., Kelleher, A. D., Walker, B. D., and Allen, T. M. (2009). Transmission and long-term stability of compensated CD8 escape mutations. *J Virol*, 83(8):3993–3997. (Cited on page 37.)

Shimizu, N., Takeuchi, Y., Naruse, T., Inagaki, M., Moriyama, E., Gojobori, T., and Hoshino, H. (1992). Six strains of human immunodeficiency virus type 1 isolated in Japan and their molecular phylogeny. *J Mol Evol*, 35(4):329–336. (Cited on page 64.)

Sidat, M. M., Mijch, A. M., Lewin, S. R., Hoy, J. F., Hocking, J., and Fairley, C. K. (2008). Incidence of putative hiv superinfection and sexual practices among hiv-infected men who have sex with men. *Sex Health*, 5(1):61–67. (Cited on page 89.)

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(1). (Cited on pages 73 and 75.)

Sironi, L., Lazzari, B., Ramelli, P., Stella, A., and Mariani, P. (2008). Avian tap genes: detection of nucleotide polymorphisms and comparative analysis across species. *Genet Mol Res*, 7(4):1267–1281. (Cited on pages 6 and 82.)

Slade, R. W. and McCallum, H. I. (1992). Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*, 132(3):861–864. (Cited on pages 6, 15, 52, 68, and 81.)

Snell, G. D. (1968). The H-2 locus of the mouse: observations and speculations concerning its comparative genetics and its polymorphism. *Folia Biol (Praha)*, 14(5):335–358. (Cited on page 10.)

Streeck, H., Li, B., Poon, A. F. Y., Schneidewind, A., Gladden, A. D., Power, K. A., Daskalakis, D., Bazner, S., Zuniga, R., Brander, C., Rosenberg, E. S., Frost, S. D. W., Altfeld, M., and Allen, T. M. (2008). Immune-driven recombination and loss of control after HIV superinfection. *J Exp Med*, 205(8):1789–1796. (Cited on page 89.)

Takeuchi, N. and Hogeweg, P. (2008). Evolution of complexity in RNA-like replicator systems. *Biol Direct*, 3:11. (Cited on page 1.)

Tanaka, K. (2009). The proteasome: overview of structure and functions. *Proc Jpn Acad Ser B Phys Biol Sci*, 85(1):12–36. (Cited on pages 6 and 82.)

Tanaka, K. and Kasahara, M. (1998). The mhc class i ligand-generating system: roles of immunoproteasomes and the interferon-gamma-inducible proteasome activator pa28. *Immunol Rev*, 163:161–176. (Cited on page 6.)

Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhütter, H.-G. (2005). 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. (Cited on pages 11, 32, 33, 39, 54, 57, 70, and 74.)

Toes, R. E., Nussbaum, A. K., Degermann, S., Schirle, M., Emmerich, N. P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T. P., Müller, J., Schönfisch, B., Schmid, C., Fehling, H. J., Stevanovic, S., Rammensee, H. G., and Schild, H. (2001). Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J Exp Med*, 194(1):1–12. (Cited on pages 6 and 7.)

Tossi, A., Sandri, L., and Giangaspero, A. (2002). New consensus hydrophobicity scale extended to non-proteinogenic amino acids. *Peptides*. (Cited on pages 36 and 44.)

Uebel, S. and Tamp, R. (1999). Specificity of the proteasome and the tap transporter. *Curr Opin Immunol*, 11(2):203–208. (Cited on page 44.)

van Baalen, C. A., Guillon, C., van Baalen, M., Verschuren, E. J., Boers, P. H. M., Osterhaus, A. D. M. E., and Gruters, R. A. (2002). Impact of antigen expression kinetics on the effectiveness of HIV-specific cytotoxic T lymphocytes. *Eur J Immunol*, 32(9):2644–2652. (Cited on page 17.)

van Ballegooijen, M., Bogaards, J. A., Weverling, G.-J., Boerlijst, M. C., and Goudsmit, J. (2003). Aids vaccines that allow hiv-1 to infect and escape immunologic control: a mathematic analysis of mass vaccination. *J Acquir Immune Defic Syndr*, 34(2):214–220. (Cited on page 60.)

van den Berg, H. A. and Rand, D. A. (2004). Foreignness as a matter of degree: the relative immunogenicity of peptide/mhc ligands. *J Theor Biol*, 231(4):535–548. (Cited on page 88.)

van der Kuyl, A. C. and Cornelissen, M. (2007). Identifying hiv-1 dual infections. *Retrovirology*, 4:67–67. (Cited on page 89.)

Vider-Shalit, T., Almani, M., Sarid, R., and Louzoun, Y. (2009). The HIV hide and seek game: an immunogenomic analysis of the HIV epitope repertoire. *AIDS*, 23(11):1311–1318. (Cited on page 64.)

Wagner, R., Leschonsky, B., Harrer, E., Paulus, C., Weber, C., Walker, B. D., Buchbinder, S., Wolf, H., Kalden, J. R., and Harrer, T. (1999). Molecular and functional analysis of a conserved CTL epitope in HIV-1 p24 recognized from a long-term nonprogressor: constraints on immune escape associated with targeting a sequence essential for viral replication. *J Immunol*, 162(6):3727–3734. (Cited on page 37.)

Waldhauer, I. and Steinle, A. (2008). NK cells and cancer immunosurveillance. *Oncogene*, 27(45):5932–5943. (Cited on page 2.)

Walker, B. A., van Hateren, A., Milne, S., Beck, S., and Kaufman, J. (2005). Chicken TAP genes differ from their human orthologues in locus organisation, size, sequence features and polymorphism. *Immunogenetics*, 57(3-4):232–247. (Cited on page 23.)

Walker, B. D. and Burton, D. R. (2008). Toward an AIDS vaccine. *Science*, 320(5877):760–764. (Cited on page 31.)

Walker, B. D., Chakrabarti, S., Moss, B., Paradis, T. J., Flynn, T., Durno, A. G., Blumberg, R. S., Kaplan, J. C., Hirsch, M. S., and Schooley, R. T. (1987). Hiv-specific cytotoxic t lymphocytes in seropositive individuals. *Nature*, 328(6128):345–348. (Cited on page 37.)

Walker, B. D. and Korber, B. T. (2001). Immune control of hiv: the obstacles of HLA and viral diversity. *Nat Immunol*, 2(6):473–475. (Cited on page 37.)

Watkins, D. I., McAdam, S. N., Liu, X., Strang, C. R., Milford, E. L., Levine, C. G., Garber, T. L., Dogon, A. L., Lord, C. I., and Ghim, S. H. (1992). New recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. *Nature*, 357(6376):329–333. (Cited on page 7.)

Wawer, M. J., Gray, R. H., Sewankambo, N. K., Serwadda, D., Li, X., Laeyendecker, O., Kiwanuka, N., Kigozi, G., Kiddugavu, M., Lutalo, T., Nalugoda, F., Wabwire-Mangen, F., Meehan, M. P., and Quinn, T. C. (2005). Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis*, 191(9):1403–1409. (Cited on pages 54, 55, and 56.)

White, S. H. and Jacobs, R. E. (1990). Statistical distribution of hydrophobic residues along the length of protein chains. implications for protein folding and evolution. *Biophys J*, 57(4):911–921. (Cited on page 46.)

Whitney, J. B., Cobb, R. R., Popp, R. A., and O'Rourke, T. W. (1985). Detection of neutral amino acid substitutions in proteins. *Proc Natl Acad Sci U S A*, 82(22):7646–7650. (Cited on page 17.)

Wilk, M. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of databases. *Biometrika*, 55(1):1–17. (Cited on pages 34 and 40.)

Willberg, C. B., McConnell, J. J., Eriksson, E. M., Bragg, L. A., York, V. A., Liegler, T. J., Hecht, F. M., Grant, R. M., and Nixon, D. F. (2008). Immunity to hiv-1 is influenced by continued natural exposure to exogenous virus. *PLoS Pathog*, 4(10). (Cited on page 89.)

Yewdell, J. W. and Bennink, J. R. (1999). Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annu Rev Immunol*, 17:51–88. (Cited on pages 3, 17, and 20.)

Yewdell, J. W., Reits, E., and Neefjes, J. (2003). Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol*, 3(12):952–961. (Cited on page 88.)

Yokomaku, Y., Miura, H., Tomiyama, H., Kawana-Tachikawa, A., Takiguchi, M., Kojima, A., Nagai, Y., Iwamoto, A., Matsuda, Z., and Ariyoshi, K. (2004). Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. *J Virol*, 78(3):1324–1332. (Cited on pages 7, 17, 18, 52, 61, 63, 68, and 81.)

Yusim, K., Kesmir, C., Gaschen, B., Addo, M. M., Altfeld, M., Brunak, S., Chigaev, A., Detours, V., and Korber, B. T. (2002). Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *J Virol*, 76(17):8757–8768. (Cited on pages 7, 8, 15, 31, 32, 37, 44, 48, 68, and 85.)

Zafiropoulos, A., Barnes, E., Piggott, C., and Klenerman, P. (2004). Analysis of 'driver' and 'passenger' CD8+ T-cell responses against variable viruses. *Proc Biol Sci*, 271 Suppl 3:S53–S56. (Cited on page 17.)

Zhang, D., Shankar, P., Xu, Z., Harnisch, B., Chen, G., Lange, C., Lee, S. J., Valdez, H., Lederman, M. M., and Lieberman, J. (2003). Most antiviral CD8 T cells during chronic viral infection do not express high levels of perforin and are not directly cytotoxic. *Blood*, 101(1):226–235. (Cited on page 16.)

Samenvatting

Het humane immuunsysteem gebruikt verschillende strategien om virale infecties te bestrijden. Een daarvan is het cellulaire immuunsysteem, wat bestaat uit cytotoxische T cellen (CTLs). Deze CTLs kunnen virus-genfecteerde cellen herkennen aan de hand van de virale eiwitfragmenten die op het celoppervlak van een genfecteerde cel worden gepresenteerd. Virussen staan onder sterke selectiedruk om te voorkomen dat de cel waarin ze zich voortplanten voortijdig vernietigd wordt door CTLs, en evolueren mutaties die ervoor zorgen dat er minder eiwitfragmenten van hun protenen op het celoppervlak gepresenteerd worden. Zo ontsnappen ze aan de immuunreactie van de gastheer.

In deze thesis bestuderen we hoe de co-evolutie tussen virussen en gastheren de klassieke antigeen presentatie route (een groep van moleculen die ervoor zorgen dat intracellulaire protenen op het celoppervlak ten toon gesteld worden), gevormd heeft. We kiezen het AIDS virus als model om te onderzoeken hoe een virus zich aanpast aan een nieuwe gastheer, en leren zodoende ook meer over hoe het menselijke immuunsysteem om kan gaan met de snelle evolutie van virussen. Allereerst bestuderen we hoe HIV-1 zich op een populatieniveau heeft aangepast gedurende de laatste 25 jaar (Chapter 2), door per HIV-1 sequentie het totaal aantal voorspelde CTL epitopen (de virale eiwitfragmenten, die door de route gepresenteerd worden) en hun intracellulaire voorlopers te bepalen, en de trends in hun aantallen over tijd te bestuderen. Ook hebben we gekeken naar de distributie van deze epitopen over de protenen van HIV-1, en die distributie vergeleken met die van de protenen van andere organismen (Chapter 3). In beide gevallen vonden we geen bewijs dat het virus zich in grote mate aan had gepast aan de humane populatie. Dit was een onverwachte bevinding, aangezien twee belangrijke moleculen (genaamd 'proteasoom' en 'TAP') in de antigeen presentatie route, die tezamen de intracellulaire voorlopers van epitopen genereren, bijna geen variatie vertonen tussen verschillende mensen. De verwachting was dat een snel evoluerend virus als HIV-1 veel ontsnappingmutaties voor deze twee monomorfe molecule zou hebben verzameld in zijn genoom.

Aan de hand van deze bevindingen, stelden we de hypothese op dat het HIV-1 virus zich niet kan adapteren aan het monomorfe proteasoom en TAP, omdat het meest specifieke molecuul in de antigeen presentatie route, het MHC, juist erg polymorf is. Er is dus een grote genetische diversiteit aan MHC moleculen in de humane populatie, en twee verschillende MHC moleculen zullen niet dezelfde set van intracellulaire voorlopers kunnen herkennen. Wanneer het virus wordt overgedragen naar een andere gastheer, verandert daarmee dus ook welke delen van het virale genoom onder selectiedruk staan van cytotoxische T cellen. Ontsnappingsmutaties in HIV-1 die het monomorfe proteasoom en TAP, alsook die het polymorfe MHC benvloedden, staan geregeld niet onder

selectiedruk van een immuunsysteem, en kunnen daarom weer terug muteren naar een oorspronkelijke staat.

Dit scenario hebben we getoetst in een computermodel waarin een virtueel HIV-1 virus zich kon adapteren aan virtuele gastheer populatie (Chapter 4). We vonden dat in een situatie waarin de gastheer populatie een hoge mate van MHC polymorfisme had, virussen geselecteerd werden om zich voornamelijk aan te passen aan de monomorfe proteasome en TAP moleculen, maar dat tegelijk door de grote variatie in MHC moleculen tussen de verschillende gastheren, dit niet leidde tot een hoge mate van aanpassing van het virus aan de gastheer populatie. Als laatste hebben we het computermodel zodanig uitgebreid dat de antigeen presentatie route van de gastheer populatie ook kon evolueren (Chapter 5). Met deze uitbreiding konden we bestuderen of de huidige structuur van de presentatie route kon worden uitgelegd puur in termen van gastheer-virus co-evolutie. Gebaseerd op onze eerder geponeerde hypothese, verwachtte we dat alleen in de meest specifieke stap van de route een polymorfisme zou evolueren. Onder de aanname dat de verschillende stappen van de route onderling nog redelijk op elkaar bleven aansluiten, evolueerde de virtuele gastheer populatie in het model inderdaad een antigeen presentatie route met een structuur gelijk aan die van de mens.

Curriculum Vitæ

The author of this thesis, Boris Valentijn Schmid, was born on July 16th, 1978 in Leiderdorp, Netherlands. From 1989 onwards he attended the Stedelijk Gymnasium te Leiden, and graduated in 1996. Boris moved to Utrecht to study biology. There was little doubt there: even though the field of cognitive artificial intelligence seemed tempting, and fitted his recent hobby of programming various individual based models, the nature documentaries of Sir David Attenborough had fascinated him since an early age: biology it was to be. During his studies, Boris predominantly followed courses in ecology, evolution, and (neuro-) ethology. As his last required course in biology, Boris followed the course 'Bioinformatic processes' of Prof. Paulien Hogeweg. The course came as a revelation, as it combined the biological topics that were dear to him with methods that allow for a far quicker cycle of hypothesis forming, testing, and understanding of the biology beneath it, than typical in field work. It also allowed him to incorporate his hobby of programming into his study of biology.

After finishing his masters projects ('Testing a model for the distribution of grooming in a group of rhesus macaques', with Dr. Annet Louwerse and 'Developing an agent-based model to study a stable behavioral polymorphism in fruit flies', with Prof. Paulien Hogeweg), Boris started a PhD study in 2004, supervised by Prof. Rob J. de Boer, which resulted in this thesis. Currently, Boris has moved more towards epidemiology, and is modelling disease spread through sexual contact networks at the Rijksinstituut voor Volksgezondheid en Milieu (Bilthoven, NL).

List of Publications

Boris V. Schmid, Can Keşmir, 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, 2008* 3(10): e3525. doi:10.1371/journal.pone.0003525

Boris V. Schmid, Can Keşmir, and Rob J. de Boer The distribution of CTL epitopes in HIV-1 appears to be random, and similar to that of other proteomes. *accepted by BMC evolutionary Biology.*

Boris V. Schmid, Rob J. de Boer Quantifying how MHC polymorphism prevents pathogens from adapting to the antigen presentation pathway. *submitted*

Boris V. Schmid, Rob J. de Boer The emergen of polymorphism in the antigen presentation pathway. *in preparation*

Acknowledgments

First and foremost, I want to thank my promotor Rob de Boer and co-promotor Can Keşmir. Together they form an amazing team, and have succesfully supervised and steered me through my PhD. The scientific freedom you granted me, and the patience you showed when we had conflicting opinions was remarkable. Both of you willingly took the time to debate with me for hours, just so that all of us would be of the same opinion again. Both of you were also willing to wrestle yourself through 15 versions of my first paper, before we called in professional help, all the while trying to supporting me in becoming a better scientist. Perhaps due to the large overlap in the way you supported me, the differences between you stand out. Rob, I am still amazed by your ability to quickly pick up and understand the ideas that I explain imprecisely and incoherently, as they are still too fresh in my mind to have fully crystallized into language. You bring valuable clarity to any discussion. Can, I am utterly grateful for knowing that during our discussions you would also keep an eye out for my emotional state, and support me when I needed it.

That brings me to the rest of you ;-). Among all the people I would like to thank, I will name two more by name. These are my mentor Hans de Cock, en my tutor in scientific writing, Linda McPhee: thanks for helping me with probably the most frustrating part of my PhD thesis. Somehow, I refused to learn by example how to write scientific papers, I needed someone to explain to me how people read papers, and how papers are written. Hans' supporting words and advice, and especially Linda's writing course helped me enormously.

Perhaps now then is a good moment to thank all my colleagues. Somewhere during the last 5 years, interacting with all of you changed my opinion of what was important in work. Previously I wanted a challenging and interesting topic, coupled with new skills and tricks to learn and add to my toolbox. These days, I find so much joy in interacting with the people I work with, that future workplaces are more and more considered in terms of who are my colleagues. I very much enjoyed my time with all of you.