Computational Immunology: Analyses of Viral Escape, Epitope Binding and T Cell Receptor Recognition

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultät III

Chemie, Pharmazie, Bio- und Werkstoffwissenschaften

der Universität des Saarlandes

von

Kirsten Rump

Saarbrücken

2011

Tag des Kolloquiums:	12. Juli 2011
Dekan:	Prof. Dr. Wilhelm F. Maier
Vorsitzender:	Prof. Jörn Walter, Ph.D.
Berichterstatter:	Prof. Dr. Thomas Lengauer, Ph.D. Prof. Dr. Volkhard Helms
Akademischer Mitarbieter:	PD Dr. Michael Hutter

In Memory of My Father

Martin Rump
1937 – 1997

Abstract

It has been shown repeatedly that infectious diseases in humans have strong associations with the human leukocyte antigen system, but an understanding of the basis of these associations remains elusive. Adaptive immune responses involving CD4 and CD8 T lymphocytes are dependent on (1) the appropriate and effective processing of a peptide from a protein source, (2) the stable binding of the peptide to the HLA molecule and (3) the recognition of this complex by the T cell receptor. In this thesis, we present work helping to better define such host-virus dynamics, examining aspects relating to each of the described steps. We examined two large patient cohorts, the first infected with HIV-1 and the second with HCV. We identified viral escape mutations and thus potential immune epitopes. Also, we examined the possible effects of HLA genotypes on the development of drug resistance mutations (HIV-1) and the success of antiviral therapy (HCV). To better understand the stable binding of peptides to HLA molecules, we evaluated the performance of diverse HLA class I prediction methods on large datasets, showing that all leading methods are capable of good to excellent performance. Finally, we developed the first algorithms, based on the interactions found in actual experimental structures, which allow for the prediction of interactions between residues in the T cell receptor's CDR loops and residues in the HLA-peptide antigen. The algorithms had good performance under cross-validation.

Kurzfassung

Wiederholt wurden viele Zusammenhänge menschlicher Infektionskrankheiten mit dem Human-Leukozyten-Antigen-System aufgezeigt, doch ein vollständiges Verständnis dieser Zusammenhänge fehlt. Adaptive Immunantworten mit CD4- und CD8-T-Lymphozyten sind abhängig von (1) einer angemessenen und effektiven Bearbeitung eines Peptids, (2) der stabilen Bindung des Peptids an das HLA-Molekül und (3) der Erkennung dieses HLA-Peptid-Komplexes durch den T-Zell-Rezeptor. In dieser Dissertation präsentieren wir Arbeiten, die helfen, diese Wirt-Virus-Dynamik besser zu definieren, indem wir Aspekte jedes dieser beschriebenen Schritte untersuchen. In zwei großen Patientengruppen (die erste mit HIV-1 und die zweite mit HCV infiziert) identifizierten wir virale Escape-Mutationen und damit potentielle Immun-Epitope. Wir untersuchten die möglichen Auswirkungen des HLA-Genotypes auf die Entwicklung von Resistenz-Mutationen (HIV-1) und den Erfolg einer antiviralen Therapie (HCV). Um die stabile Bindung von Peptiden an HLA-Moleküle besser zu verstehen, untersuchten wir die Leistung verschiedener HLA-Klasse I-Prognoseverfahren und zeigten, dass alle führenden Methoden gute bis sehr gute Ergebnisse liefern können. Abschließend haben wir die ersten Algorithmen entwickelt, die die Interaktionen zwischen den Aminosäuren der CDR-Schleifen des T-Zell-Rezeptors und Aminosäuren des HLA-Peptid-Komplexes vorhersagen. Diese Algorithmen zeigten gute Leistung unter Cross-Validierung.

Extended Abstract

It has been shown repeatedly that infectious diseases in humans have strong associations with the human leukocyte antigen system, but an understanding of the basis of these associations remains elusive. Adaptive immune responses involving CD4 and CD8 T lymphocytes are dependent on (1) the appropriate and effective processing of a peptide from a protein source, (2) the stable binding of the peptide to the HLA molecule and (3) the recognition of this complex by the T cell receptor. In this thesis, we present work helping to better define such host-virus dynamics, examining aspects relating to each of the described steps.

Viruses with high mutation rates are capable of adapting to their host's HLA profile if the fitness costs for the virus are not too high. We examined two large patient cohorts, the first infected with HIV-1 and the second with HCV, in order to identify viral escape mutations and thus potential immune epitopes which help define the host's adaptive immune response. In the case of HIV-1, we examined the persistence of HLA-associated mutations over time and the possible effects of HLA genotypes on the development of drug resistance mutations. In the case of HCV, we examined the effect of the host genetic HLA profile on the success of antiviral therapy. We also developed a DAS server which serves as a reference and annotation server for HIV-1 and HCV, targeted at the genome annotation community.

To better understand the stable binding of peptides to HLA molecules, we evaluated the performance of diverse HLA class I prediction methods on large datasets which recently became available. Overtraining was avoided by making use of methods which had been, wherever possible, reimplemented. This methodology permits the control of both learning and testing of the prediction methods on a variety of datasets, thus allowing for a more accurate comparison of the methods. While all leading methods are capable of achieving good to excellent performance for clear binders and non-binders in cases where sufficient data is available, intermediate binders remain difficult to categorize.

Finally, we examined interactions, which are still not well understood, between T cell receptors and the HLA-peptide complex. We identified important differences to previous studies and, based on the interactions found in actual experimental structures, developed the first algorithms which allow for the prediction of interactions between residues in the T cell receptor's CDR1, CDR2 and CDR3 loops and residues in the HLA-peptide antigen. The algorithms had good performance under cross-validation.

Acknowledgements

This work was carried out in the Department for Computational Biology and Applied Algorithmics at the Max Planck Institute for Informatics in Saarbrücken, Germany. I would like to thank my supervisor Thomas Lengauer for his encouragement, advice, and support, and for providing me with the freedom to pursue my own ideas in a stimulating atmosphere. I am also very grateful to Volkhard Helms for his ongoing support throughout this time.

Thank you to those that also guided me during my work, in particular Niko Beerenwinkel, Jörg Rahnenführer, Iris Antes, and Francisco Domingues. Also to my colleagues in the Lengauer Group for many inspiring discussions on my research and for creating such a pleasant work environment, especially Ingolf Sommer, Jochen Maydt, André Altmann, Tobias Sing, Oliver Sander, Jasmina Bogojeska, Laura Tolosi, Adrian Alexa, Christoph Welsch, Priti Talwar and Dorothea Emig.

I am very grateful to the clinicians and virologists who collected and provided data for this work and who also gave invaluable input on the study design, in particular Martin Däumer, Rolf Kaiser, Golo Ahlenstiel, Ulrich Spengler, Christian Lange, and Christoph Sarrazin.

I would also like to thank Achim Büch and Georg Friedrich for crucial technical support. Uwe Brahm, Bernd Färber, Jörg Herrmann, Wolfram Wagner, and the rest of the Information Services & Technology team provided much friendly and prompt help over the years. And thank you to Ruth Schneppen-Christmann for her dedication in helping with administrative problems and keeping the group running smoothly.

Finally, I would like to thank family and friends for their support and understanding, especially Kevin, my mother and Kristof. I have dedicated this thesis to my father who passed away in 1997, who always encouraged my interest in the life sciences.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

1.1 HIV-1

1.1.1 The pandemic

Human immunodeficiency virus 1 (HIV-1) is the pathogen which currently causes the most extreme case of immunosupression in humans. Over time, infection with HIV-1 leads to the gradual loss of immune competence, in the form of CD4 T cell depletion, allowing for the development of cancer and the infection of the host with organisms that are not normally pathogenic (Douek 2003; Douek et al. 2003).

The earliest scientific reports of cases of previously healthy young homosexual men dying of Pneumocystis carinii pneumonia (PCP) and Kaposi's sarcoma in New York and California were made in mid-1981 (CDC 1981a; CDC 1981b; Hymes et al. 1981). Acquired immunodeficiency syndrome (AIDS) was first properly defined by the Centers for Disease Control (CDC) in Atlanta in September 1982 (CDC 1982c). Later that year it also became clear that this new disease was probably caused by an infectious agent: a 20-month old child receiving multiple transfusions of blood and blood products died from infections related to AIDS (CDC 1982a) and the first case of mother to child transmission was described (CDC 1982b). In 1983, the virus we now call HIV-1 was first isolated. (Controversy surrounds who first isolated HIV-1, but many view the awarding of the Nobel Prize in Medicine or Physiology to Luc Montagnier and Françoise Barré-Sinoussi from the Pasteur Institute in Paris, France as a final scientific verdict on the matter (Lever and Berkhout 2008; Vahlne 2009)).

By the end of 2009, the number of people who were living with HIV-1/AIDS had grown to 33.3 million and more than 25 million people had died of the disease since 1981. It has been estimated that 2.6 million new infections and 1.8 million deaths occurred in 2009 (Figure 1.1). Women accounted for 50% of all adults living with HIV-1 world-wide and Africa had 16.6 million AIDS orphans. Those given access to antiviral drugs have rose to 5.2 million, but 10 million were still waiting for treatment in developing and transitional countries. At present, the virus is spreading most rapidly in Central Asia and Eastern Europe (UNAIDS 2009; UNAIDS 2010).

Assuming that 80% of infected individuals receive antiviral therapy by 2012, it has been projected that the number of AIDS deaths will continue to rise and reach 6.5 million per year by 2030. This rise is due to projected demographic changes in sub-Saharan Africa, where population growth is highest and HIV/AIDS incidence rates are assumed to remain largely constant. Under this model, the total number of deaths due to AIDS is projected to be 117 million between 2006 and 2030 (Mathers and Loncar 2006; UNAIDS 2010). The importance of tackling the spread of HIV-1/AIDS was already recognized by the United Nations in 2000 when it defined its nine most important development targets for the 21st century: one of these targets was to reduce the HIV-1 infection rates for persons 15–25 years of age by 25% within 10 years (Annan 2000).

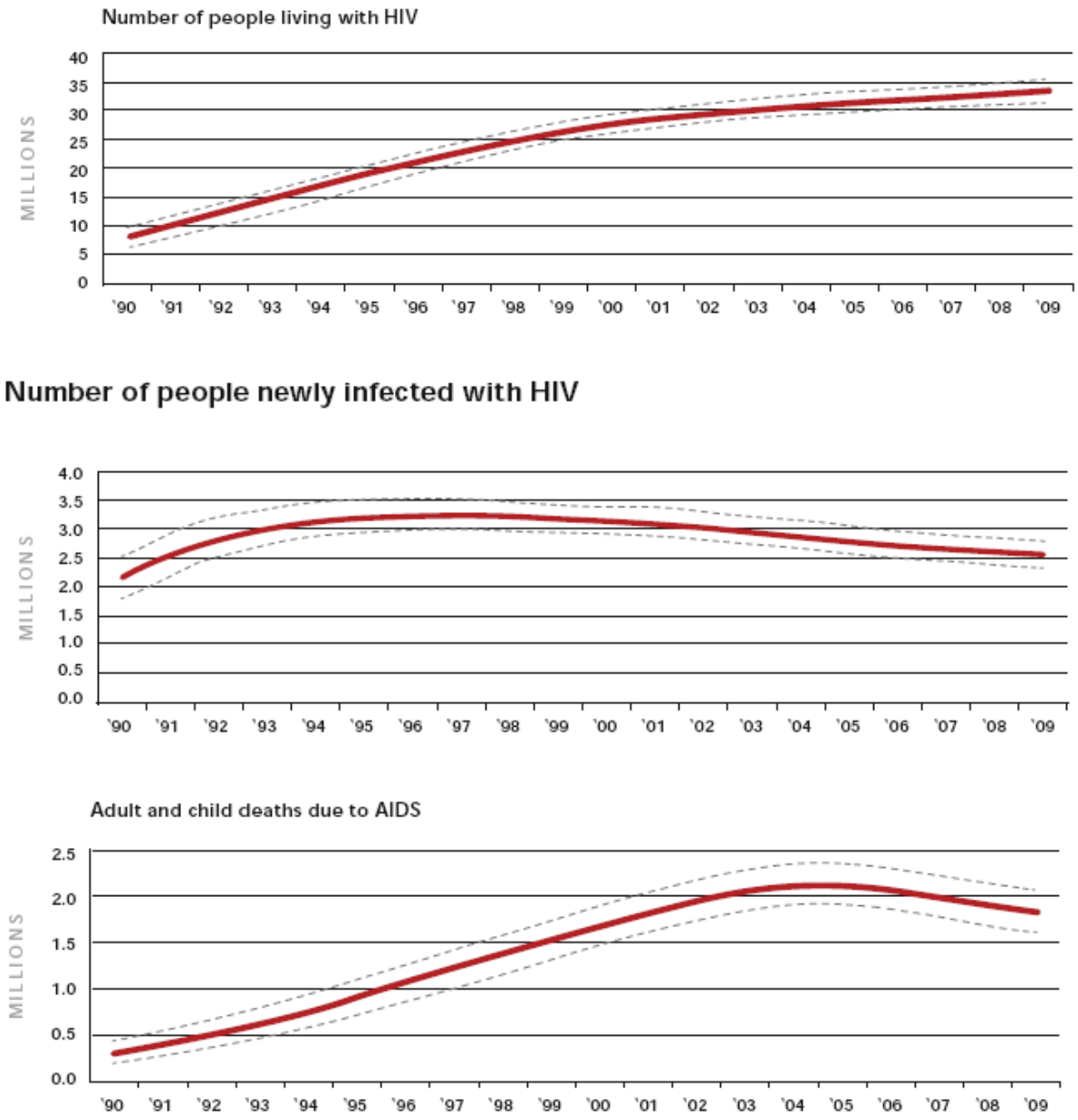

Figure 1.1 Numbers of people living with, newly infected with, and killed by HIV-1 from 1990-2009 (UNAIDS 2010)

1.1.2 HIV-1 groups, subtypes and nomenclature

HIV-1 is a retrovirus and a member of the primate lentivirus family. It is derived from retroviruses indigenous to African primates. Therefore, the HIV infection can be considered to have originally been a zoonotic disease introduced to humans by primates. HIV-1 shows great genetic variability and has been classified into four major groups, based on nucleotide sequence: M (main), O (outlier), N (non-M and non-O), and the newly identified P group. These groups are only distantly related to each other and therefore thought to be the result of four separate zoonotic transmission events from chimpanzees (M, N, O) and gorilla (P). The O group currently appears to be restricted to west-central Africa and the N group is extremely

rare and was discovered in Cameroon in 1998. The P group was identified recently in a woman in Cameroon (DeFranco et al. 2007; Murphy et al. 2008; Plantier et al. 2009).

In contrast to the O, N and P groups which have essentially remained restricted to Africa, the M group is responsible for the global HIV-1 pandemic. The M group contains several lineages (also called subtypes or clades) which are designated with the letters A, B, C, D, F, G, H, J and K. Lineages have been known to recombine into circulating recombinant forms (CRFs) and, for example, the CRF A/B is a mixture of subtypes A and B. The lineages and CRFs are distributed unevenly across the globe, with the most widespread being A and C. Historically, subtype B has been the common subtype in Europe, the Americas, Japan and Australia (Robertson et al. 2000; Perrin et al. 2003).

For the sake of completion, it should be noted that the second major type of HIV is designated HIV-2 and shares 42% sequence identity with HIV-1. HIV-2 is thought to have originated through a zoonotic transmission from sooty mangabeys, with eight independent transmission events occurring from this monkey to humans. Recombination between the two virus types has not been reported and probably cannot occur, because of differences in the RNA dimer hairpin sites. HIV-2 is also distributed worldwide, but has a lower transmission rate than HIV-1 and a lower rate of progeny virus production. The course of infection is also less pathogenic, resembling in part the infection of sooty mangabeys and African Green Monkeys (AGMs) with simian immunodeficiency virus (SIV) (Levy 2009; Wertheim and Worobey 2009).

1.1.3 Structure and genome

Mature virions display a broad range of diameters, but the average size is approximately 145 nm. A virion is generally spherical in shape and consists of a lipid bilayer which surrounds a cone-shaped core (also called nuleocapsid) (Figure 1.2). A core contains two copies of a single-stranded viral RNA genome, as well as virus-specific enzymes. Most virions contain a single core, but approximately one-third contain two or more such cores. These multi-core virions are larger in diameter. There is also a small subset of virions (about 7%) which are not conical in shape but tubular, although it is thought that all virions share a similar internal organization (Briggs et al. 2003).

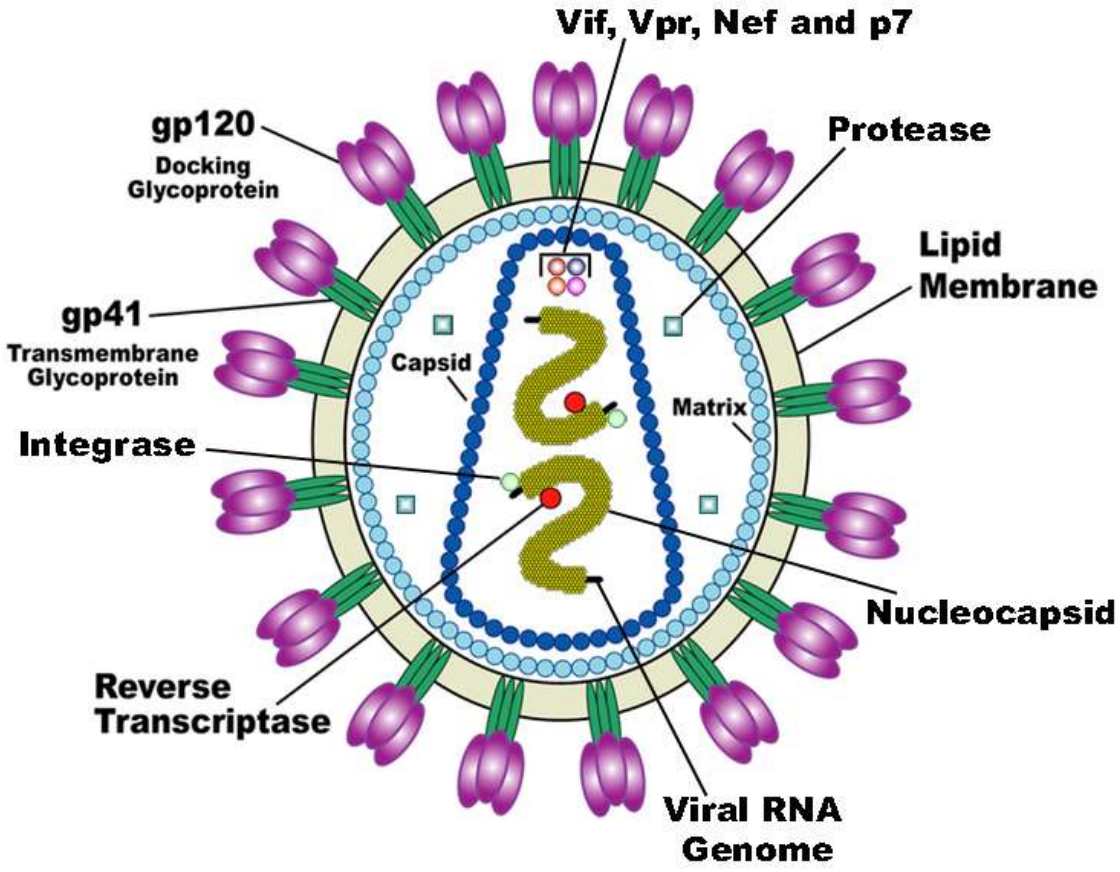

Figure 1.2 Schematic diagram of a mature HIV-1 virion (www.wikipedia.org)

The genome of the virus is 9.2 kb in length (Ratner et al. 1985) (Figure 1.3). The three structural genes are *gag* (group-specific antigen), *pol* (polymerase) and *env* (envelope). The *gag* gene produces a precursor protein, which is processed into the matrix protein (sometimes also described as p17 or MA), the capsid protein (p24 or CA) and the nucleocapsid protein (p7 or NC). These proteins package genomic viral RNA into new virions and participate in viral uncoating. The *pol* gene encodes the protease (PR), reverse transcriptase (RT), RNAse and integrase (IN). These proteins are required for viral replication. The *env* gene encodes membrane-exposed proteins on the virion that mediate attachment and fusion; the precursor gp160 protein is cleaved into these membrane-exposed proteins which are named gp41 and gp120. Furthermore, *tat* and *rev* are two regulatory genes which are essential for *in vivo* and *in vitro* replication. And finally, the accessory genes *vpu*, *vpr*, *nef* and *vif* are important for viral pathogenicity *in vivo*. The vpu protein is required for viral assembly and budding, vpr protein is important for nuclear import, nef protein disturbs the morphology of the endosomal recycling compartment by several mechanisms and downregulates CD4 and MHC class I expression, and vif protein is a viral infectivity factor which acts by circumventing the potent antiviral activity of the host's APOBEC3G cell protein. Like all retroviruses, HIV-1's genome is flanked by long terminal repeats (LTRs) which are involved in viral integration and in the regulation of transcription (Madrid et al. 2005; DeFranco et al. 2007; Paul 2008).

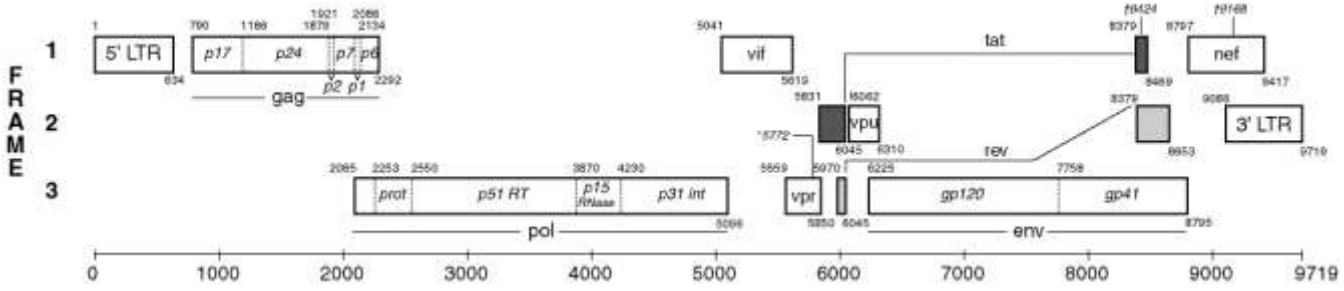

Figure 1.3 The HIV-1 gene map of the reference strain HXB2 (www.hiv.lanl.gov)

1.1.4 Viral replication

HIV-1 is capable of infecting all cells which express CD4 on the cell surface. These cells include CD4 T cells, dendritic cells, macrophages and monocytes, but the production of infectious virus particles primarily occurs in activated T cells. Initially, the trimeric gp120 subunit located on the surface of the HIV-1 virion binds to the CD4 membrane protein which is located on the surface of the host CD4 T cells, dendritic cells or macrophages. gp120 subsequently undergoes a significant conformational change, which increases the interactions with cellular coreceptors and exposes the trimeric transmembrane fusion protein gp41. These cellular coreceptors can consist of different chemokine receptors, including CXCR4 and CCR5. This process is followed by membrane fusion (DeFranco et al. 2007).

Upon entry into the host cell, the viral genome is converted to DNA by the viral reverse transcriptase: the HIV-1's inner core acts as a template for the reverse transcription of RNA into double-stranded DNA. The viral cDNA is transported into the host cell's nucleus by making use of nuclear localization signals. The proviral genome is integrated into the host cell's genome via the viral integrase, generating a provirus which becomes the primary template for subsequent transcription of a new generation of virions. Expression is greatly enhanced by the activation of host transcriptional regulators and the viral tat protein. Subsequently, mature viral RNAs are transported into the cytoplasm where the viral core proteins and the viral reverse transcriptase are synthesized. Gp120 and gp41 are synthesized in the endoplasmic reticulum. Membrane-associated envelope proteins recruit structural proteins and two copies of the full-length viral RNA genome. This is followed by membrane budding on the cell surface of T cells. In macrophages, mature virions assemble on endosomal membranes and bud into multivesicular bodies (MBVs) which are vesicles of the endosomal pathway. MBVs fuse with the plasma membrane and subsequently release mature virions from the macrophage. The cellular endosomal sorting complex required for transport (ESCRT) mediates the budding process in both cell types (DeFranco et al. 2007; Paul 2008). A general overview of this process can be seen in Figure 1.4.

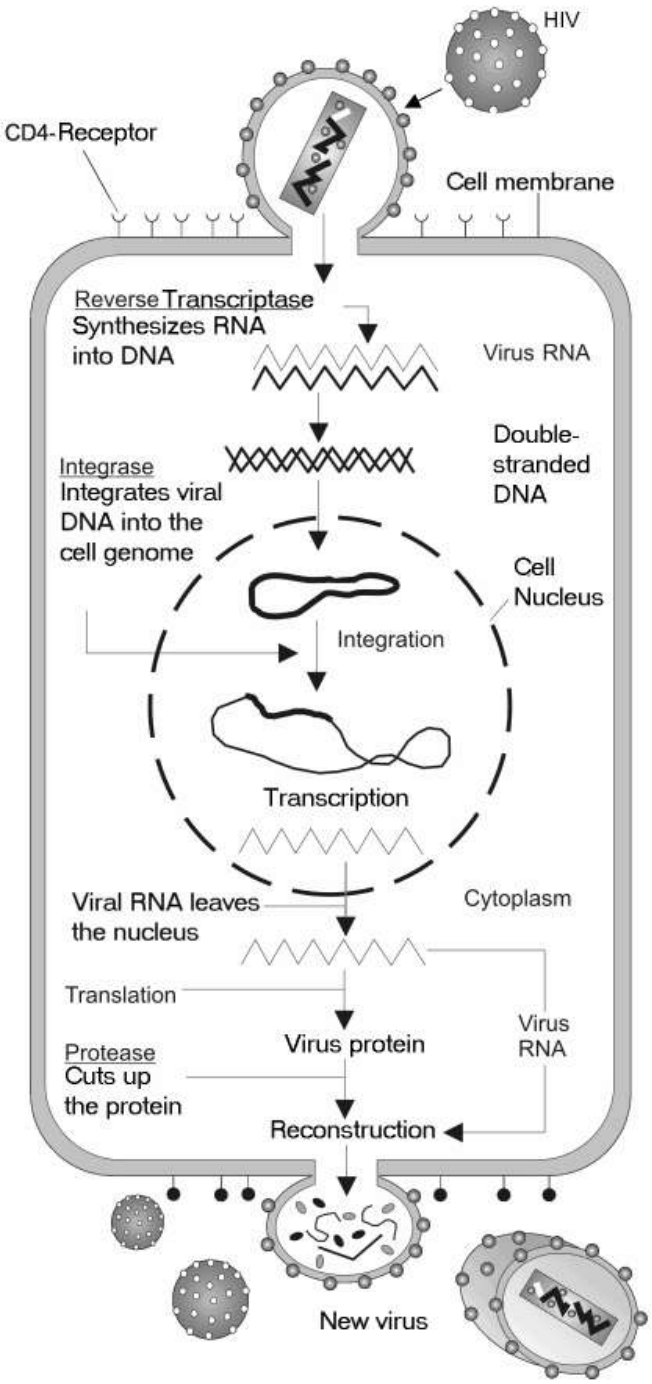

Figure 1.4 The replication cycle of HIV-1 (www.wikipedia.org)

1.1.5 Transmission

HIV-1 is primarily transmitted by sexual contact. A further route of transmission is provided by the sharing of needles among intravenous drug users, particularly in Eastern Europe. Mother to infant transmission remains a problem in many regions world-wide, but intrapartum (during labor and birth) and peripartum (during the first months after birth) transmission can be substantially reduced by treating infected women with a short course of antiretroviral drugs. However, subsequent breast feeding of the infant is still a major source of infection in the developing world. Transmission via the transfusion of infected blood

products has been largely eliminated through the screening of blood products, procedures which were developed in the mid-1980s (Cohen et al. 2008).

1.1.6 Course of an untreated infection

Typically the course of an untreated infection with HIV-1 can be divided into four phases: (1) primary infection and the acute period, (2) the asymptomatic or clinical latency period, (3) the symptomatic period during which constitutional symptoms occur and (4) AIDS during which opportunistic diseases occur which lead to death (Figure 1.5).

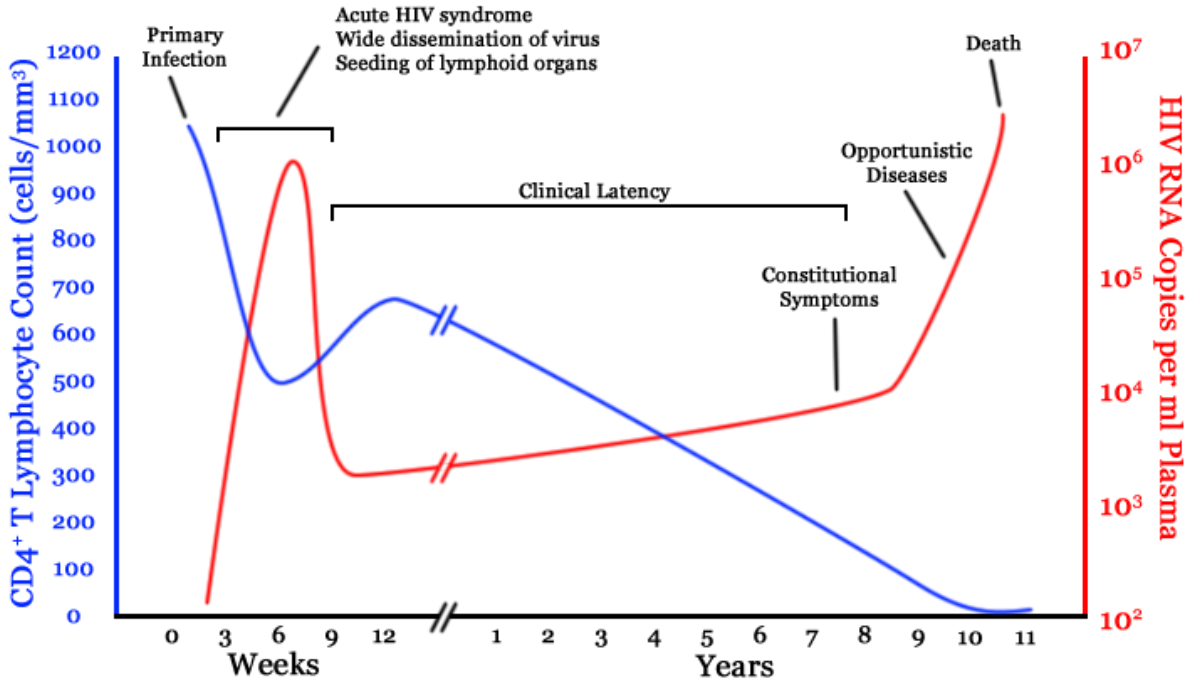

Figure 1.5 The typical course of an untreated HIV-1 infection (www.wikipedia.org)

The CD4 T cell count (cells per µl) is shown in blue; the HIV-1 RNA copies per ml of plasma is shown in red.

Subsequent to infection 50-80% of individuals have flu-like symptoms which clinically characterize the acute phase. There is a sharp increase in viremia (the abundance of the virus) and a marked decrease in the number of circulating CD4 T helper cells. In virtually all patients, CD8 cytotoxic T cells are activated which kill HIV-1 infected cells, leading to a rebound in the number of CD4 T cells. This period of acute infection typically lasts from 2 to 6 weeks. Diagnosis of HIV-1 infection at this stage is usually missed, unless there is a specific cause for suspicion (Murphy et al. 2008; Paul 2008).

Antibodies begin to be produced (seroconversion) and this characterizes the transition from the acute to the asymptomatic phase. A balance occurs between the level of viral destruction of immune cells and the host response, which is designated the viral set-point. This viral set-point is generally considered the best indicator of disease progression. Numerous viral mutants appear due to the high level of viral replication, the error-prone copying action of the viral reverse transcriptase and the absence of a viral nucleotide repair system. Specific mutations, called escape mutations, are selected for which help the virus evade immune recognition. This asymptomatic period can last anywhere from 6 months to 20

years. The number of CD4 T cells declines progressively in the asymptomatic phase, until a point is reached where constitutional symptoms begin to appear and the patient moves into the symptomatic phase. Constitutional symptoms include low grade fevers, chronic fatigue, general weakness (Murphy et al. 2008; Paul 2008).

A diagnosis of AIDS can be made if the patient has fewer than 200 CD4 cells / µl, the patient's CD4 cells account for less than 14% of all lymphocytes or the patient has been diagnosed with one or more of 25 so called AIDS defining clinical events. Such AIDS defining infections include Pneumocystitis carinii pneumonica, toxoplasmosis, cryptococcal meningitis and Candida esophagitis. AIDS defining cancers include Kaposi's sarcoma and non-Hodgkin lymphoma (CDC 1992). The life span of a person living with AIDS who is not receiving specific therapy is typically less than 2 years (Vella et al. 1992).

The course of untreated infection can vary widely; while most infected untreated individuals will eventually develop AIDS and die of its consequences, a small percentage of people seroconvert, but maintain high levels of immunocompetence and therefore do not develop progressive disease (such people are generally termed long-term non-progressors). There is also a small group of patients that remain virus-free despite repeated high levels of exposure to HIV-1 (Murphy et al. 2008; Paul 2008).

1.1.7 Antiretroviral therapy

Antiretroviral therapy is based on the hypothesis that maximal suppression of viral replication will prevent or delay the deterioration of the immune system, thus avoiding the development of later stage disease and thus morbidity and mortality. Advances in antiviral therapy (ART) continue to shift the balance of therapeutic risk-benefit balance to earlier treatment as concerns regarding the long-term risks of untreated viremia have increased. There have been significant advances in the areas of potency, toxicity, tolerability and pill burden allowing for better treatment options and compliance (Thompson et al. 2010).

As of May 2010, 24 compounds from different mechanistic classes had been approved for antiretroviral therapy by the US Food and Drug Administration (FDA, http://www.fda.gov) (Table 1.1). A summary of the classes and their method of action follows:

- *Nucleoside/Nucleotide Reverse Transcriptase Inhibitors (NRTIs):* This group of drugs consists of competitive substrate inhibitors, which are analogs of naturally occurring deoxynucleotides. They compete with the natural deoxynucleotides for incorporation into the viral cDNA by the viral reverse transcriptase. NRTIs do not have a 3'-hydroxyl group and once such an analog has been added to a growing viral DNA chain, this chain cannot be extended (a process called chain termination).
- *Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs):* This drug class consists of non-competitive inhibitors which bind directly to the viral reverse transcriptase. The movement of critical protein domains of the enzyme is restricted and thus the ability of the enzyme to synthesize viral cDNA is interrupted.
- *Protease Inhibitors (PIs):* This class of drugs inhibits the activity of the viral protease, which cleaves protein precursors for final assembly into new virions. While treatment with a PI does not prevent the formation of new virions, it prevents the formation of mature and thus infectious virions. The already infected host cell is still destroyed, but the infection of new cells is prevented.

- *Fusion or Entry Inhibitors:* These drugs interfere with the ability of the virus to enter the host cell. By blocking extracellular rather than intracellular interactions, this class of drugs does not have to be transported across the cell membrane to operate. Drugs of this class function by preventing attachment or membrane fusion.
- *Integrase Inhibitors:* Drugs in this group interfere with the viral integrase whose function is to integrate the viral cDNA into the host cell's genome.

	Abbreviation	**Generic name**	**Brand name**	**Date of FDA approval**
Nucleoside/Nucleotide Reverse Transcriptase Inhibitors (NRTIs):	3TC	lamivudine	Epivir	17-Nov-95
	ABC	abacavir	Ziagen	17-Dec-98
	AZT or ZDV	zidovudine	Retrovir	19-Mar-87
	d4T	stavudine	Zerit	24-Jun-94
	ddI	didanosine	Videx EC	31-Oct-00
	FTC	emtricitabine	Emtriva	02-Jul-03
	TDF	tenofovir	Viread	26-Oct-01
Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs):	DLV	delavirdine	Rescriptor	04-Apr-97
	EFV	efavirenz	Sustiva (US)	17-Sep-98
			Stocrin (Europe)	
	ETR	etravirine	Intelence	18-Jan-08
	NVP	nevirapine	Viramune	21-Jun-96
Protease Inhibitors (PIs):	APV	amprenavir	Agenerase	15-Apr-99
	FOS-APV	fosamprenavir	Lexiva (US)	20-Oct-03
			Telzir (Europe)	
	ATV	atazanavir	Reyataz	20-Jun-03
	DRV	darunavir	Prezista	23-Jun-06
	IDV	indinavir	Crixivan	13-Mar-96
	LPV/RTV	lopinavir + ritonavir	Kaletra	15-Sep-00
			Aluvia (developing world)	
	NFV	nelfinavir	Viracept	14-Mar-97
	RTV	ritonavir	Norvir	01-Mar-96
	SQV	saquinavir	Invirase (hard gel capsule)8	06-Dec-95
	TPV	tipranavir	Aptivus	22-Jun-05
Fusion or Entry Inhibitors:	T-20	enfuvirtide	Fuzeon	13-Mar-03
	MVC	maraviroc	Celsentri (Europe)	18-Sep-07
			Selzentry (US)	
Integrase Inhibitors:	RAL	raltegravir	Isentress	12-Oct-07

Table 1.1 Anti-HIV-1 drugs currently approved by the FDA

The choice of the initial drug regimen has long-standing consequences for the patient. In order to optimally treat a given patient, resistance testing should be performed and thereby the predicted virologic efficacy assessed. Typical initial treatment regimens for adults include a combination of two NRTIs and a potent third agent from another class. Such a therapy has been commonly referred to as highly active antiretroviral therapy (HAART), although combination antiretroviral therapy (CART) is currently the preferred term. A common starting regimen is tenofovir and emtricitabine, combined with efavirenz, atazanavir/r or darunavir/r. Another common regimen for treatment-naïve patients is abacavir and lamivudine, plus one of the already mentioned third agents. For the latter regimen, it is important to screen for the HLA allele HLA-B*5701 in the patient to reduce the risk of abacavir hypersensitivity reaction (HSR), as a strong association between this HLA allele and abacavir HSR has been demonstrated in a number of studies. Abacavir HSR occurs in 4-9% of patients and is characterized by multisystem involvement and can be fatal in rare cases. The exact mechanism of this reaction is still unknown (Hetherington et al. 2002; Mallal et al. 2002; Martin et al. 2004). Overall, fixed-dose formulations and once-daily regimens are currently generally preferred (Thompson et al. 2010).

The HIV-1 virus mutates readily and thereby produces drug resistant variants. Three non-independent factors are responsible for this process (Beerenwinkel 2004):

- *Replication and recombination:* HIV-1's reverse transcription is a highly error prone process as it lacks a proof reading mechanism. The mutation rate has been estimated to be 10^{-4} to 10^{-6} substitutions per base pair per replication cycle. This, combined with high titers of virus in untreated individuals and a viral generation time of 1 to 3 days, drives the development of genetic variation. Another source of variability is recombination, where new genomes are formed by template switching by the reverse transcriptase when it encounters two different genomes in a cell infected multiple times.
- *Diversity:* Each individual HIV-1 patient has a complex mixture of genetically related variants.
- *Anti-HIV drug selection pressure:* If the patient is not receiving antiretroviral therapy, the wild type virus and closely related variants are likely to dominate, whereas drug resistant viruses are common in patients which are being treated with antiretroviral therapies. If drug therapy is not successful in completely suppressing replication, the replicating HIV-1 quasi-species can develop new mutations. Some of these new variants will carry mutations which confer drug resistance and these will eventually predominate under drug therapy.

In addition to the development of drug resistant variants, multi-drug resistance is a significant problem. Mutations that confer resistance to one drug are frequently likely to confer resistance to other drugs of that class.

Currently, choosing a drug regimen that produces the maximal and durable suppression of viral replication to <50 RNA copies/ml is considered to be the most important therapeutic goal, as it minimizes the development of drug resistance mutations, preserves CD4 T cell counts and confers clinical benefits to patients (Este and Cihlar 2010).

1.2 HCV

1.2.1 Discovery of the virus and current global burden

The hepatitis C virus (HCV) has low titers in human blood, which made it difficult to detect using conventional methods which existed prior to the late 1980s. Additional complicating factors were that the virus is genetically heterogeneous and the infection has a long asymptomatic phase during the early years of chronic infection (Worman 2002; Houghton 2009).

HCV was first discovered in 1989 using molecular biology techniques i.e. recombinant DNA technology; chimpanzees were infected with the virus and DNA and RNA was isolated from blood specimens of the infected animals. RNA was reverse transcribed to cDNA, a DNA library was constructed from both the DNA and cDNA. Clones of the library were subsequently sequenced and portions of the virus' genome elucidated. Initially this virus was termed non-A, non-B hepatitis, as it was different from both hepatitis A and hepatitis B which had already been described. The same group of researchers at Chiron Corporation also showed in a separate publication in the same issue of *Science* that 85% of patients diagnosed with non-A, non-B hepatitis had antibodies against parts of the newly discovered virus (Choo et al. 1989; Kuo et al. 1989; Worman 2002; Houghton 2009).

HCV is currently recognized as being a major cause of chronic liver disease both in the industrialized and developing world. It has been shown to cause chronic hepatitis, liver cirrhosis and hepatocellular carcinoma. Available data suggests that the prevalence is approximately 2.2-3.0% world-wide, which is approximately 130-170 million people. The highest prevalence of infection is currently found in Africa and the Eastern Mediterranean. It should also be noted that in many countries, HCV infection rates and disease burden have not been examined with population-based epidemiological studies. This makes accurate estimations of the true world-wide impact of HCV impossible (Suzuki et al. 2007; Lavanchy 2009).

1.2.2 HCV genotypes, subtypes and quasi-species

The hepatitis C virus belongs to the family *Flaviviridae*. More specifically, it belongs to the genus *Hepacivirus*, of which it is the only member. Although it shares a similar name with hepatitis A and hepatitis B viruses, HCV is genetically and clinically distinct from these two viruses (2006).

HCV is classified into six major genotypes based on phylogenetic analyses. These genotypes are numbered 1 through 6 according to a standardized nomenclature first proposed in 1994 and later revised and updated (Simmonds et al. 1994; Simmonds et al. 2005). There are several subtypes for each genotype and these are represented by letters. Additionally, quasi-species have been identified based on genetic comparisons. There is substantial sequence variation among the different genotypes (epidemiologically distinct subtypes differ by 20-25%), but all of them share the same structure of linear genes which are also almost of identical size.

Subtype 1a is widely distributed in Northern Europe and the USA, and subtype 1b is the most common world-wide. Subtypes 2a, 2b and 2c are predominantly found in older HCV infected individuals in countries surrounding the Mediterranean and in the Far East. Subtype 3a is also widely distributed in Europe among intravenous drug users. Subtypes 4a and 5a are mainly found in the Middle East and South Africa, respectively. Finally, subtype 6a is

typically found among intravenous drug users in Hong Kong, Vietnam and more recently Australia. A number of additional subtypes have been identified, but they occur more rarely (Simmonds et al. 2005; Timm and Roggendorf 2007).

1.2.3 Structure and genome

The hepatitis C virus contains a positive-sense single-stranded viral RNA genome and is enveloped by a host cell derived lipid bilayer in which two envelope glycoproteins are anchored. The bilayer surrounds the nucleocapsid or core, which is composed of multiple copies of the same core protein and this core in turn contains the viral RNA genome (Figure 1.6). Determining the size of bona fide infectious particles has been hampered by the low amount of virus in blood and tissues. A large number of studies have been performed in a variety of systems and the particle size found was typically 20-60 nm. Infectious virions are likely to be 50-60 nm in diameter (Wakita et al. 2005; 2006).

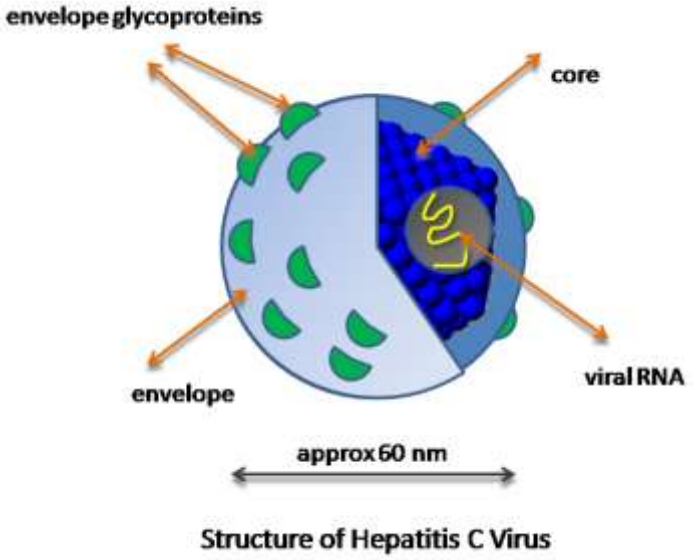

Figure 1.6 Structure of the hepatitis C virus (www.wikipedia.org)

The genome of HCV is approximately 9600 nucleotides in length (Figure 1.7). It is structured as one long open reading frame which is flanked by 5' and 3' non-translated regions (NTRs). The 5' NTR is important for the first step of HCV polyprotein translation. It contains several highly structured domains which contain stem-loops and a pseudo-knot, a subset of which forms an internal ribosomal entry site (IRES). The IRES forms a stable pre-initiation complex by directly binding to the cellular 40S ribosomal subunit and thus initiates the cap-independent translation (i.e. without the need of canonical translation initiation factors) of the viral RNA. The 3' NTR is the initiation site for the synthesis of the negative-strand viral replication and is also involved in translational regulation (2006; Foster et al. 2010).

One long polyprotein is then generated which is subsequently cleaved by host-cellular proteases, as well as virally encoded proteases, into the individual proteins which are either classified as being structural or non-structural. The structural proteins encompass core (C), envelope 1 (E1) and envelope 2 (E2) and the non-structural proteins are p7 (only rarely called NS1), NS2, NS3, NS4A, NS4B, NS5A and finally NS5B (2006; Timm and Roggendorf 2007).

The HCV core protein is highly basic and binds to RNA and is therefore thought to form the viral capsid. The envelope glycoproteins E1 and E2 are located on the surface of the viral particle and partially embedded in the lipid bilayer. They are necessary for viral entry and fusion. E1 and E2 are highly glycosylated, containing up to 5 and 11 glycosylation sites, respectively. E1 is thought to be involved in intra-cytoplasmic virus membrane fusion, while viral attachment is thought to occur via E2. The p7 protein is absolutely required for viral reproduction and behaves like an ion channel when reconstituted in artificial lipid membranes. Recent work has shown that p7 functions as an H(+) ion channel (also termed permeation pathway), acting to prevent acidification in otherwise acidic intracellular compartments (Wozniak et al. 2010). NS2 acts as a protease which cleaves the site between NS2 and NS3. It is a short-lived protein, which subsequently loses its protease activity and is degraded by the proteasome. NS3 in a multifunctional protein: the N-terminal NS3 protease cleaves the remaining non-structural proteins, the C-terminal RNA helicase/NTPase unwinds RNA and DNA. NS4A is a co-factor of NS3 protease activity. *In vitro* studies have indicated that the NS3-NS4A complex may play additional roles in the host cell. NS4B is an integral membrane protein which is thought to act as a scaffold for replication complex assembly, induce membrane association and have several other putative properties. The phosphorylated zinc-metalloprotein NS5A is multifunctional, playing an important role in virus replication by interacting with multiple partners. It is absolutely required for virus replication (Foster et al. 2010). NS5B encodes the viral RNA-dependent RNA polymerase (RdRp) which replicates the viral genome (2006; Timm and Roggendorf 2007).

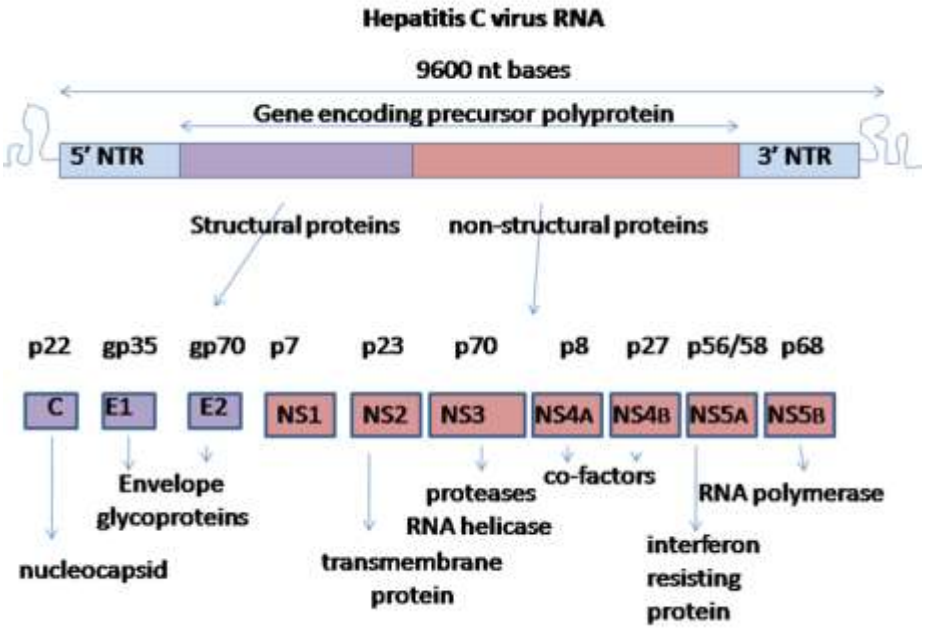

Figure 1.7 The genome of the hepatitis C virus (www.wikipedia.org)

The gene names beginning with "p" or "gp" are indicative of the predominant kDa size of the respective proteins and are part of an older naming system. The current naming system is shown in the boxed, colored genes (e.g. C).

1.2.4 Viral replication

HCV interacts with a number of known and as yet unidentified cell surface receptors leading to binding followed by internalization of the virus particle. The E2 protein of the virus binds to the extracellular loop of CD81 (a tetraspanin expressed on different cell types including hepatocytes) with high affinity. Additionally, the scavenger receptor class-B type-I (SR-BI), low-density lipoprotein receptor, glycosamino-glycans and the tight junction proteins claudin-1 and occluin are necessary for viral entry into the cell (Sabahi 2009; Gouttenoire et al. 2010).

Once attachment has occurred, fusion takes place between the viral membrane and the cellular membrane. Alternatively, particles can also be internalized via endosomes, after which membrane fusion occurs (Figure 1.8). Decapsidation of the viral nucleocapsid occurs next, allowing the now free positive-strand genomic RNA to enter the cell's cytoplasm. These RNAs and newly synthesized RNAs serve as mRNAs for the synthesis of the viral polyprotein. The viral 5' NTR's IRES controls the translation of the HCV genome and recruits both cellular and viral proteins to perform this task. Once the large precursor polyprotein (approximately 3011 amino acids in length) has been generated, it is targeted to the ER membrane via an internal signal sequence located between the core and E1 proteins. There the ectodomain portion of E1 protein is translocated into the ER lumen. Cleavage of the polyprotein by cellular and viral proteases follows, producing the viral proteins. The non-structural HCV proteins then recruit the viral genome into an RNA replication complex, which is associated with rearranged cytoplasmic membranes. NS4B is responsible for inducing a specific membrane alteration, termed the membranous web, which serves as a scaffold for the replication complex. The replication complex is composed of viral proteins, replicating RNA and altered cellular membranes and is considered to be a hallmark of all positive-strand RNA viruses. The precise mechanism of HCV replication is not well understood, but it takes place via HCV's RNA-dependent RNA polymerase NS5B that produces a negative strand RNA intermediate. This intermediate is the template for the generation of positive strand RNA genomes which are assembled into new viral particles or serve as templates for further translation or replication. The mechanisms of viral assembly and the release of newly formed viral particles are also poorly understood. Of the non-structural proteins, most if not all are thought to be involved in this process. Additionally, lipid droplets and the very low density lipoprotein pathway have been recently shown to play a central role in the production of mature virus particles (2006; Gouttenoire et al. 2010).

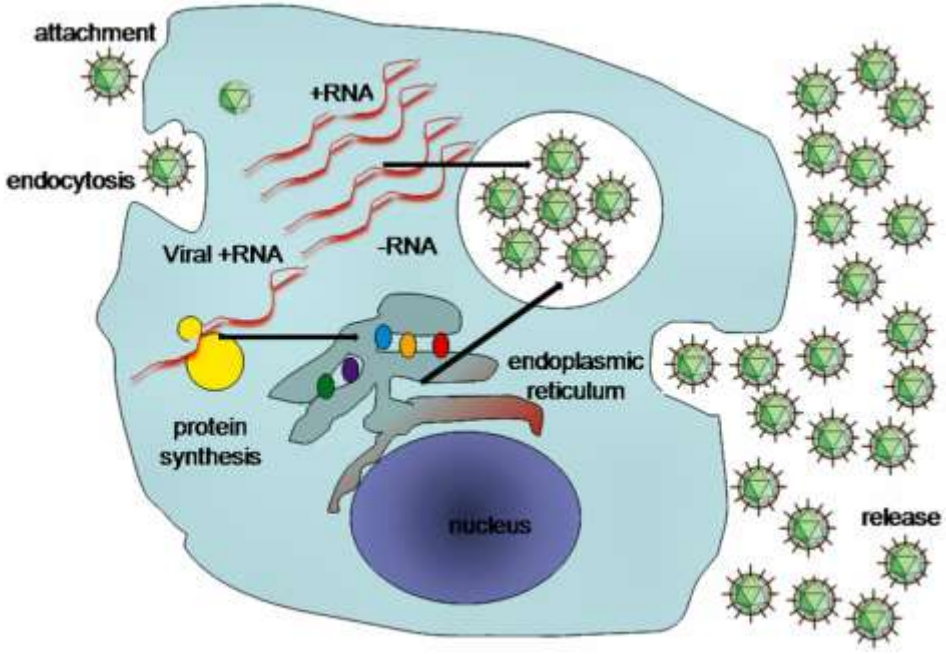

Figure 1.8 Replication of the hepatitis C virus (www.wikipedia.org)

1.2.5 Transmission

HCV is primarily transmitted via parenteral routes, i.e. through piercing the skin or mucous membranes. Other methods of transmission are the mucosal exposure to blood or serum-derived fluids, as well as environmental sources (e.g. exposure to dried blood). Before 1990 and the effective screening of donated blood, many cases occurred due to the transfusion of HCV-contaminated blood products. In some countries such as the USA, intravenous drug use is currently a common method of transmission. High-risk sexual activity has also been associated with transmission, as has mother to infant transmission. However, 20-40% of patients do not have an identifiable risk factor for HCV (2009).

It has been shown that different viral subtypes are associated with particular transmission routes. The transmission of genotypes 1b and 2 is associated with transfusions, 1a and 3a with intravenous drug use and 4a with unsafe injections in Egypt (Magiorkinis et al. 2009).

1.2.6 Chronic infection

Most patients who contract HCV are generally asymptomatic in the phase of acute infection. Once the HCV RNA viremia persists for six months or more, a patient is considered to be suffering from a chronic infection. 54-90% of patients progress from acute to chronic infection (some of this variability is likely due to differences in study design). Patients suffering from chronic HCV infection have an increased risk of liver cirrhosis, hepatocellular carcinoma and death. In retrospective studies, 20-51% of patients with chronic HCV infection developed liver cirrhosis within 20 to 30 years of infection, 2-11% developed hepatocellular carcinoma and 4-15% died of liver-related causes (2009).

1.2.7 Antiviral therapy

The current standard of care for treating chronic HCV infection is a combination of peginterferon and ribavirin. Peginterferon is an antiviral drug consisting of interferon alpha-2a to which polyethylene glycol (PEG) has been added. The process of adding PEG (also termed pegylation) extends the half-life of interferon. Peginterferon acts as an antiviral, blocking viral entry into cells and viral replication. It also enhances the immune response. Ribavirin is a guanosine analog which is also thought to inhibit viral replication. However, the mechanism by which this combination therapy eradicates the virus is not yet well understood (Ghany et al. 2009; Shiffman 2010).

Peginterferon/ribavirin combination therapy yields a sustained virologic response (SVR) in 40-45% of patients with HCV genotype 1a/1b and an SVR of 75-80% in patients with genotype 2. The other HCV genotypes have SVRs somewhere between those of genotypes 1a/1b and 2. As most patients are infected with genotype 1a/1b, the majority of patients will fail to achieve SVR using the current treatment method (Shiffman 2010).

Today, patients are treated using response-guided therapies in which the duration of therapy is adjusted based on the time to response (as opposed to older dosing regimens of fixed-duration therapy based on genotype). For example, patients with a rapid SVR and a low rate of relapse can be treated with peginterferon/ribavirin for 24 weeks irrespective of viral genotype. Patients that take longer to respond receive a longer duration of therapy. Non-responders can be identified after 12 weeks and their treatment stopped (2009; Shiffman 2010).

Currently, two protease inhibitors are being evaluated in phase 3 clinical trials and a large number of both protease and polymerase inhibitors are in phase 2 clinical trials. While protease inhibitors are highly potent and frequently reduce RNA viremia by 3 to 6 logs, the rate of viral resistance is high. In contrast, polymerase inhibitors reduce viremia by 2 to 4 logs and have a lower rate of viral resistance when used as monotherapy. It is anticipated that a protease or polymerase inhibitor (also termed direct-acting antiviral agent or DAA) will receive US FDA approval in mid to late 2011. It is likely that patients will then begin to receive triple therapy consisting of peginterferon, ribavirin and a DAA as a first line therapy. It is expected, based on current data and data from ongoing clinical trials, that an SVR of 70% will be achievable with triple therapy (Shiffman 2010).

1.3 MHC molecules

In the aftermath of World War II, the need to treat extensively burned airmen kick-started the modern era of transplantation biology. It was observed that the second graft from a given donor (allograft) was rejected more rapidly and vigorously by its host than the first graft, whereas skin transplanted from a new donor would be rejected with the kinetics of a first set reaction (Delves et al. 2006). A number of scientists, most notably Peter Medawar, began to investigate the immunological basis of this process (Male et al. 2006). This eventually led to the discovery that a family of molecules existed which were so variable that it was unlikely that two unrelated individuals would have the same combination of variants. The human major histocompatibility complex (MHC) is called the human leukocyte antigen (HLA) system and is used by the immune system to recognize non-self, to which it generally responds with great ferocity (DeFranco et al. 2007).

The HLA system consists of a 4-Mb gene region on the short arm of chromosome 6 (6p21.3) which contains over 220 genes of diverse function (Stephens et al. 1999; Horton et al. 2004; Robinson et al. 2009). These genes encode proteins which display antigen to T

lymphocytes. Of central importance in the protection against pathogens are the MHC class I genes (particularly HLA-A, HLA-B and HLA–C) and MHC class II genes (particularly HLA-DR, HLA-DQ and HLA–DP). Class I genes encode molecules that are expressed on the surface of all nucleated cells at varying levels. These molecules bind peptide fragments of cytoplasmic and secreted proteins (Figure 1.9) and present them to CD8 T cells, leading to a cytotoxic T cell response. An important function of these MHC molecules is to signal viral infection.

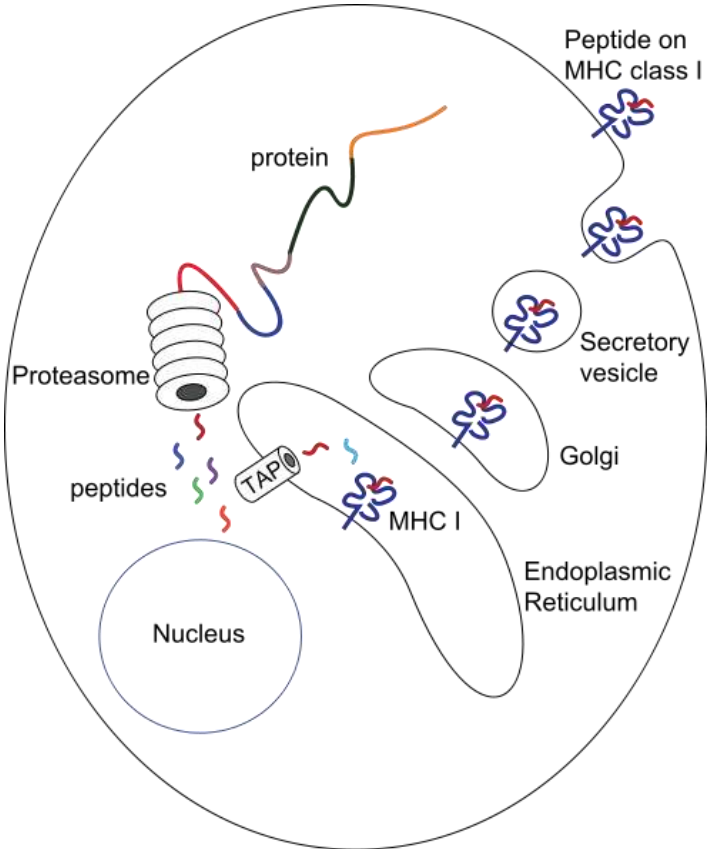

Figure 1.9 The process by which peptides are generated, loaded onto and presented by MHC class I molecules (www.wikipedia.org)

Cytoplasmic protein degradation by the proteasome, transport into endoplasmic reticulum by the TAP complex, loading on MHC class I molecules, and transport to the surface for presentation to CD8 T cells

MHC class II molecules are mainly expressed on antigen presenting cells (APCs) such as B lymphocytes, macrophages and dendritic cells. Class II molecules are specialized to sample the endosomal-lysosomal system and therefore present peptides of extracellular origin to CD4 T cells, resulting in cytokine production and T cell help in antibody production (Figure 1.10). Supporting the activities of these molecules are the class III complement proteins and the inflammatory cytokine genes, which are also located in the HLA region in humans (Beerenwinkel et al. 2007; DeFranco et al. 2007).

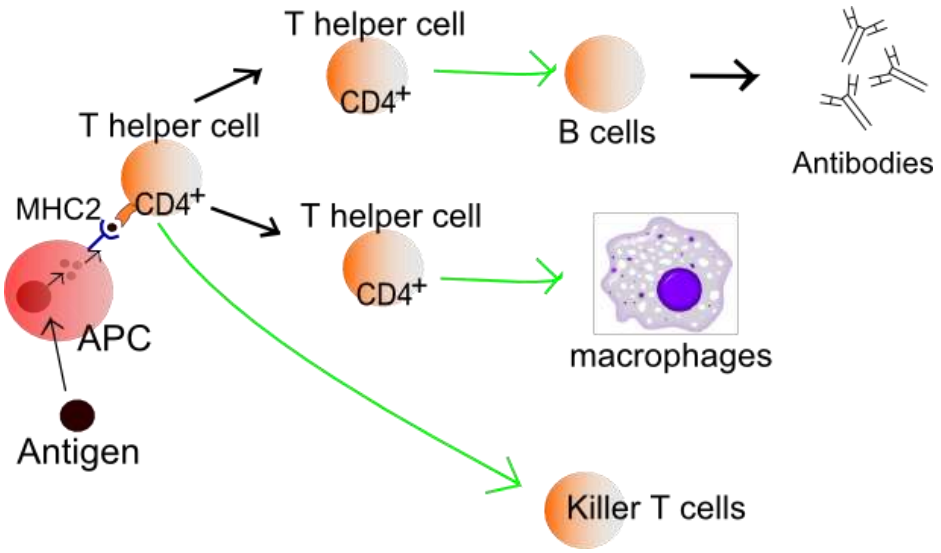

Figure 1.10 MHC class II activation of CD4 T helper cells (www.wikipedia.org)

As already noted, HLA class I and class II genes are extremely polymorphic (Robinson et al. 2009). As of July 2010, 3,937 class I alleles and 1,253 class II alleles have been named and the growth in the total number of observed alleles is projected to continue for some time. Additionally, HLA-B has been confirmed as the most polymorphic gene in the human genome (Mungall et al. 2003). Allelic variation occurs both within and between different ethnic groups (Middleton et al. 2003; Solberg et al. 2008). Linkage disequilibrium, where the alleles at one HLA locus are not randomly distributed with respect to the alleles at another HLA locus, has been described between different loci in the HLA region (Miretti et al. 2005; Beerenwinkel et al. 2007).

Due to the high number of polymorphisms in the HLA genes, most individuals are likely to be heterozygous at most polymorphic loci. The expression of HLA alleles is co-dominant, as both alleles of a locus are expressed on a cell's surface and each allele is able to present peptides to T cells (Figure 1.11). Polymorphisms in HLA molecules are primarily concentrated among amino acids which are responsible for the binding of the foreign peptide. These amino acids are located on the floor or on the inner walls of the peptide-binding cleft. The polymorphisms cause the clefts to have different sizes and chemical characteristics in different allele variants. Therefore, although all HLA molecules can bind large and diverse sets of peptides, different HLA molecules have preferences in their binding affinities and specificities (Beerenwinkel et al. 2007; Murphy et al. 2008).

Groups of HLA-A, -B and -DR alleles, termed supertypes, have been identified which share specific binding preferences for peptides or supermotifs of a similar size, charge and amino acid composition. Clustering HLA molecules into such supermotifs facilitates epitope identification and vaccine design (Sette and Sidney 1999; Lund et al. 2004; Sidney et al. 2008).

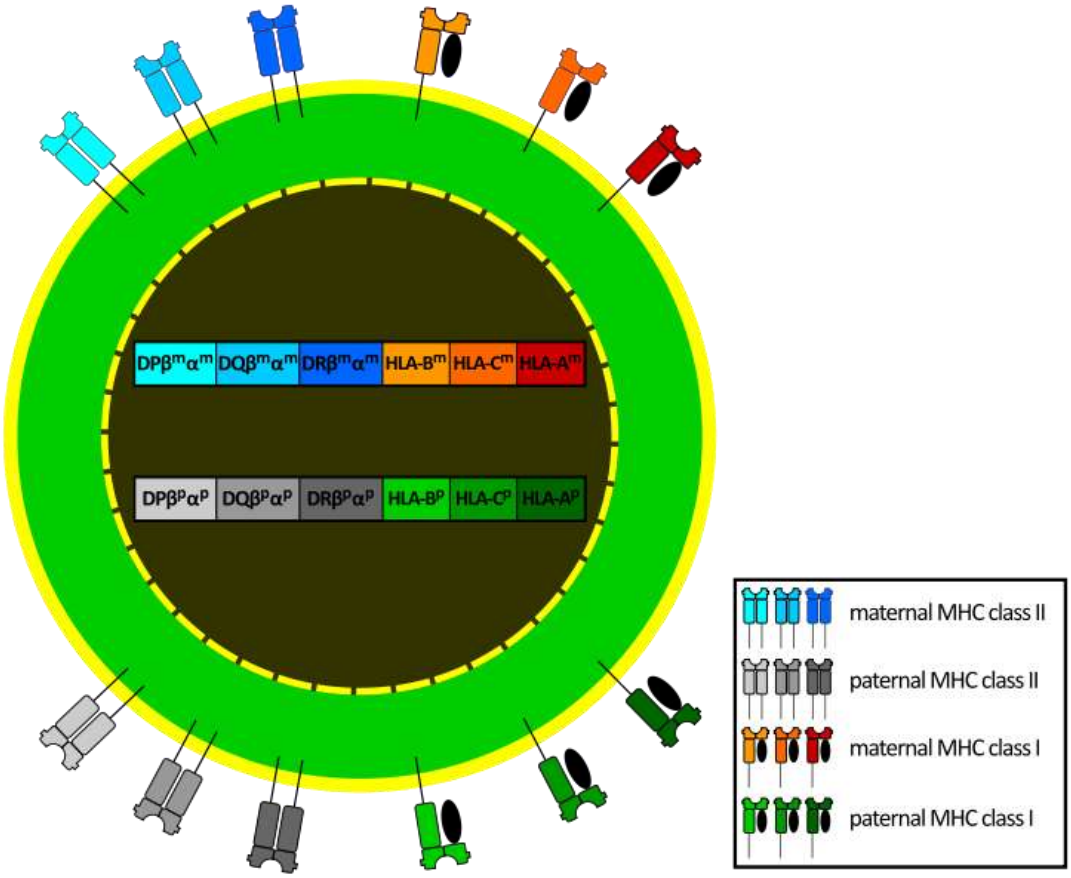

Figure 1.11 The co-dominant expression of MHC alleles (www.wikipedia.org)

1.4 T Cell Receptors and T Cell Development

1.4.1 Lymphocyte development

In mammals, lymphocyte development mostly occurs in specialized areas called the central lymphoid organs: bursa-dependent (B) cells develop in the bone marrow and in the liver of fetuses, thymus-dependent (T) cells develop in the thymus (Murphy et al. 2008). Bursa-dependent refers to the bursa of Fabricius in birds, which is an outpouching of the cloaca, where these cells were originally described. Both fetuses and juvenile mammals produce large numbers of lymphocytes which move into peripheral lymphoid tissues and populate them. In adult mammals, B cells are still continually formed in the bone marrow, but T cells are only formed at a much reduced rate in the thymus. T cell numbers in adults are instead maintained by mature T cells dividing outside of the central lymphoid organs (Paul 2008).

T cell precursors originate in the bone marrow where they develop from a common lymphoid progenitor, which also gives rise to B cells (Murphy et al. 2008). Subsequently, these T cell precursors move from the bone marrow to the thymus, where they eventually mature (von Boehmer et al. 2003; Werlen et al. 2003; Hogquist et al. 2005; Siggs et al. 2006). The thymus contains a network of epithelia called the thymic stroma, which provides a microenvironment for the developing bone marrow-derived T cell precursors which are committed to the T cell lineage. These T cell precursors, having arrived in the thymus, undergo differentiation for one week prior to undergoing a period of intense proliferation.

Only a small proportion of these cells (less than 1-3%) survive apoptosis and leave the thymus as mature T cells (Scollay et al. 1980; Egerton et al. 1990; Goldrath and Bevan 1999); the vast majority are ingested by macrophages in the thymic cortex. This widespread apoptosis is a crucial aspect of T cell development, as it represents the process of negative selection (also called clonal deletion). During this process, double-positive (having the surface markers CD4 and CD8) or single-positive (having only the surface marker CD4 or CD8) thymocytes which express T cell receptors (TCRs) with a high affinity for self antigens are eliminated (Starr et al. 2003; Klein et al. 2009). Contact occurs between the TCRs of the developing thymocytes and the cortical stroma, which is composed of epithelial cells with long branching processes that express both MHC class I and MHC class II molecules on their surface. This intense screening allows each new T cell to be tested for its ability to recognize MHC-self peptide complexes and if the affinity is too strong the cell is destroyed, thereby building self tolerance. Additionally, thymocytes which express TCRs with little or no affinity for MHC-self peptide complexes die of neglect. Therefore, only thymocytes with TCRs which have an intermediate affinity allow for positive selection: immature double-positive thymocytes with intermediate affinity and/or avidity for self-peptide-MHC complexes are induced to differentiate into mature single positive thymocytes (DeFranco et al. 2007; Paul 2008).

Commitment to the CD4 or CD8 lineage is accompanied by commitment to either the helper or cytotoxic T cell function, which occurs due to the differential expression of the T-helper-inducing POZ/Kruppel-like factor (Th-POK) or runt-related transcription factor (RUNX) transcription factors (Collins et al. 2009). After CD4 or CD8 lineage commitment, where double positive thymocytes become single positive thymocytes, cells are exported outside of the thymus. Cells of the CD4 helper T cell lineage recognize MHC class II molecule complexed with antigenic peptides, whereas the CD8 cytotoxic T cell lineage recognizes MHC class I molecule peptide complexes. Each lineage therefore retains the co-receptor necessary for recognizing the type of MHC molecule preferentially bound by the TCR of that cell type. When cells of both lineages are exported to the periphery they become part of the peripheral T cell repertoire. The time period between the entry of the progenitor T cells into the thymus and the export of mature cells into the periphery has been estimated to be 12 days (DeFranco et al. 2007).

Recent work suggests that maturing thymocytes need to have repeated engagements with antigen presenting cells (APCs) as they travel through the thymus. Positive selection occurs in cases where there is continuous low-level stimulation through a series of brief encounters with MHC-self peptide complexes presented by the thymic stromal cells. This way, developing thymocytes sample many cell surfaces, increasing the chance that they may encounter a negatively selecting ligand (Ebert et al. 2008).

1.4.2 Gene segment rearrangements in maturing T cells

Each mature T cell expresses a unique antigen receptor on its surface; during the maturing of the cells in the thymus, a sequential rearrangement of antigen receptor gene segments occurs which ultimately generates this unique antigen receptor. All jawed vertebrates appear to contain the same four homologous TCR isotypes: TCRα, TCRß, TCRγ and TCRδ (Rast et al. 1997). There are two subsets of T cells, differentiated by the exact pair of receptor chains that they express. These are either the TCRα and TCRß pair or the TCRγ and TCRδ pair. Most cells express the α:ß heterodimeric receptor and hence the focus of this work will be on this more common receptor type. Cells expressing the α:ß TCR typically recognize peptide antigens which are presented on the MHC encoded molecules, while cells

expressing the γ:δ heterodimer are thought to be either restricted to non-classical MHC class I molecules or bind free antigen, such as invariant molecules released by damaged cells (Brenner et al. 1986; Clevers et al. 1988; Hayday 2000).

The human loci which encode the T cell receptor subunits consist of separate variable (V), diversity (D), joining (J) genes, as well as constant (C) genes. The locus for human TCRα is located on chromosome 14q11.2 and spans 1000 kb. The germline organization consists of 54 V genes, of which only 44 to 47 genes are functional. Further 3' in the locus, a cluster of 61 J genes exists, followed by a single C gene. The genes of TCRδ are nestled within the TCRα locus, located between the V and J genes (Lefranc 2001; Giudicelli et al. 2005).

The locus for TCRß is located on 7q35 and spans 620 kb. The germ-line organization of the TCRß locus is somewhat different from TCRα: 76 V genes exist, not all of which are functional, followed by a single D gene, 7 J genes, a single C gene, another D gene, 7 J genes and finally a single C gene. The C genes of the TCRß locus consist of separate exons which encode a number domains. The genes of TCRδ are nestled within the TCRα locus, located between the V and J genes. (Lefranc 2001; Giudicelli et al. 2005).

The nomenclature for human T cell receptor genes which will be used here is consistent with that proposed by the developers of the International IMmunoGeneTics database (IMGT) (Lefranc 2001; Giudicelli et al. 2005) and which was also approved by the Human Genome Organization (HUGO) Nomenclature Committee in 1999. For example, TRAV7 indicates V gene 7 of TCRα, whereas TRGC1 indicates C gene 1 of TCRγ.

In both human and mice, TCRδ is the first locus to rearrange during the maturation of T cells in the thymus. This is closely followed by the rearrangement of TCRγ and TCRß loci and finally TCRα rearranges (Blom et al. 1999; Livak et al. 1999; Joachims et al. 2006).

A mature T cell expresses either α:ß or γ:δ TCR heterodimers on its surface and the process by which a bipotent precursor cell commits to one of these two lineages is termed lineage commitment. The exact process through which this occurs is unknown and a number of models have been proposed to explain it (Joachims et al. 2006). These models include instructive model in which precursor cells are bipotent, but the development of a functional γ:δ TCR leads to the development of a cell with the γ:δ heterodimer on its surface. Alternatively, if a so-called pre-TCR complex forms, the cell becomes a α:ß cell (Dudley et al. 1995; Livak et al. 1999). In the stochastic or separate lineage model, the commitment is thought to precede gene rearrangement (Winoto and Baltimore 1989; Kersh et al. 1995; Sim et al. 1995). And finally, the more recent signal strength model states that the type of cell which develops is determined by the strength of signals transduced by TCRs: weak signals lead to the development of α:ß cells and strong signals produce γ:δ cells (Haks et al. 2005; Hayes et al. 2005).

Once germline rearrangement is completed the TCRα locus consists of a V gene, a J gene and C gene, which produces the mRNA encoding the TCRα chain. The rearranged TCRß consists of V, J, D and C genes which also produce one mRNA encoding the TCRß (Krangel 2009). The hypervariable regions in both TCRα and TCRß are known as CDR1, CDR2 and CDR3, where CDR stands for complementarity determining region. These hypervariable regions participate in antigen binding and determine the structural complementarity of the receptor to the antigen. CDR1 and CDR2 are encoded by the V gene, whereas CDR3 is encoded by the J gene (TCRα) or the D and J genes (TCRß). The constant region (encoded by the C gene) does not participate in antigen binding and does not vary between cells which have different antigen specificities (DeFranco et al. 2007).

TCR diversity is generated in a number of different ways, one of which is the assembly of different V-J or V-D-J combinations. Also, the dimerization of different TCRα and TCRß chains is promiscuous, although not all TCRα and TCRß isotypes are able to form

dimers due to structural constraints. However, the most significant source of TCR diversity is a process by which the sequences at the junctions of the V, D and J genes are partially digested and random nucleotide sequences of variable length are inserted during junction assembly via an enzyme complex called the V(D)J recombinase. Proteins that make up the V(D)J recombinase complex and that are important for this process are RAG-1 and RAG-2, which are specifically found in lymphoid cells (Melek et al. 1998; Gellert 2002; Krangel 2009).

Interestingly and controversially, some authors believe that TCR revision is possible once the mature T cells have left the thymus. TCR revision is the process by which mature peripheral self-reactive T cells (which are thought to only exist in very small numbers, having escaped the selection process in the thymus) lose surface expression of their TCR, reinduce the expression of the recombinase machinery, then rearrange the genes which encode extrathymically generated TCRs for antigen and finally express these new receptors on the cell surface (Hale and Fink 2010).

1.4.3 Antigen binding

Because the activation of naïve T cells in the face of infection requires sustained signaling from the TCR, the TCR must bind the MHC-peptide complex with high enough affinity to ensure such sustained signaling is possible (Carreno et al. 2006). A number of structures of TCRs bound to their MHC-peptide ligands have been reported, where the interaction occurs between the MHC-peptide and CDR loops of the TCR: the CDR1α, CDR2α, CDR1β and CDR2β loops which are germ-line encoded and the CDR3α and CDR3β loops, which are only partially germ-line encoded. TCRs often lie in a diagonal, across the face of the MHC-peptide antigen with the interaction partners typically being as follows: CDR1α, CDR2α are usually located facing the α2 helix of the MHC class I or the β1 helix of the MHC class II molecules; CDR1β and CDR2β are usually located facing the α1 helix of the MHC class I or MHC class II molecules; CDR3α and CDR3β have interactions with the peptide. The angle and pitch of the TCR versus the MHC-peptide varies because of the differing bound peptides and the variable CDR3 sequences (Marrack et al. 2008).

The interaction sites between the TCR and MHC-peptide are inherently flexible and can be difficult to identify. Rules that govern the reactions between TCR and MHC-peptide, apart from the usual diagonal orientation and the placement of the TCR V region loops over the α-helices of the MHC, have not been apparent. In a detailed study of CDR1 and CDR2 loops in human and mouse complexes, Marrack et al., identified built-in biases in the TCR variable regions which bind to MHC-peptide antigens. The authors state that nearly all structures in the analysis would have been subjected to normal negative selection and therefore unlikely to demonstrate all their built-in abilities to react to MHC-peptide. Fourteen MHC class I complexes (8 human and 6 mouse) were examined, as were 8 MHC class II complexes (4 human and 4 mouse). This dataset included all structures which were available at the time, excluding duplicates of particular TCR-MHC-peptide combinations and ones analyzed in a previous publication (Feng et al. 2007). The dataset was also heavily biased with respect to the use of specific TCRβ subunits, both in the case of human (TRBV6 or Vβ13 using an older nomenclature system) and mouse (Vβ8), which can be expected to skew the identified similarities. Positions in the TCR CDR loops which commonly interacted with the MHC-peptide commonly were:

- CDR1α amino acid positions 28, 29 (frequently Y), 31 (frequently Y/F)
- CDR2α amino acid positions 50 (frequently Y), 51 (frequently S), 52
- CDR1β amino acid positions 28, 29 (frequently N/Y)
- CDR2β amino acid position 46, (frequently Y/F), 48 (frequently Y), 54 (frequently D/E)

Marrack et al. also observed that some of these TCR amino acids tended to interact with particular MHC class I amino acids:

- CDR1α amino acids positions 29, 31 tended to interact with MHC class I α2 domain's amino acid in position 158
- CDR2α amino acids positions 50, 51 tended to interact with MHC class I α2 domain's amino acid in position 158
- CDR1β amino acids positions 28, 29 tended to interact with MHC class I α1/α domain's amino acid in position 69
- CDR2β amino acids positions 46, 48 tended to interact with MHC class I α1/α domain's amino acid in position 69

For MHC class II amino acids the pattern was similar:

- CDR1α amino acids positions 29, 31 tended to interact with MHC class II β1 domain's amino acid in position 73
- CDR2α amino acids positions 50, 51 tended to interact with MHC class II β1 domain's amino acid in position 73
- CDR1β amino acids positions 28, 29 tended to interact with MHC class II α1 domain's amino acid in position 64
- CDR2β amino acids positions 46, 48, 54 tended to interact with MHC class II α1 domain's amino acid in position 64

The authors theorized that the rules did not apply consistently in all cases due to differences in the CDR3 regions of the TCR (which were not included in the study).

The roughly diagonal binding orientation (±75º) of α:β TCR on the surface of the MHC-peptide is similar in most structures that have been crystallized so far. However, because there is some variation in the binding angle it has been difficult to identify all residues involved in the interaction of the complex (Wucherpfennig et al. 2010). Yet some researchers believe that gaining such an understanding may not even be necessary, as some TCR V genes may have coevolved with MHC genes and thus the TCR V-domains contain residues which have preferred contacts with the MHC molecule. This hypothesis has been around for some time (Jerne 1971) and a number of cellular, functional and biochemical studies support it, yet it remains controversial.

It is thought that invariant germ-line V genes, which encode both CDR1 and CDR2, are involved in such germ-line encoded recognition, while somatically rearranged CDR3, which is encoded by the J gene (TCRα) or the D and J genes (TCRß), is not. Thus, both CDR1 and CDR2 are susceptible to evolutionary selection, unlike the somatically generated CDR3 loops; the "invariant" regions of the TCR and MHC are paired, as are the "variable" regions of the CDR3 and the peptide. It has been calculated that 75-80% of the binding surface involves contacts between CDR1, CDR2 and the MHC which may also have encouraged such a coevolution (Garcia et al. 2009).

Garcia et al. analyzed structures produced by two previous studies (Feng et al. 2007; Dai et al. 2008). They examined the seven complexes:

- 3 V_β8.2 TCRs complexed with mouse I-A^u MHC molecules
- 2 V_β8.2 TCRs complexed with mouse I-A^b MHC molecules
- 1 V_β8.1 TCR complexed with mouse I-A^b MHC molecule
- 1 V_β8.2 TCR complexed with mouse I-A^k MHC molecule

Not only do the MHC alleles vary, as described above, but there are different peptides (MPB1-11, Con A and 3K) and the TCR Vα chains (Vα4.2, Vα4.1, Vα3.1 and Vα2.3) in the complexes differ as well. The Protein Data Bank accession numbers of six structures in the study were 1U3H, 2Z31, 2PXY, 1D9K, 3C60 and 3C61 (one structure did not have its accession number specified). Superposition shows close conversion of the CDR1β and CDR2β contacts with the I-A α1 chain helix. All complexes show a similar "knob-in-the-hole" complementarity in which the Y48 of CDR2β occupies a depression in the α1 helix of the I-A MHC molecule. This interaction is surrounded by a halo of side-chain interactions and the authors found this coincidence very striking, despite the differences in the Vα chain, the peptides and the differing MHC alleles in each complex. Garcia et al. consider this to be the strongest evidence to date that this interaction is encoded at the level of pair-wise interactions.

Work confirming this hypothesis was performed in mice which expressed a single rearranged TCRβ chain with individual mutations of the Y46, Y48 and E54 residues (Scott-Browne et al. 2009). Mutations of these residues into alanine substantially reduced the development of the entire T cell repertoire. The phenotype was particularly strong in the case of the Y46 or Y48 mutations. Interestingly, the lower binding affinity of the Vβ segment led to the pairing of TCRβ chains with higher affinity TCRα chains. The authors conclude that their work demonstrates that thymic selection is controlled by specific germ-line encoded contact points in the MHC and that the near obligate requirement for these amino acids indicate that these amino acids might confer "generic" MHC reactivity to TCRs.

1.4.4 T cell receptor cross-reactivity

It has been shown that there are far more potentially immunogenic peptides in the environment than there are T cells in an individual at any point in time (Mason 1998). The stringent positive selection process in the thymus of developing T cells leads not only to the survival of cells which have maximum sensitivity for recognizing non-self peptides, but also to T cells which have a low affinity for MHC molecules bearing peptides derived from host proteins (Starr et al. 2003). Therefore the mechanism of TCR cross-reactivity (also sometimes referred to as alloreactivity or degeneracy) was proposed to explain the greater effective size of the TCR repertoire (Yin and Mariuzza 2009).

There are currently five conceptually distinct mechanisms which have been identified, which help explain how cross-reactivity works at the atomic level (Yin and Mariuzza 2009): (1) induced fit where the binding site is conformationally flexible allowing it to bind to multiple MHC peptide antigens (Mazza et al. 2007), (2) differential TCR docking where the same TCR binds to different MHC-peptide antigens using different docking orientations (Colf et al. 2007), (3) structural degeneracy where suboptimal complementarity between the TCR and peptide can be improved by variations in the peptide (Li et al. 2005), (4) molecular mimicry where different MHC-peptide ligands with very similar interfaces can be bound by a cross-reactive TCR (Harkiolaki et al. 2009; Macdonald et al. 2009) and (5) antigen-dependent

tuning of MHC-peptide flexibility where the conformational dynamics in the ligand allow for it to be recognized upon TCR binding (Borbulevych et al. 2009).

The "codon hypothesis" (Garcia et al. 2009) states that germ-line TCR-MHC interactions can occur in two ways. First, each TCR can bind to different MHCs with low affinity, yet in a specific manner and therefore the binding orientation can differ, dependent on the binding partner; a TCR gene product (Vx) can bind multiple MHC molecules (Y, X or Z) and engages them in different highly specific ways. Second, a given TCR can bind to the same MHC molecule with entirely unrelated chemistries; a TCR gene product (Vx) can bind to the same MHC (Y) using several distinct codons (A, B or C) and this is thought to be influenced by the CDR3 sequence and the peptide. This would explain how different peptides bound to the same MHC, produce very different docking footprints with the same TCR (Feng et al. 2007; Dai et al. 2008). In this manner, the peptide is involved in determining the final docking geometry as it selects from a limited menu of predefined codons.

1.4.5 TCR/MHC-peptide binding geometry

In MHC class I and class II structures, the peptides are presented in an extended conformation in a vice-like groove. The bound peptides are flanked by two α-helices and the floor of the binding groove is composed of an anti-parallel β-strand. The ends of the peptide binding groove are closed in class I structures, limiting the length of the peptide which can be bound, while in class II structures the groove is open, which allows for much longer peptides to be bound (Murphy et al. 2008). The TCR heterodimer is generally oriented diagonally relative to the long axis of the MHC peptide binding groove (Garboczi et al. 1996; Garcia et al. 1996). The Vα of the TCR is positioned above the N-terminal half of the peptide and the Vβ of the TCR is located above the C-terminal half of the peptide. Previous analyses have shown that most contacts between the TCR and the peptide occur via the highly variable CDR3 loops. Contacts between the TCR and MHC are mediated mostly through the CDR1 and CDR2 loops (especially those of the Vα domain), whereas CDR3 make a smaller contribution in the number of conserved contacts. The binding orientation is thought to be driven through long-range electrostatic steering or through a low-affinity binding event, subsequent to which the CDR loops maximally mold to the MHC-peptide providing the final docking orientation (Rudolph et al. 2006).

There has been a number of different approaches for describing the TCR to MHC-peptide binding orientation or how the TCR crosses the MHC binding platform (also referred to as the crossing angle). One widely applicable recent approach has been proposed by Rudoph et al. in which relative rather than absolute values are used in the calculations to define the general principles. The crossing angle is determined by calculating a vector along the MHC axis, drawing a best-fit straight line through the Cα atoms of the two MHC α-helices. For MHC class I, these are the Cα atoms A50-A86 and A140-A176. A second vector describing the long axis of the TCR binding site is calculated by constructing a vector between the centroids of the conserved Ig disulfide-forming sulfur atoms in TCR α:β heterodimers (L22-L90 and H23-H92). The crossing angle between the MHC and TCR vectors is the dot product of the two vectors. Using the described method, the crossing angles in TCR-MHC class I peptide complexes were found to be roughly diagonal and range from 21º to 70º. Additionally, there is also significant variation in the tilt and role of the TCR on top of the MHC (Rudolph et al. 2006).

1.4.6 T cell activation

T cell activation occurs in the secondary lymphoid tissue, where recirculating naïve T cells bind to mature activated dendritic cells. Dendritic cells are specialized cells which ingest debris and infectious agents in peripheral tissues and then migrate into the lymphoid tissue. Some dendritic cells are resident in the lymphoid tissue and function as phagocytes before maturing into APCs (DeFranco et al. 2007).

Once TCR ligation has occurred an intracellular signaling cascade is initiated; several pathways are activated within the T cell leading to T cell activation and proliferation (Smith-Garvin et al. 2009). The exact mechanism of initial TCR triggering is unknown and is still hotly disputed (Fooksman et al. 2010; Ma and Finkel 2010). Subsequent to initial triggering, TCR signaling is mediated via the immunoreceptor tyrosine-based activation motifs (ITAMs) located in the cytoplasmic tails of the CD3 molecule. Then, several major pathways in the T cell are activated: (1) the protein tyrosine kinase Lck phosphorylates the ITAM tyrosines, which then recruit and phosphorylate the zeta chain associated protein of 70 kDa (ZAP70), eventually leading to the assembly of a large protein complex built on the linker for activation of T cells (LAT) protein (2) an increase in the intracellular calcium levels and (3) the production of diaglycerol (DAG) followed by the activation of pathways involving the key signaling molecule Ras and protein kinase C isoforms (PKCθ). Subsequently, a proliferation of activation and a clonal expansion of the T cells occurs via the induction of cytokine gene transcription. CD8 cytotoxic T cells mature and acquire the capacity to lyse target cells through the release of cytotoxic molecules. CD4 helper cells activate other cells including phagocytes, mast cells, basophiles, eosinophils and B cells, which then go on to differentiate into antibody producing cells (Murphy et al. 2008).

1.4.7 Modeling the thymic selection of T cell receptors

T cell immune responses are dependent on the appropriate and effective processing of peptides from a protein source, the stable binding of the peptide to the MHC molecule and recognition of this complex by the TCR. Of these three steps, the binding of the peptide to the MHC is the most selective event. There is a large body of work from last 20 years, which has examined MHC-peptide binding (Lafuente and Reche 2009). The study of a representative cross-section of such methods has been shown them to have high prediction accuracies (Roomp et al. 2010).

However, there is still a significant proportion of stable MHC peptide complexes which elicit no TCR response. Comparatively little work has been performed on predicting which TCR will bind to MHC-peptide complexes. Simplified representations of amino acids, represented as a string of numbers or bits, have been used to study negative selection and TCR cross-reactivity (Detours et al. 1999; Detours and Perelson 1999; Chao et al. 2005).

A more recent approach has been developed in which the problem was studied numerically but still in a significantly simplified form (Kosmrlj et al. 2008; Kosmrlj et al. 2009; Kosmrlj et al. 2010) . The TCR and MHC plus peptide (subsequently referred to as MHC-peptide) complexes are each modeled numerically as strings of amino acids. The strings (typically both 10 amino acids in length) represent the amino acids at the interface between the TCR and MHC-peptide complex and the model assumes that only one site on the TCR interacts with a corresponding site which consists of the bound peptide. Both the segment of the TCR and the segment of the peptide are modeled as being of equal sequence length. The TCR segment represents the highly variable CDR3 and interactions between the TCR and MHC are not explicitly considered.

Subsequently, *in silico* thymic selection experiments were carried out where an HLA-dependent number of thymic self-peptides where chosen, according to their frequency of appearance in the human proteome. Similarly, immature thymocytes (CD8 T cells) were generated by choosing amino acids which contacted the peptide according to the probabilities with which amino acids appear in the human proteome. A mature T cell emerged from the thymus only when its TCR had bound to at least one self-peptide with an affinity that exceeded the positive selection threshold, but did not exceed the negative selection threshold. After emergence from the thymus, the mature T cell was challenged by a viral peptide bound to the same MHC type (which, as already noted, was not explicitly considered). If the interaction strength exceeded that of the negative selection threshold the TCR was considered to successfully recognize the peptide. Additionally, TCR cross-reactivity was assessed by mutating each TCR contact residue to the other 19 amino acid types. Sites on the viral peptide were considered to be important, if half or more of the mutations caused the recognition by the TCR to be abrogated. The authors concluded that negative selection skews the mature T cell repertoire to TCRs composed of peptide contact residues enriched in amino acids that bound weakly to amino acids in the peptide.

The authors (Kosmrlj et al. 2008) also qualitatively compared their *in silico* results with those of 18 unspecified crystallized TCR-MHC-peptide complexes from the PDB: the *in silico* predictions were assessed for their strength of interaction with other amino acids and compared to the strength of such interactions in solved crystal structures, which were divided into two classes (weak and strong). The authors assert that data obtained from the crystal structures was in qualitative agreement with the theoretical prediction that weakly interacting amino acids on the TCR are enriched, whereas strongly interacting amino acids are attenuated. Overall, it can be said that the current *in silico* approaches predicting which TCR residues are likely to interact with the MHC-peptide antigen are still highly simplified, but the models do give important insights into the population dynamics of the T cell repertoire and its interaction with viruses such as HIV-1 (Kosmrlj et al. 2010).

1.5 Host-Virus Dynamics: The Interplay of Host Genetic Factors and Viruses

The most common recurring infections in humans are due to viruses. The major innate response to viruses is by type 1 interferons, complement and natural killer (NK) cells. The major adaptive protection against viruses is via antibodies and cytotoxic T cells. CTLs are able to recognize infected host cells because these cells present viral protein fragments in the form of peptides bound to HLA class I molecules on their cell surface. Once the HLA/peptide complex has been recognized by the CTL's T cell receptor, the CTL will release cytokines, chemokines and molecules such as perforin and granzyme that result in the lysis and apoptosis of the infected cell (DeFranco et al. 2007).

Host genetic profiles play an important role in the course and outcome of viral diseases. These genetic profiles have a significant impact on the susceptibility or resistance to infection, the rate of disease progression and therefore the clinical manifestations of the disease. However, the rapid rate of viral evolution in viruses such as HIV-1 and HCV allows the virus to adapt to specific host genetic profiles. The increased understanding of the interplay between the host and the virus has provided important insights into the large variation in responses to viral infection, where some patients are resistant to infection whereas others progress to more advanced forms of disease in a very short time period. Genetic resistance is an important aspect of the viral adaptation to the host and its analysis is

an important aspect of the treatment of viral disease. The HLA genotype of an individual appears to play an important role in the progression of the disease (Beerenwinkel et al. 2007).

When regions of viral proteins are presented as epitopes by HLA molecules, an immune response may be triggered. Therefore, the less likely it is for a particular peptide to be presented by an HLA molecule, the lower the overall immune response of the patient against that particular peptide will be. In certain HLA genotypes, a mutation away from the viral consensus sequence might be beneficial in promoting immune escape. The ability of the viruses such as HIV-1 and HCV to mutate rapidly enables them not only to develop resistance mutations to anti-retroviral therapy, but also to adapt to the HLA genotype of the infected individual. This type of mutation is frequently termed an escape mutation (Perelson et al. 1996; Beerenwinkel et al. 2007; Bostan and Mahmood 2010).

Once such an escape mutation has occurred it has three potential evolutionary fates. It may revert to the wild-type amino acid on transmission to an individual of another HLA genotype. In this case, the wild-type amino acid will remain a target for the immune system. Alternatively, the escape mutation might be stable and the epitope that contains it will no longer be a target for the immune system. Such an escape mutation will reach fixation in a population over time as all other amino acid variants at that position are eliminated leaving only the escape mutation. The original epitope therefore also no longer exists and has become extinct. Finally, the escape mutation might not revert to wild-type, but could still be contained in a recognized epitope. In this case, the frequency of the mutation will equilibrate in the population and thus continue co-exist with the wild-type amino acid (Leslie et al. 2005; Beerenwinkel et al. 2007).

A number of large, population-based HLA class I association studies have been conducted with HIV-1. Studies were performed on cohorts of differing ethnicity and focusing on differing HIV-1 subtypes. They have associated some HLA molecules with better responses to viral infection (e.g. HLA-B*27) and others to faster progression of disease (e.g. HLA-B*3502). Several smaller studies have been performed analyzing the effects of HLA class II genes on disease progression. Also, numerous distinct viral epitopes have been identified which contain viral escape mutations (Kaur and Mehra 2009; Kawashima et al. 2009; Ferre et al. 2010a; Shankarkumar et al. 2010).

An analysis of the percentages of viral amino-acid positions, at which polymorphisms away from population consensus were significantly associated with the HLA-A, -B and -C allelic groups, found that there is great variability across the HIV-1 genome. Some viral genes have no statistically significant associations, whereas others have over 2% of their amino acids involved in significant associations (Kiepiela et al. 2004). Other work has shown that defined CTL epitopes tend to cluster in regions with distinct characteristics: conserved regions appear to have more epitopes, highly variable regions that lack epitopes bear cumulative evidence of past immune escape that may make them relatively refractive to CTL cells, and epitopes are more highly concentrated in α-helical regions of HIV-1 proteins (Yusim et al. 2002).

The association between immune control of HIV-1 and the HLA genotype of an individual is particularly striking for alleles of the HLA-B locus. The alleles HLA-B*27, HLA-B*51 and HLA-B*57 have been associated with the control of infection, the alleles HLA-B*58 and some HLA-B*35 alleles lead to more rapid disease progression. The HLA-B locus is not only the most polymorphic of the three major HLA class I loci, but also the most polymorphic in the human genome. Work showing the critical factor linking the protective HLA-B alleles is that they present a relatively large number of HIV-1 Gag protein epitopes, whereas the HLA-B alleles that lead to a more rapid progression of disease present few to no Gag protein epitopes. Further population-based studies have shown that an increased breadth of Gag-responses in patients is correlated with decreased viral load and therefore effective

control of infection, and such a correlation has not been seen with other non-Gag-responses. It is thought that there are several reasons why Gag protein epitopes are important for immune control of HIV-1. Firstly, the Gag protein displays very little inter-individual sequence variation when compared to other HIV-1 proteins, which implies that sequence changes in this gene are not well tolerated because they lead to significant reduction in fitness cost for the virus. And secondly, Gag is considered to be highly immunogenic probably because of its high abundance in cells infected by the virus (Matthews et al. 2008; Webster 2009).

A recent comparison of elite HIV-1 controllers (viral load of fewer than 75 copies/ml), viremic controllers (viral load of 75 to 2000 copies/ml) and non-controllers (viral load of more than 10,000 copies/ml) showed that HIV-specific CD8 T-cell responses are common to both blood and mucosa, but that Gag-specific responses dominate in the rectal mucosa of controllers. Controllers had responses to both HLA-B*27- and HLA-B*57-restricted epitopes in both tissues and the magnitude of the response, measured by testing the spot-forming cells (SFC) per million, was greater than those epitopes recognized by other alleles. In contrast, the responses of non-controllers were more evenly distributed among the HIV-1 genes examined (Gag, Env and Nef) (Ferre et al. 2010b).

An important new study has examined host genetic effects on the outcome of chronic HIV-1 infection. The authors performed a genome-wide association study comparing HIV-1 controllers and progressors of multiple ethnic groups. Controllers were defined as having had their plasma viral load measured at least three times and the resulting levels measured fewer than 2000 RNA copies/ml over at least a one year in the absence of antiviral therapy (median viral load 241 copies/ml, CD4 count 699 cells/mm^3, disease duration 10 years). Such individuals also typically maintained stable CD4 cell counts, did not develop clinical disease and were less likely to transmit disease to others. Progressors were chronically infected treatment-naïve individuals with a median viral load of 61,698 copies/ml and CD4 count 224 cells/ mm^3. In the largest ethnic group, which consisted of individuals of European ancestry, 313 single nucleotide polymorphisms of genome-wide significance were identified (i.e. the *P*-values for these SNPs remained significant after correction for multiple testing and applying other experimental quality control measures). Every one of these SNPs was located in the MHC region of chromosome 6. The African American and Hispanic ethnic groups had similar results. Specific examination of the CCR5-CCR2 locus, which had been associated with HIV-1 disease progression in previous studies, only showed nominal (uncorrected) statistical significance in this new analysis. This work underscores that HLA-viral peptide interaction is the major factor modulating durable control of HIV-1 infection (Pereyra et al. 2010).

It is well established that MHC class II alleles bind epitopes promiscuously (Panina-Bordignon et al. 1989; O'Sullivan et al. 1991). More recent work, systematically examining a broad selection of HIV-1 and EBV epitopes known to bind to MHC class I alleles, was also able to show that there is significant promiscuity for this allele class (Frahm et al. 2007). This confirmed work which examined small numbers of epitopes that had been evaluated in the context of HLA supertypes i.e. the identified epitopes bound to differing alleles which belonged to the same supertype (Threlkeld et al. 1997; Burrows et al. 2003; Masemola et al. 2004), but also showed that certain epitopes are capable of binding diverse alleles which do not belong to the same supertype. Of the 242 well-defined viral epitopes tested in 100 individuals, half of all detected responses were in the absence of the originally described allele. In fact, only three percent of epitopes were exclusively recognized in the presence of the original allele.

In comparison to HIV-1, relatively little work has been invested analyzing the impact of host genetic profiles on HCV at a population level. In the studies so far (HLA class I and

class II), conflicting results have been obtained which are probably due to the small sample size of some of the cohorts studied. Recent work in a large multi-racial cohort in the United States showed that certain HLA alleles are associated with increased clearance (DRB1*0101, HLA-B*5701, HLA-B*5703, and HLA-Cw*0102) and another allele was associated with HCV RNA positivity (HLA-DRB1*0301) (Kuniholm et al. 2010).

TCR cross-reactivity has interesting implications for the ability of the immune system to fight infections. Experiments (Huseby et al. 2005; Huseby et al. 2006) and simulation studies (Kosmrlj et al. 2008; Kosmrlj et al. 2009) have shown that T cells that develop in mice only expressing one type of MHC-bound peptide in the thymus are more cross-reactive to point mutants of peptide epitopes, when compared with the T cells of mice which express diverse self peptides. Additionally, predictions have been made helping to explain the differences in the ability of different MHC molecules to bind to peptides (Kosmrlj et al. 2010): of the $\sim 10^7$ peptides in the human proteome which can be bound by various MHC molecules, only 70,000 peptides were predicted to bind to HLA-B*5701 and 180,000 peptides were predicted to bind to HLA-B*0701. Therefore, humans expressing HLA alleles such as HLA*B-5701 that bind fewer peptides are expected to have more cross-reactive TCRs and therefore a better immune response. This is indeed the case for HIV-1, where the HLA-B*5701 allele has been found to be enriched in elite controllers, i.e. patients who maintain very low levels of HIV-1 RNA without therapy, while HLA-B*0701 has not been found be not protective against HIV-1 (Kosmrlj et al. 2010).

Due to the critically important influence of host genetic profiles on the level of viral control achieved by a patient, future treatment strategies for patients infected with such viruses as HIV-1 and HCV should not only involve the analysis of drug resistance mutations in order to help define treatment options, but to also take into account patient HLA profiles as host genetic effects can also play a causal role in the loss of immune control.

2 Selective Pressures of HLA Genotypes and Antiviral Therapy on HIV-1 Sequence Mutation at a Population Level[1]

2.1 Introduction

Human immunodeficiency virus infection has become a major global human health issue and a major challenge to natural or vaccine-induced immune control of HIV-1 is the ability of the virus to mutate rapidly when it comes under the pressure of the host's immune system (Koenig et al. 1995; Borrow et al. 1997; Goulder et al. 1997; Barouch et al. 2002). Antiviral cytotoxic T lymphocytes kill HIV-1 infected cells upon recognition of specific viral epitopes. HIV-1 escape mutations interfere with processing of viral antigens by proteasomes (Allen et al. 2004; Yokomaku et al. 2004; Zimbwa et al. 2007) or evolve at critical binding sites within the human leukocyte antigen restricted CTL epitope, thereby abrogating binding to the HLA molecule or inhibiting efficient recognition by the T cell receptor (McMichael and Rowland-Jones 2001; Draenert et al. 2004). Thus, HIV-1 escapes antiviral immune responses and eradication by the host's immune system. Such selection pressure as well as viral adaptation to antiretroviral drugs should lead to consistent changes in the amino acid sequence of the proteins expressed by dominant subpopulation of the viral quasispecies.

Moore et al. studied the selection pressure exerted by HLA restricted immune responses on the evolution of the HIV-1 sequence at the population level (Moore et al. 2002). A cohort of 473 HIV-1 infected patients was genotyped for the HLA-A and HLA-B loci. The most recent sequence of the HIV-1 reverse transcriptase between amino acid positions 20 and 227 was aligned to an HIV-1 consensus sequence and viral mutations were identified. These mutations were then tested for association with distinct HLA-A or -B alleles. The authors identified 64 positive and 17 negative associations, although only 12 remained after correction for multiple testing. Several of these mutations were located in known CTL epitopes. In a second study of the same cohort, the aforementioned group identified interactions between antiretroviral drugs, HLA alleles and diversity in the RT and protease viral sequences (John et al. 2005). These interactions led to higher frequencies of antiviral drug resistance mutations in patients with certain HLA alleles in some cases, but also to lower frequencies in other cases. This indicates that HLA dependent specific immune responses can support, but also prevent the evolution of drug resistance.

The previous studies have analyzed the HLA-driven evolution of HIV-1 in only a fragment of the RT and protease. Therefore we wanted to examine if this phenomenon can be confirmed in the entire first half of the RT.

We were also interested in extending the analysis to include the MHC class II locus HLA-DRB1 to better understand selection pressure by CD4 T helper cells at the population level. MHC class II molecules are specialized to monitor the endosomal-lysosomal system of cells which in turn are specialized to internalize extracellular antigen (i.e. phagocytes, dendritic cells and B cells). MHC class II molecules display peptides derived from these antigens to CD4 T cells and may thus produce a different form of immune pressure against HIV-1 infection (DeFranco et al. 2007).

[1] The work reported in this section was performed in collaboration with Golo Ahlenstiel and Ulrich Spengler (University of Bonn, Germany) and Martin Däumer and Rolf Kaiser (University of Cologne, Germany). It has been published in the journal *Clinical and Vaccine Immunology*, appearing in this thesis with the journal's permission (Ahlenstiel et al. 2007). My own contribution to the publication comprises the construction of the HLA-HIV database, identification of escape mutations and their analysis and the phylogenetic analysis resulting 4 of the 4 published tables (Tables 2.1, 2.2, 2.3 and 2.4) and 1 of 1 published figures (Figure 2.3).

In order to minimize the influence of founder effects on the HLA associations found (Bhattacharya et al. 2007), we the analysis to only those patients infected by clade B HIV-1 viruses. Previously, viral lineage effects in studies encompassing more than just one HIV-1 clade led to the identification of escape mutations which were in reality related to the viral clade and thus not true escape mutations. Also, we performed an analysis of potential viral linage effects within the cohort.

Furthermore, we wanted to assess for "hot spots," where the sequence mutates more easily/rapidly due to immune pressure and how mutations persist over time.

Finally, to understand the clinical significance of our findings we analyzed whether HLA-driven mutations in the RT and/or protease sequence of HIV-1 lead to antiviral drug resistance and if the patient's HLA-type has an impact on whether drug resistance mutations are accumulated in a specific order in the case of thymidine analogue mutations.

2.2 Patients and Methods

2.2.1 Patients

We studied 179 clade B HIV-1 positive patients being treated at a single hospital in Bonn, Germany. The patients were monitored every three months and any complications were recorded and classified according to the European modification of the 1986 Centers for Disease Control and Prevention staging (CDC 1986). Antiviral therapy was advised according to updated recommendations on antiretroviral treatment for HIV-1 infection by the International AIDS Society panel. HIV-1 sequence data was collected between March 1999 and May 2003, with some patients having sequences collected at more than one time point. Additional information on HIV-1 transmission, ethnicity, sex and HBV and HCV infection status was also acquired (Table 2.1).

Description		Number (%)
Total number of patients		179
Age	Average	42
	Range	22 - 70
Female / Male		15 / 164
Ethnicity	Caucasian	161 (90.0%)
	Other	13 (7.3%)
	Unknown	5 (2.7%)
Risk Code	Homosexual	99 (55.3%)
	Heterosexual	12 (6.7%)
	IV drug abuse	12 (6.7%)
	Residency in epidemic region	3 (1.7%)
	Exposure to blood products	1 (0.5%)
	Haemophelia	40 (22.3%)
	Unknown	12 (6.7%)
HIV-1 Viral Subtype	B	179 (100%)
HCV Status (Chronic)		55 (30.7%)
HCV Virus Type	1	29 (16.2%)
	2	8 (4.5%)
	3	10 (5.6%)
	4	2 (1.1%)
	Multiple	1 (0.5%)
	Unknown	5 (2.7%)
HBV	HBs antigen +	15 (8.4%)
	Anti-HBs-, Anti-HBc+	23 (12.8%)
	Anti-HBs+, Anti-HBc+	45 (25.1%)
Anti-HIV Therapy	Naive	14 (7.8%)
	ART (< 3 antiviral drugs)	18 (10.1%)
	HAART (>= 3 antiviral drugs)	147 (82.1%)

Table 2.1 Characteristics of the patient cohort

For the laboratory methods used for HLA genotyping and HIV-1 sequences, see Appendix A.

2.2.2 Database design

An extensive and diverse dataset was accumulated for the patient cohort, which included personal patient data, therapy histories, virologic and immunologic data, further clinical test data derived from patient tissues, and sequence data. The HIV-1 sequences are clinically derived and thus sequenced with population-based or bulk approaches. The initial database schema was based on that of the HIV-1 Arevir database (Roomp et al. 2006), in which the data is captured in different modules consisting of a few tables each. These Arevir modules include, for example, "Patients" which contains personal patient data, treatment location and diagnoses.

The schema of the Arevir database was significantly expanded to include data on the HLA-typing of patients, HLA genes, alleles, supertypes, haplotypes, motifs, known epitopes, etc. Additionally, tables were added to reflect the complex medical histories of patients which were often co-infected with viruses such as HCV and HBV. Therefore, a data model was

developed which is specifically tailored to analyze host immune responses to viruses (Figure 2.1).

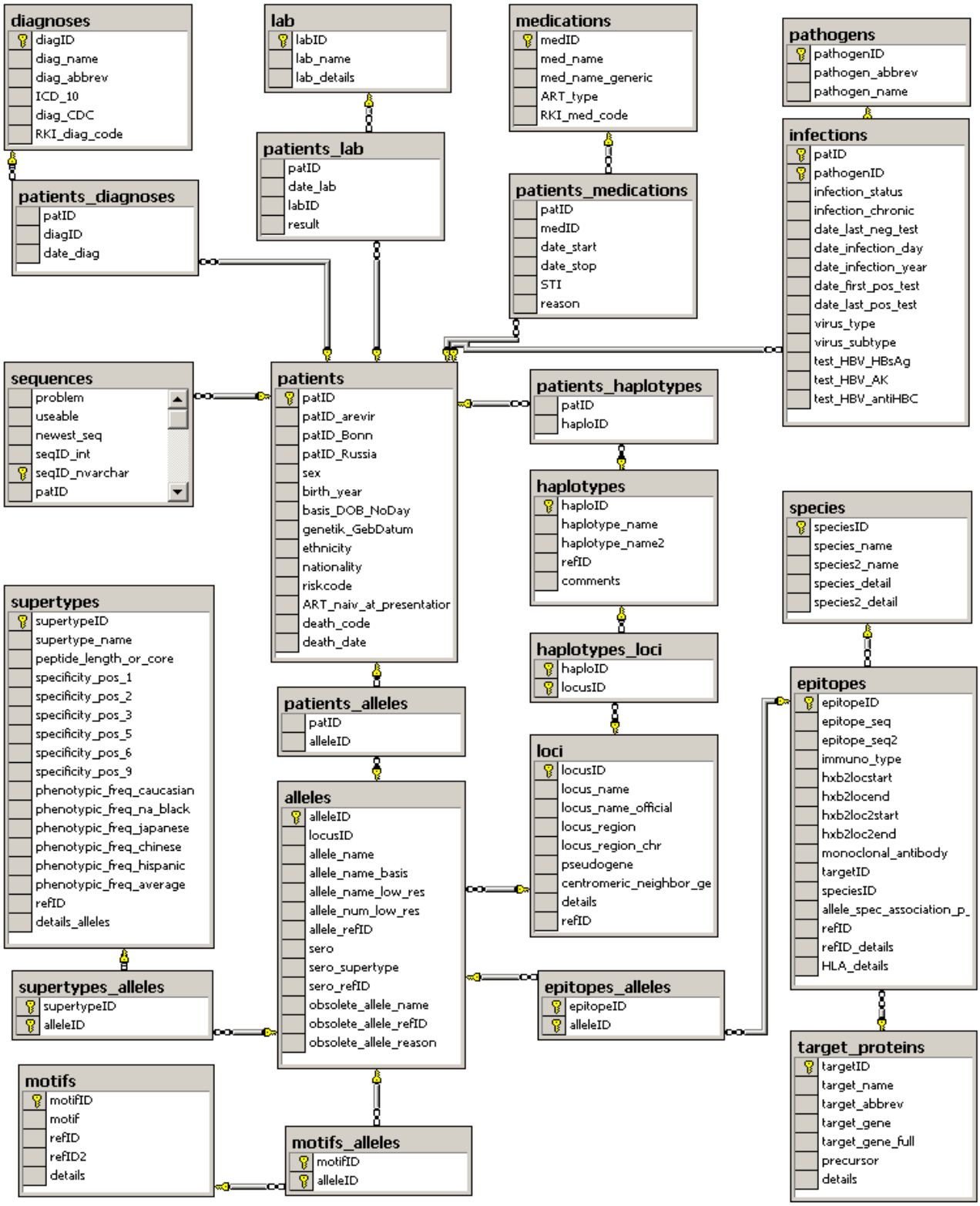

Figure 2.1 Database diagram of the HLA-HIV-1 database

Dependencies between important database tables are shown, as well as all primary and foreign keys.

The data model was implemented in the relational database management system (RDBMS) Microsoft SQL Server. It was chosen because rich features in manipulating, securing and managing data and ease of use.

2.2.3 HLA-associated mutations in the reverse transcriptase and protease

All amino acids in the complete HIV-1 protease and between positions 1 and 330 in the viral reverse transcriptase were examined in the most recent sequences from all patients and compared to population consensus sequences (see Appendix B). As a first step, we analyzed each amino acid position using Fisher's exact test for associations with HLA alleles (Mehta and Patel 1986). In order to raise the power of the calculations, very rare alleles (less than or equal to 4% of the cohort) were excluded from the analysis. All HLA allele covariates with *P*-values of less than 0.05 were identified and fitted in a subsequent multivariate analysis to logistic regression models. We used binomial models, where the linear predictor consisted of all significant alleles and the response was a factor, in which variant amino acids were classified as successes. Correction for multiple-testing used the false discovery rate (FDR) method (Benjamini and Hochberg 1995). Due to the low numbers of deletions and insertions in the sequences, no separate analysis was done to take these into account.

2.2.4 Distribution and persistence of HLA-associated mutations

A previous study has described CTL-epitope "hotspots," in which immunodominant epitopes cluster within distinct regions of the HIV-1 gp120 protein (Brown et al. 2005). The CTL epitopes collected in the HIV Molecular Immunology Database at Los Alamos also appeared to be localized in particular regions of HIV-1 proteins (2005). We were interested in determining, if the distribution of HLA-associated mutations was uniform across the RT and protease. Therefore, the distribution of known mutations was compared to 10,000 randomly generated distributions according to the uniform model, each with an equal number of sequence mutations and tested for statistical significance.

We also analyzed the persistence of HLA-associated mutations over time in our cohort. For 70 patients, HIV-1 sequences were available for at least two individual time points that were a minimum of 6 months apart (the sequences of 50 patients were collected more than 1 year apart). None of these sequences was acquired during the acute stage of HIV-1 infection, but rather at a later time point. Persistence of mutations was examined at each of the already identified positions.

2.2.5 Impact of HLA on drug resistance mutation pathways

During antiretroviral therapy the virus is exposed to strong selective pressure, which can result in an accumulation of mutations conferring drug resistance. These mutations are usually persistent, provided that there is continuous drug-induced selection pressure (Shafer et al. 2000). Therefore, the development of resistance within the HIV-1 genome can be regarded as the accumulation of such mutations. This accumulation has been modeled by weighted branchings or directed trees, which provide an intuitive model of directed dependencies between events and their time of occurrence. The single tree model has been extended to mixtures of trees (so-called mutagenetic trees mixture models) in order to capture more complex evolutionary scenarios, for which the software package *Mtreemix* has been developed (Beerenwinkel et al. 2005).

Of particular interest were the thymidine analogue mutations (TAMs) in the RT that can arise after treatment with nucleoside reverse transcriptase inhibitors (NRTIs) zidovudine,

stavudine and abacavir. Studies have suggested that HIV-1 develops TAMs by one of two distinct pathways TAM1 (41L, 210W and 215F/Y) or TAM2 (67N, 70R and 219E/Q) (Hanna et al. 2000).

We identified all patients in our cohort that had undergone NRTI treatment and examined their HIV-1 sequences for TAM mutations to determine, whether particular HLA alleles can be associated with either the TAM1 or TAM2 pathway.

2.2.6 HLA-driven selection at antiretroviral drug resistance sites

Several HLA allele-specific mutations have been reported to be located in positions of known drug resistance mutations (John et al. 2005; Johnson et al. 2005). An analysis of the sequences of our patient cohort was performed in order to determine whether similar associations could be identified.

2.2.7 Phylogenetic analysis

In order to explore the impact of viral lineage founder effects, we have applied *ProtTest* version 1.3 (Drummond and Strimmer 2001; Guindon and Gascuel 2003; Abascal et al. 2005) to find the best-fitting model of protein evolution for the HIV-1 sequence alignment of the cohort (the alignment was made using the bulk/consensus sequences generated for each patient). The best-fitting model according to both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Hastie et al. 2001) was the JTT model (Jones et al. 1992) with gamma rates, variable amino-acid frequencies and invariable sites. We then used *PhyML* (Guindon and Gascuel 2003) for estimating a maximum-likelihood phylogeny for the given protein evolution model and performed 100 bootstrap replicates in order to obtain bootstrap support values. Subsequently, we used neighbor-net (Bryant and Moulton 2004; Huson and Bryant 2006) to get a better visualization of the noise in the phylogenetic signal.

2.3 Results

2.3.1 Patients

During the observation period, 147 (82.1%) patients were treated with highly active anti-retroviral therapy (HAART), which was based on a protease inhibitor (PI; indinavir, saquinavir, nelfinavir, amprenavir or ritonavir) in 117 patients and a non-nucleoside reverse-transcriptase inhibitor (NNRTI; nevirapine, efavrienz, or delavirdine) in 30 patients. HAART was the first-line treatment in 61 patients, and 86 had previously had therapy with nucleoside analogues. HAART was changed to different drug combinations in 85 of 117 patients who received HAART based on PIs and in 23 of 30 patients who received HAART based on NNRTIs, to address side-effects or emerging viral resistance.

The antiretroviral-treatment group consisted of 18 (10.1%) patients who received 2 or fewer antiretroviral drugs. One patient received zidovudine monotherapy and 17 were treated with several nucleoside analogues (zidovudine, didanosine, zalcitabine, stavudien or laminvudine). The remaining 14 patients (7.8%) did not receive any retroviral drugs.

With regard to the length of treatment, 18 patients had been treated for less than 1 year, 62 patients had been treated between 1 and 5 years, and 88 patients for more than5 years. The staging of HIV-1 disease according to the European modification of the 1986 Centers for Disease Control and Prevention staging (CDC 1986) in the cohort was A in 58 patients, B in 59 patients and C in 57 patients. Five patients were not classified.

2.3.2 HLA-associated mutations in the reverse transcriptase and protease

HIV-1 RT sequences (amino acid positions 1 to 330) were initially aligned with the reference sequence HXB2 (2005). Associations were subsequently confirmed using the population consensus sequence, which was generated by assigning the most common amino acid for each position of all sequences pooled from the cohort. Overall, 15 associations with uncorrected *P*-values of less than 0.005 were found, which also had *P*-values of less than 0.05 in the logistic regression models (Table 2.2). Ten of these associations had *P*-values of less than 0.05 after correction for multiple-testing.

We were able to confirm four of the associations that Moore et al. reported as significant after correction for the total number of residues examined across the entire region (Moore et al. 2002): HLA-B*51 at amino acid position 135, HLA-B*07 at position 162, HLA-A*11 at position 166 and HLA-B*35 at position 177. We also found a number of novel HLA allele-specific associations with 4 of those associations lying outside the part of the RT sequence analyzed by Moore et al. Seven of the mutations in the RT sequence are located in previously defined epitopes with six of these being located in known epitope anchor positions.

Interestingly, six of the identified associations were negative associations indicating that mutations in the RT were less likely, if the patient carried that particular allele: amino acid position 177 was negatively associated with HLA-B*35, position 178 with HLA-B*35, position 188 with HLA-DRB1*12, position 207 with HLA-B*15, position 277 with HLA-A*03 and position 291 with HLA-B*27. A possible explanation for such negative associations was given by Leslie et al: a negative association can arise as a result of positive selection of an escape mutation by high frequency alleles, which is stably transmitted, accumulating in the population to the point at which it defines the consensus sequence (Leslie et al. 2005).

In the protease, we found HLA-associated mutations at seven positions with an uncorrected *P*-value of less than 0.005, which also had *P*-values of less than 0.05 in the logistic regression models (Table 2.3). Four associations had *P*-values of less than 0.05 after correction for multiple-testing, with three of these associations being previously defined epitopes. No negative associations were found.

For the two HLA-DRB1 associated mutations in the RT (position 67 for HLA-DRB1*08 and position 188 form HLA-DRB1*12), the alleles are not known to be in linkage disequilibrium with any class I alleles which were included in the multivariate analysis at each position.

Overall, HLA-B alleles were involved in more associations (n = 15 or 68%) than alleles from either the HLA-A (n = 5 or 23%) or HLA-DRB1 (n =2 or 9%).

Polymorphism	HLA type	Absent (n)	Present (n)	OR	P-value (unadjusted)	P-value (adjusted, FDR)	Known Epitopes	Persistent?	New?
1 E6x	HLA-B*40	23	11	11.08	0.00001	0.00047	B*4001 aa5-aa12: IETVPVKL	persistent: 9 multiple pops: 3 definite change: 0	y
	Non-HLA-B*40	139	6						
2 V35x	HLA-A*24	19	17	3.23	0.00314	>0.05		persistent: 6 multiple pops: 1 definite change: 1	y
	Non-HLA-A*24	112	31						
3 D67x	HLA-DRB1*08	1	6	13.85	0.00501	>0.05		persistent: 3 multiple pops: 2 definite change: 0	y
	Non-HLA-DRB1*08	120	52						
4 V75x	HLA-B*57	6	5	9.17	0.00244	>0.05		persistent: 4 multiple pops: 2 definite change: 1	y
	Non-HLA-B*57	154	14						
5 I135x	HLA-B*51	2	22	18.40	<0.00001	0.00002	B*5101/B51 aa128-aa135:TAFTIPSI	persistent: 8 multiple pops: 0 definite change: 1	n
	Non-HLA-B*51	97	58						
6 E138x	HLA-A*24	31	5	-	0.00026	0.00859		persistent: 7 multiple pops: 0 definite change: 1	y
	Non-HLA-A*24	143	0						
7 S162x	HLA-B*07	21	27	6.04	<0.00001	0.00004	B7 aa153-aa165: WKGPAIFQSSMT aa153-aa165: WKGSPAIFQSSMT aa156-aa164: SPAIFQSSM aa156-aa165: SPAIFQSSMT	persistent: 14 multiple pops: 1 definite change: 3	n
	Non-HLA-B*07	108	23						
8 K166x	HLA-A*11	13	7	11.69	0.00016	0.00521	A*1101/A11 aa158-aa166: AIFQSSMTK aa158-aa166: SIFQSSMTK	persistent: 5 multiple pops: 0 definite change: 0	n
	Non-HLA-A*11	152	7						
9 D177x	HLA-B*35	16	18	0.14	<0.00001	0.00013	B*3501/B35 aa175-aa183: NPDIVIYQY aa175-aa183: HPDIVIYQY	persistent: 10 multiple pops: 0 definite change: 1	n
	Non-HLA-B*35	125	20						
10 I178x	HLA-B*35	22	12	0.23	0.00125	0.04121	B*3501/B35 aa175-aa183: NPDIVIYQY aa175-aa183: HPDIVIYQY	persistent: 10 multiple pops: 0 definite change: 1	y
	Non-HLA-B*35	129	16						
11 Y188x	HLA-DRB1*12	5	3	0.04	0.00191	>0.05		persistent: 3 multiple pops: 0 definite change: 1	y
	Non-HLA-DRB1*12	167	4						
12 Q207x	HLA-B*15	10	20	0.17	0.00003	0.00090		persistent: 8 multiple pops: 1 definite change: 2	y
	Non-HLA-B*15	111	38						
13 K277x	HLA-A*03	1	38	0.02	<<0.00001	<<0.00001	A3 aa269-aa277: QIYPGIKVR	persistent: 14 multiple pops: 1 definite change: 1	y
	Non-HLA-A*03	89	51						
14 K277x	HLA-B*44	26	9	3.61	0.00225	0.02616		persistent: 12 multiple pops: 2 definite change: 1	y
	Non-HLA-B*44	64	80						
15 E291x	HLA-B*27	10	3	0.06	0.00534	>0.05		persistent: 4 multiple pops: 0 definite change: 0	y
	Non-HLA-B*27	163	3						

Table 2.2 HIV-1 sequence mutations in the HIV-1 reverse transcriptase.

Polymorphism shows the variant position with the wild type amino acid; *HLA type* the specific HLA type for which a mutation was found; *Absent (n)* the number of patients with the wild type amino acid; *Present (n)* the number of patients with the HIV-1 sequence mutation; OR the odds ratio; *P-value (unadjusted)* the unadjusted *P*-value of the contingency table using Fisher's Exact Test; *P-value (adjusted)* the adjusted *P*-value using FDR; *Known Epitopes* any known epitopes for the HLA-allele and their sequences (2005); *Persistent?* the persistence of the HIV-1 sequence mutation in individual patients at a minimum of two different time points (*persistent* = all dominant viral sequences from a patient were identical at the different time points, *multiple pops* = dominant viral sequences identified in a patient were different at the various time-points without a consistent change from one type to another, *definite change* = the patient's dominant viral sequence showed a clear switch at this position from one amino acid to another amino acid); *New?* indicates whether the HIV-1 sequence mutation is new or has already been described in the literature.

2.3.3 Distribution and persistence of HLA-associated mutations

The HLA-associated mutations that have been identified in our patient cohort do not appear to be uniformly distributed across the protease and RT, but are more frequent in regions known to have many epitopes. However, an analysis of the non-random distribution of the HIV-1 sequence mutations showed only a weak level of significance in the protease (*P*-value equal to 0.06) and none in the RT (data not shown).

In most patients, sequence mutations were consistently found in all available sequences. In relatively few cases, the mutations were not persistent with the wild type amino acid being substituted by a variant amino acid in most cases (Tables 2.2 and 2.3). Drug therapy did not appear to influence the gain or loss of HIV-1 sequence mutations, as most patients had HAART composed of diverse drug combinations, and yet their mutations have remained stable throughout.

Polymorphism	HLA type	Absent (n)	Present (n)	OR	P-value (unadjusted)	P-value (adjusted, FDR)	Known Epitopes	Persistent?	New?
1 E35x	HLA-B*44	11	24	10.91	<0.00001	<0.00001	B*44 aa34-aa42: EEMNLPGRW	persistent: 14 multiple pops: 1 definite change: 0	y
	Non-HLA-B*44	120	24						
2 E35x	HLA-A*68	11	13	4.05	0.00238	0.03922	A*6802 aa30-aa38: DTVLEDINL aa30-aa38: DTVLEEMNL aa30-aa38: DTVLEEWNL	persistent: 7 multiple pops: 0 definite change: 0	y
	Non-HLA-A*68	120	35						
3 N37x	HLA-B*44	9	26	6.15	<0.00001	0.00022	B*44 aa34-aa42: EEMNLPGRW	persistent: 11 multiple pops: 3 definite change: 1	y
	Non-HLA-B*44	98	46						
4 I54x	HLA-B*57	5	6	7.57	0.00285	>0.05		persistent: 4 multiple pops: 0 definite change: 3	y
	Non-HLA-B*57	145	23						
5 V82x	HLA-B*57	5	6	6.55	0.00499	>0.05		persistent: 5 multiple pops: 1 definite change: 1	y
	Non-HLA-B*57	142	26						
6 Q92x	HLA-B*15	22	8	17.70	0.00003	0.00099		persistent: 10 multiple pops: 1 definite change: 0	y
	Non-HLA-B*15	146	3						
7 I93x	HLA-B*15	11	19	3.63	0.00184	>0.05		persistent: 10 multiple pops: 0 definite change: 1	y
	Non-HLA-B*15	101	48						

Table 2.3 HIV-1 sequence mutations in the HIV-1 protease

Polymorphism shows the variant position with the wild type amino acid; *HLA type* the specific HLA type for which a mutation was found; *Absent (n)* the number of patients with the wild type amino acid; *Present (n)* the number of patients with the HIV-1 sequence mutation; *OR* the odds ratio; *P-value (unadjusted)* the unadjusted *P*-value of the contingency table using Fisher's Exact Test; *P-value (adjusted)* the adjusted *P*-value using FDR; *Known Epitopes* any known epitopes for the HLA-allele and their sequences (2005); *Persistent?* the persistence of the HIV-1 sequence mutation in individual patients at a minimum of two different time points (*persistent* = all dominant viral sequences from a patient were identical at the different time points, *multiple pops* = dominant viral sequences identified in a patient were different at the various time-points without a consistent change from one type to another, *definite change* = the patient's dominant viral sequence showed a clear switch at this position from one amino acid to another amino acid); *New?* indicates whether the HIV-1 sequence mutation is new or has already been described in the literature.

2.3.4 Impact of HLA on drug resistance mutation pathways

Of our patient cohort, 165 patients were treated with an NRTI regimen containing zidovudine, stavudine or abacavir before the sequence used for analysis was generated. Of these patients, 72 had no TAM mutations reported. For TAM1 (41L, 210W and 215F/Y), 9 patients matched the pathway, and 25 had either one mutation too few or too many. For TAM2 (67N, 70R and 219E/Q), 10 patients had matching mutations and 10 patients had either one mutation too many or one too few. The remaining patients had even rarer combinations of these mutations.

We looked for associations between either the TAM1 or TAM2 pathway and a particular HLA type using *Mtreemix*. This also allowed for imperfect matches to a particular pathway, because the software supports the analysis of complex evolutionary scenarios. The optimal number of trees was estimated to be k = 3 (Figure 2.2). However, no significant associations were found, i.e. patients belonging a particular tree could not be associated with also having particular HLA allele. We suspect that a large part of the difficulty of performing this analysis stems from the fact that such a large proportion of patients (72 of 165 treated with the NRTI regimen) had no TAM mutations whatsoever.

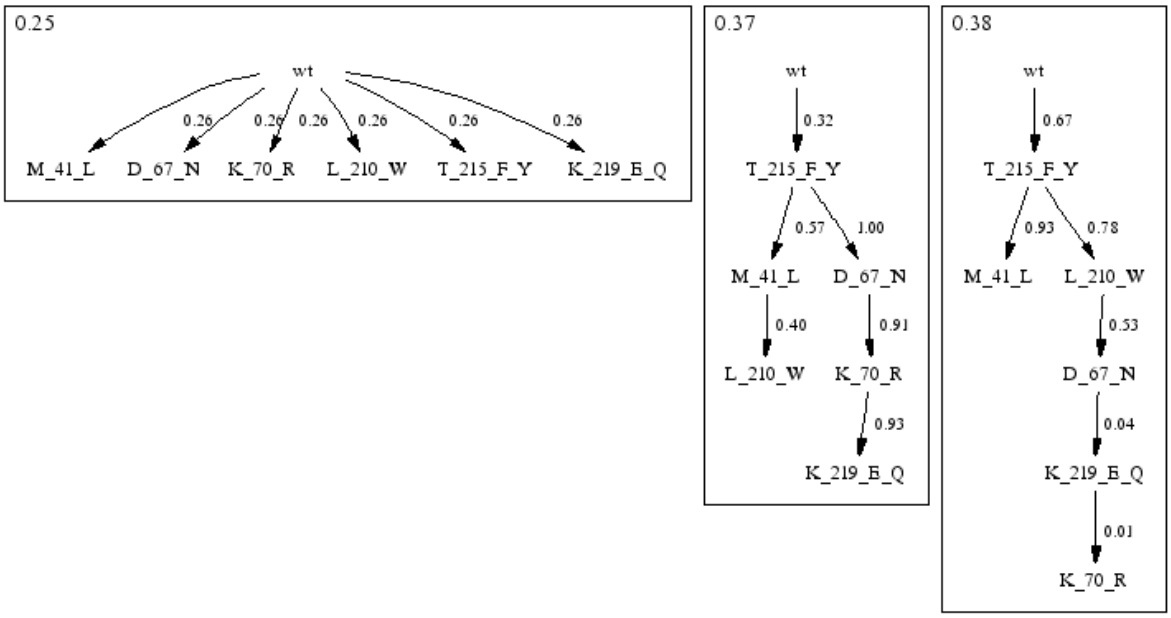

Figure 2.2 *Mtreemix* analysis, optimal number of trees

2.3.5 HLA-driven selection at antiretroviral drug resistance sites

We examined all associations in our cohort, which are also known drug resistance mutations (Table 2.4). Our results differ somewhat from those of John et al. (John et al. 2005), who reported drug resistance mutations at protease amino acid residues 20, 32, 36 and 48 to have positive HLA associations. In our cohort we could not confirm any of these associations. John et al. also reported the RT amino acid residues 41, 67, 70, 118, 210 and 215 to have positive HLA associations matching NRTI-associated mutations. Of these, only residue 67 was associated with HLA-DRB1*08 in our cohort (HLA-A*10 in John et al).

	Drug or Reistance Type	Polymorphism	HLA type	Absent (n)	Present (n)	Patients who received drugs	New?
1	Tipranavir/Ritonavir	PRO-E35x	HLA-B*44	11	24	0 of 24 (0%)	y
			Non-HLA-B*44	120	24	0 of 155 (0%)	
2	Tipranavir/Ritonavir	PRO-E35x	HLA-A*68	11	13	0 of 35 (0%)	y
			Non-HLA-A*68	120	35	0 of 144 (0%)	
3	Many currently used PIs	PRO-I54x	HLA-B*57	5	6	10 of 11 (90.9%)	y
			Non-HLA-B*57	145	23	117 of 168 (69.6%)	
4	Many currently used PIs	PRO-V82x	HLA-B*57	5	6	10 of 11 (90.9%)	y
			Non-HLA-B*57	142	26	117 of 168 (69.6%)	
5	Atazanavir	PRO-Q92x	HLA-B*15	22	8	0 of 30 (0%)	y
			Non-HLA-B*15	146	3	0 of 149 (0%)	
6	multi-nRTI Resistance	RT-D67x	HLA-DRB1*08	1	6	7 of 7 (100%)	y
			Non-HLA-DRB1*08	120	52	158 of 172 (91.8%)	
7	multi-nRTI Resistance	RT-V75x	HLA-B*57	6	5	11 of 11 (100%)	y
			Non-HLA-B*57	154	14	154 of 168 (91.6%)	
8	NNRTIs	RT-Y188x	HLA-DRB1*12	5	3	5 of 8 (62.5%)	y
			Non-HLA-DRB1*12	167	4	84 of 171 (49.1%)	

Table 2.4 HIV-1 sequence mutations at positions which are also known drug resistance mutations, and the number of patients of the HLA-type which received the drug type

Drug or Resistance Type shows the drug or drug resistance type; *Polymorphism* the HIV-1 sequence mutation in the protease or RT; *HLA type* the specific HLA type for which the mutation was found, *Absent (n)* the number of patients with the wild type amino acid; *Present (n)* the number of patients with the HIV-1 sequence mutation; *Patients who received drugs* the counts and percentages of patients who received the drug(s); *New?* indicates whether the association is new or was previously described in the literature.

We also found eight new associations, five in the protease and three in the RT. For three of the associations in the protease (residue 35 for HLA-A*68 and HLA-B*44, residue 92 for HLA-B*15), none of the patients in our cohort had ever been treated with the relevant antiviral therapy (tipranavir/ritonavir or atazanavir). In contrast, for the remaining five associations (residues 54 and 82 in the protease, residues 67, 75 and 188 in the RT), most patients were treated with the relevant antiviral therapy. These residues are associated with resistance against most currently available PIs, NNRTIs or multi-nucleoside and nucleotide reverse transcriptase inhibitors. However, statistically significant differences in patients taking these drugs could not be found between individuals with or without the HLA allele.

We observed that hepatitis C co-infected patients were less likely to have received antiretroviral therapy, which included PIs.

2.3.6 Phylogenetic analysis

The shape of the maximum-likelihood phylogeny of the viral sequences is almost star-like: the sequence diversity within clusters is higher than the diversity between clusters.

Another indicator for the star-likeness is the low bootstrap support for all branches in the interior of the phylogeny. The resulting neighbor-net network (Figure 2.3) shows a lot of netting in the center of the phylogeny and rejects a clustering into distinct groups. Hence, the neighbor-net method also suggests a star-like shape for the phylogeny. Given that the sequences evolved along a star-like phylogeny, we can rule out founder effects or other artifacts due to a shared evolutionary history.

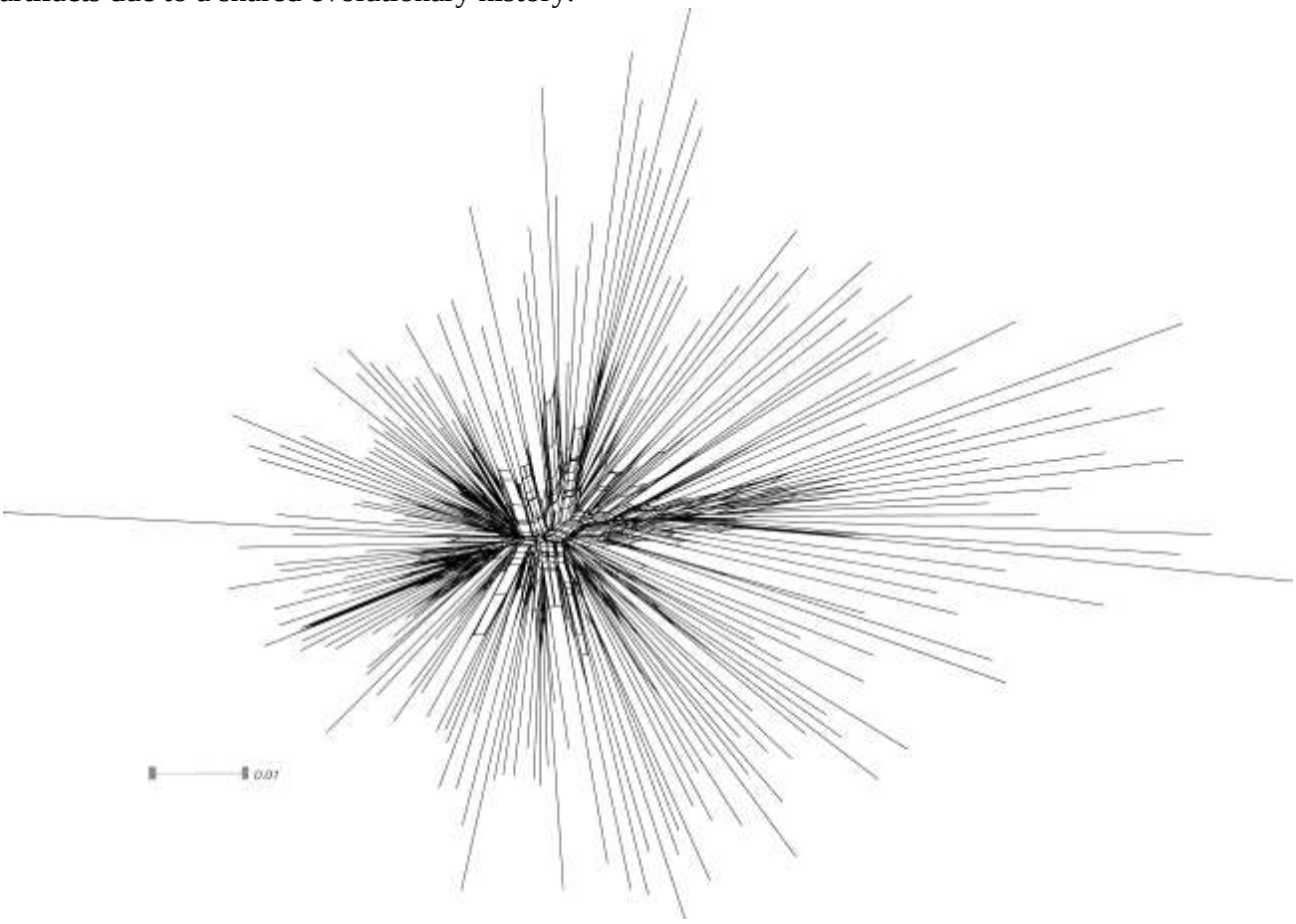

Figure 2.3 Neighbor-net network of the viral sequences drawn from the cohort

The network illustrates a star-like shape and a high level of netting in the center of the phylogeny.

2.4 Discussion

Selection of the dominant HIV-1 population among the viral quasispecies is likely to reflect the combined effects of viral adaptation to the host's immune response and antiretroviral drugs. MHC class I alleles have been described to affect the evolution of the HIV-1 sequence on the population level in a single previous cohort from Western Australia (Moore et al. 2002; John et al. 2005). We were interested in examining whether these findings could be confirmed in our Western European cohort.

Four out of a total of twelve significant associations previously reported in the RT were confirmed in our cohort, all of which corresponded to known epitopes (Table 2.2). Moreover, we describe 11 new associations with four of these lying outside the region of the RT analyzed previously (Moore et al. 2002). Several of these associations also correspond to known T cell epitopes.

The identified CTL escape mutations, for which known epitopes exist, do not consistently lie in known anchor positions. Due to the complex pathway by which an epitope is processed before being displayed on the cell surface by the HLA-molecule and the

incompletely understood nature of T cell receptor binding and T cell activation, HIV-1 sequence mutations in non-anchor positions could very well still have a negative impact on the recognition of such epitopes and hence be beneficial for the virus.

These data support the hypothesis that the associations found are indeed epitopes targeted by CTL and that continuous pressure by CTL responses leads to the selection of survival strains that have mutations in these regions. The fact that we cannot reproduce all associations described in the previous publications and that we found a number of additional associations may be due to several reasons: the first cohort consisted of patients from Western Australia, whereas our patients are mainly Caucasians from Germany. Thus, the respective collected HLA sequences may differ and the access to and type of antiviral drugs may be different. However, the fact that we were able to confirm a number of associations despite these differences proves that our main conclusions, although affected by local factors, are generally true. Furthermore, HLA-associated mutations, once acquired, were stable and found in later viral isolates from the same patient. This supports the hypothesis that the continuing selection pressure by T cell recognition prevents the virus from reverting back despite changes in the antiviral drug regimen.

MHC class I dependent CTL responses shape the viral sequence by direct killing of infected cells that present the right epitope in their HLA molecules. So far, the impact of MHC class alleles II on viral evolution has only been studied in a small patient group (Harcourt et al. 1998). Harcourt et al. described HIV-1 sequence variation in the p24 GAG epitope of HLA-DR1. Our data from a large patient cohort support their observation of a role of MHC class II alleles and CD4 T cell responses in the evolution of the HIV-1 sequence. We found two associations in the RT region, also demonstrating that even class II alleles can exert selection pressure on viral sequences.

We also studied the effect of MHC class I and class II on the evolution of the full-length protease sequence. Interestingly, we did not find any of the associations which were described in the Western Australian cohort (John et al. 2005). The underlying reason for this is unclear. However, since most patients in our cohort and the Western Australian cohort were undergoing drug treatment it is possible that the type, time point and length of use of certain protease inhibitors may have affected the results in our study as well those in the previous studies. Nevertheless, three associations were found within previously described epitopes for the same HLA allele and four associations remained significant after correction for multiple testing, supporting the validity of the associations reported.

Although we were not statistically able to confirm the existence of "hotspots," as the effect reached only a weak level of statistical significance (*P*-value equal to 0.06), it appears that certain regions in the RT and protease are targeted by a greater number of epitopes and therefore CTLs, thus driving viral evolution in this region.

A study by Kiepiela et al. reported that the relative contribution of HLA-B alleles outweighs the contribution of HLA-A alleles in influencing HIV-1 disease outcome (Kiepiela et al. 2004). This is also reflected in our results, as more significant associations were found with HLA-B alleles (n = 15 or 68%) than with HLA-A (n = 5 or 23%) alleles or HLA-DRB1 alleles (n = 2 or 9%).

In most patients HLA-associated HIV-1 sequence mutations were consistently found in all sequences of that patient, while only relatively few cases had HLA-associated mutations that were not stable. These changes were generally from a wild type amino acid to the variant amino acid in the subsequent sample. This observation is consistent with other studies showing that CTL escape mutations develop soon after infection (Jones et al. 2004).

We also analyzed whether these HLA-associated mutations were influenced by drug therapy, and found that the impact of drug therapy on the associations in our patient cohort was not statistically significant. In some cases, patients in the cohort had not received a particular drug treatment (tipranavir or atazanavir), yet had HLA-associated mutations at positions where known drug resistance mutations to these particular drugs can also occur. As

tipranavir and atazanavir were first approved for antiviral therapy shortly after our patients' viral sequences were acquired, it is not possible for these patients to have been infected with an already resistant viral strain. This indicates that either these mutations have become locked in our population, so that the virus cannot revert back, or HLA-driven selection of mutations may predispose some patients to the development of some drug-resistance mutations. In order to optimize the therapy of patients carrying particular HLA-types, the choice of future therapy regimens should take into account the fact that some patients are more likely to develop particular HLA-associated mutations, which co-exist at residues of drug resistance mutations. This consideration could be made for positions in the protease that are linked to specific drugs (tipranavir and atazanavir), but not the majority of PIs, NRTIs or NNRTIs.

In summary, we were able to confirm immune-driven selection pressure not only by MHC class I alleles, but also by MHC class II alleles on the development of escape mutations at a population level. Further, we have described a number of HLA-associated mutations in the RT as well as protease and, finally, have analyzed their possible impact on the success of antiviral drug treatment.

3 HLA Class I Allele Associations with HCV Polymorphisms and Outcome of Antiviral Therapy in Patients with Chronic Hepatitis C[2]

3.1 Introduction

HCV is currently recognized as being a major cause of chronic liver disease both in the industrialized and developing world and thus is a significant global health problem. The virus has been shown to cause chronic hepatitis, liver cirrhosis, hepatocellular carcinoma and is a leading cause of liver transplantation worldwide. Available data suggest that the prevalence is approximately 2.2-3.0% world-wide, which is approximately 130-170 million people (McHutchison 2004; Suzuki et al. 2007; Lavanchy 2009).

A number of factors are thought to contribute to failure of the host to control infection. These include the impairment of cellular effector functions, suppression of antigen-specific cells by regulatory T cells, dendritic cell dysfunction, T cell exhaustion and the deletion of antigen-specific cells in the liver (Timm et al. 2007). Additionally, in the chimpanzee model, a further mechanism has been described leading to the failure of the host to control infection: the development of escape mutations in HCV led to viral persistence (Weiner et al. 1995; Erickson et al. 2001). Early evidence for such viral escape in humans came from chronically infected patients (Chang et al. 1997).

Studies suggest that immune control of HCV in humans is possible and that host factors play an important role in the ability of a host to mount a sufficient immune response against the virus. A response against a broad spectrum of viral epitopes by CD4 and CD8 T cells is necessary for viral clearance, but most patients are only able to provide a narrow repertoire of T cells specific for HCV antigens (Lechner et al. 2000; Lauer et al. 2004; Timm et al. 2007; Thimme et al. 2008).

The repertoire of T cells targeting differing viral epitopes in patients critically depends on the patient's HLA profile. HLA class I molecules are expressed on the surface of every nucleated cell in the human body and their role is to present fragments of cytosolic proteins eliciting a CD8 T cell response. While CD8 T cells do not respond to peptide fragments from healthy cells, they will recognize and respond to foreign protein fragments from viruses or cancers. HLA class I genes are highly polymorphic and thus there is enormous intra- and inter-individual allelic diversity. Differing HLA molecules have different peptide binding grooves which lead to a different composition of the HCV-derived epitope antigens presented to CD8 T cells by different patients. Patients with alleles such as HLA-A*03 and HLA-B*27 seem to have an advantage in clearing HCV by being able to produce a strong immune response, while HLA-B*08 appears to occur more often in those with chronic infections (McKiernan et al. 2004). The identification of specific sequence positions associated with viral escape may be used to identify new potential HCV epitopes, particularly for HCV genotypes such as 1b were little experimental work has yet been done.

While the HLA class II allele profile of patients infected with HCV has been correlated with aspects of disease outcome such as viral persistence, RNA viral load or liver

[2] The work reported in this section was performed in collaboration with Christian Lange and Christoph Sarrazin (University of Frankfurt, Germany). Most of the work has been published in the *Journal of Hepatology* and appears in this thesis with the journal's permission (Lange et al. 2010). My own contribution to the publication comprises the identification of escape mutations, the association of HLA alleles with clinical parameters and the analysis of phylogenetic relatedness resulting in 2 of the 3 published tables (Tables 3.2 and 3.3).

fibrosis progression, the influence of HLA class I alleles has been less clearly defined (Jiao and Wang 2005; Hraber et al. 2007; Wang et al. 2009).

Studies examining CD8 escape mutations have generally been limited in the terms of the cohort sizes and also in the number of HCV proteins/epitopes examined (Seifert et al. 2004; Timm et al. 2004; Cox et al. 2005; Ray et al. 2005; Tester et al. 2005; Urbani et al. 2005; Timm et al. 2007; Neumann-Haefelin et al. 2008a; Salloum et al. 2008). Rauch et al. were the only group so far to examine a large cohort of patients infected with HCV genotype 1a or 3 showing substantial differences between the two genotypes (Rauch et al. 2009).

In this study, one goal was to identify novel associations between HLA alleles and polymorphisms within the HCV genes which encode the proteins E2, NS3 and NS5B in a large patient cohort where patients were infected with either HCV genotype 1a or 1b. These escape mutations are thought to negatively influence the binding strength of the peptide to the HLA allele and therefore reduce the potential immune response against the virus; a detailed analysis was performed to analyze the binding strength of identified potential variant epitopes. Furthermore, it has been hypothesized that certain HLA alleles may be associated with successful antiviral therapy and therefore, treatment success was also correlated with the host HLA type. It was possible to show for the first time that the sustained virologic response rates of patients differed substantially between patient groups and some HLA alleles are associated with a successful antiviral therapy.

3.2 Patients and Methods

3.2.1 Patients

The cohort examined in this study consisted of 159 adult patients of both sexes monoinfected with HCV genotype 1 and suffering from chronic hepatitis C. They were selected before the initiation of antiviral therapy and had never received antiviral therapy before.

Patients whose liver disease was caused by other factors were excluded from the study. All patients had compensated (stable) liver disease. A liver biopsy was performed in 146 of 159 patients, prior to the initiation of treatment in order to grade and stage the level of chronic hepatitis. Serum aminotransferase (both aspartate and alanine aminotransferase) levels were also determined prior to antiviral therapy to assess liver function.

Antiviral therapy consisted of pegylated interferon alpha-2b combined with ribavirin within a randomized controlled treatment study (Berg et al. 2009). Patients received 1.5 µg/kg body weight pegylated interferon alpha-2b per week in addition to 800-1,400 mg ribavirin daily for 24 to 48 weeks. Close monitoring of patients occurred for safety, tolerance and efficacy of the antiviral therapy. All patients were outpatients. The treatment protocol was approved by a review board and all patients gave written informed consent.

A number of further variables were examined in the cohort:

- **Inflammation** indicates the degree of inflammation of the patient's liver. It is graded by a scoring system (Berg et al. 2009) where a score of 0 indicates no inflammation, 1 indicates minimal inflammation, 2 indicates mild inflammation and 3 indicates severe inflammation.
- **Fibrosis** indicates the degree of liver fibrosis. It is also graded by a scoring system where a score of 0 indicates no fibrosis, 1 indicates minimal fibrosis (only portal fibrosis), 2 indicates intermediate fibrosis (portal fibrosis with few septa), 3 indicates advanced fibrosis (severe portal fibrosis and many septa) and 4 indicates cirrhosis.

- **Baseline** is the viral load measured prior to beginning antiviral therapy and the unit of measurement is IU/ml.
- **Basl8** is the initial viral load prior to treatment under or over 800,000 IU/ml.
- **Basl4** is the initial viral load prior to treatment under or over 400,000 IU/ml. 400,000 IU/ml is currently considered to be the optimal baseline viral load cut-off to predict response in treatment-naive patients with genotype 1 HCV and thus differentiate relatively high and low viral loads.
- **Early treatment response** (ETR): the patient responds quickly and tests negative for HCV RNA in week 12 of the therapy.
- **Sustained virologic response** (SVR): the patient tests negative for HCV RNA 48 weeks after completing therapy.

3.2.2 HLA genotyping

HLA antigens were determined for all patients prior to beginning antiviral therapy. For the methods used to HLA genotype the patients, see Appendix C.

3.2.3 HCV RNA detection, quantification and genotyping

Before antiviral therapy was initiated, HCV RNA was quantified and the HCV genotyping was performed. For the methods used to genotype the HCV viruses, see Appendix C.

3.2.4 HCV gene sequencing

For the methods used to generate HCV gene sequences, see Appendix C.

3.2.5 HLA-associated mutations in the E2, NS3 and NS5B

All patient data, as well as HCV sequence data were entered into the database described in Section 2.2.2.

For the E2 protein, amino acid positions 464 to 576 (112 amino acids) were examined, for the NS3 protease amino acid positions 1026 to 1221 (195 amino acids) were examined and in NS5B RNA-dependent RNA polymerase amino acid positions 2670 to 2987 (317 amino acids) were examined in the sequences from all patients for which they were available.

Patients were separated into two groups for this analysis, depending on whether they were infected with HCV subtype 1a or 1b. Patients who were coinfected with HCV subtype 1a and 1b simultaneously were excluded from this analysis. The population consensus sequence of each genotype was used as reference sequence. The population consensus sequence was generated by assigning the most common amino acid for each position of all sequences pooled from the cohort of the indicated genotype (Yusim et al. 2005). HCV E2, NS3 and NS5B protein sequences from the patients were then aligned with the appropriate consensus sequence. The position of the identified polymorphism was then indicated by using the reference sequence H77 from the Los Alamos HCV Sequence Database (Kuiken et al. 2005; Kuiken et al. 2006; Kuiken et al. 2008). See the Appendix D and E for consensus sequences and the alignments of these protein sequences with reference sequences.

We analyzed each amino acid position using an extension of Fisher's exact test for associations with HLA alleles. As it requires very strong statistical power to make statements

about associations with rare alleles and our cohort was relatively small, we excluded very rare alleles from our analysis (less than or equal to 4% of the cohort). All HLA allele covariates with *P*-values of less than 0.05 were identified and fitted in a subsequent multivariate analysis to logistic regression models. We used binomial models, where the linear predictor consisted of all significant alleles and the response was a factor, in which variant amino acids were classified as successes. Correction for multiple-testing used the false discovery rate (FDR) method. Due to the low numbers of deletions and insertions in the sequences, no separate analysis was done to take these into account.

3.2.6 Association of HLA alleles with clinical parameters

Testing for significant associations between specific HLA alleles and a variety of clinical parameters was performed in which the parameters were treated as categorical data: inflammation, fibrosis, viral load less than 400,000 IU/ml before treatment, viral load less than 800,000 IU/ml before treatment, viral load greater than 800,000 IU/ml before treatment, achievement of SVR and EVR. Each case was assigned to one cell in a 2x2 contingency table and Fisher's exact test was used to identify significant differences.

3.2.7 Assessment of phylogenetic relatedness

An assessment for the presence of a founder effect type bias in the cohort was made, where an HLA allele is overrepresented in subgroup of individuals that have viral sequences sharing a recent common ancestor. We identified clusters of possibly related sequences and assessed the potential impact of such relatedness by performing analyses stratified by clusters (similar to Rauch et al. (Rauch et al. 2009)). In a cohort where no founder effect exists, the distributions of HLA alleles should be random across clusters of relatively homogeneous possibly related sequences. Protein sequences were aligned, sequence similarity scores were calculated, and subsequently a dissimilarity matrix, using the Euclidean metric, was generated. We used a robust generalization of k-means clustering known as partitioning around medoids (PAM) (Kaufman and Rousseeuw 1990). Optimal clustering was determined by assessing the silhouette widths for 2-20 medoids and selecting the maximal average silhouette width. For NS3, the maximal average silhouette width was 8, and for E2 it was 16 (Figure 3.1). NS5B was not included in the analysis because not all patients were sequenced for the gene (86 of 159 patients were sequenced). Individuals were assigned to their nearest cluster and the proportion of HLA alleles in the clusters was assessed using Fisher's exact test. Genotype 1a was aligned against genotype 1b to maximize comparability between genotypes. There were no statistically significant associations found, indicating that in this cohort there is a random distribution of HLA alleles across clusters of relatively homogenous sequences.

The method for assessing the phylogenetic relatedness of the HCV cohort differed from the method used for the HIV-1 cohort in Section 2.2.7. The reason for this change in approach were recommendations made by a reviewer of the manuscript, as the approach used here has been used in previous HCV studies and thus may be more easily understood by the medical community.

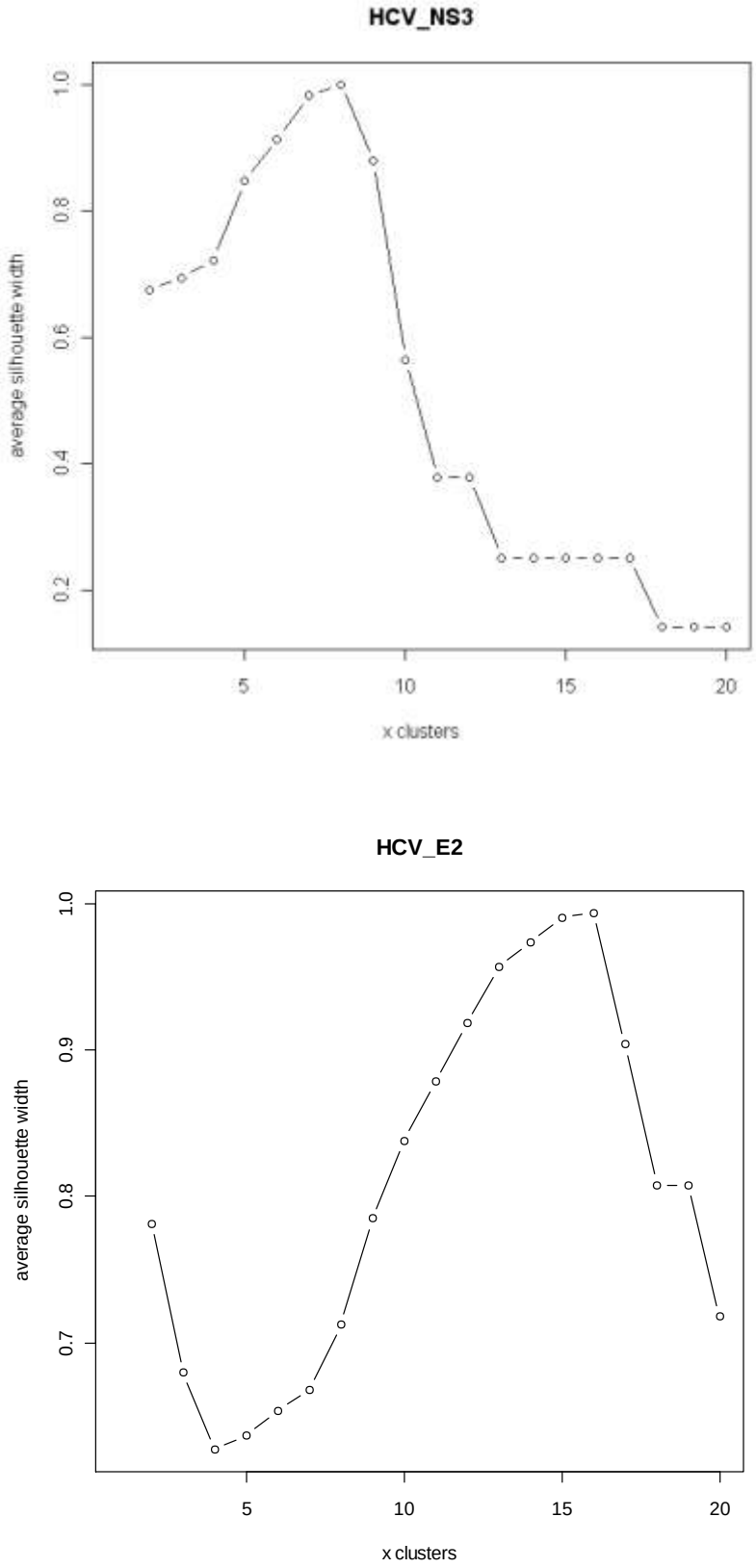

Figure 3.1 Silhouette widths for 2-20 clusters for NS3 and E2 gene sequences calculated using PAM

3.3 Results

3.3.1 Patient characteristics

The demographic and baseline characteristics of the cohort are summarized in Table 3.1. Although there is no specific data on ethnicity, the physicians involved in the study assert that most of the patients were Caucasian and representative of a Western European population. Most patients were male and the mean age of the cohort was 43 years. In the cohort, 43% were infected with HCV genotype 1a, 55% with genotype 1b and 2% were coinfected with both subtypes. During antiviral treatment, 68% achieved an early treatment response and after treatment 50% achieved a sustained virologic response.

Description		**Number (%)**
Total number of patients		159
Age	Mean	43
	Range	18-68
Female / Male		75 / 84
HCV Genotype	1a	69 (43%)
	1b	87 (55%)
	1a/1b	3 (2%)
HCV RNA, log10 IU/ml	Mean	5.74
	Range	2.79-6.90
ETR		108 (68%)
SVR		80 (50%)
Inflammation Stage	0	12 (8%)
	1	64 (40%)
	2	55 (34%)
	3	11 (7%)
	Unknown	17 (11%)
Fibrosis Stage	0	26 (16%)
	1	61 (38%)
	2	36 (23%)
	3	21 (13%)
	4	1 (1%)

Table 3.1 Baseline characteristics of the cohort

Abbreviation: SVR, sustained virologic response.

3.3.2 HLA distribution of the cohort

The frequencies of the HLA-A and HLA-B alleles in the cohort were examined (Figure 3.2). Of the most frequently occurring HLA-A alleles were broadly representative of Western European population studies found in the *dbMHC* database (Sayers et al. 2010), with the HLA-A*01 allele being slightly underrepresented and HLA-A*24 being somewhat overrepresented. Of the HLA-B alleles, HLA-B*07 was somewhat overrepresented and HLA-B*08 was somewhat underrepresented.

The highly polymorphic nature of the HLA-B locus is reflected by the large number of alleles occurring at a very low frequency. Previous work associated a number of alleles with viral clearance in patients with acute hepatitis C: HLA-A*03, HLA-B*27 and HLA-B*54. The frequency of these alleles in the cohort examined here was similar to the previous study (Neumann-Haefelin et al. 2008a).

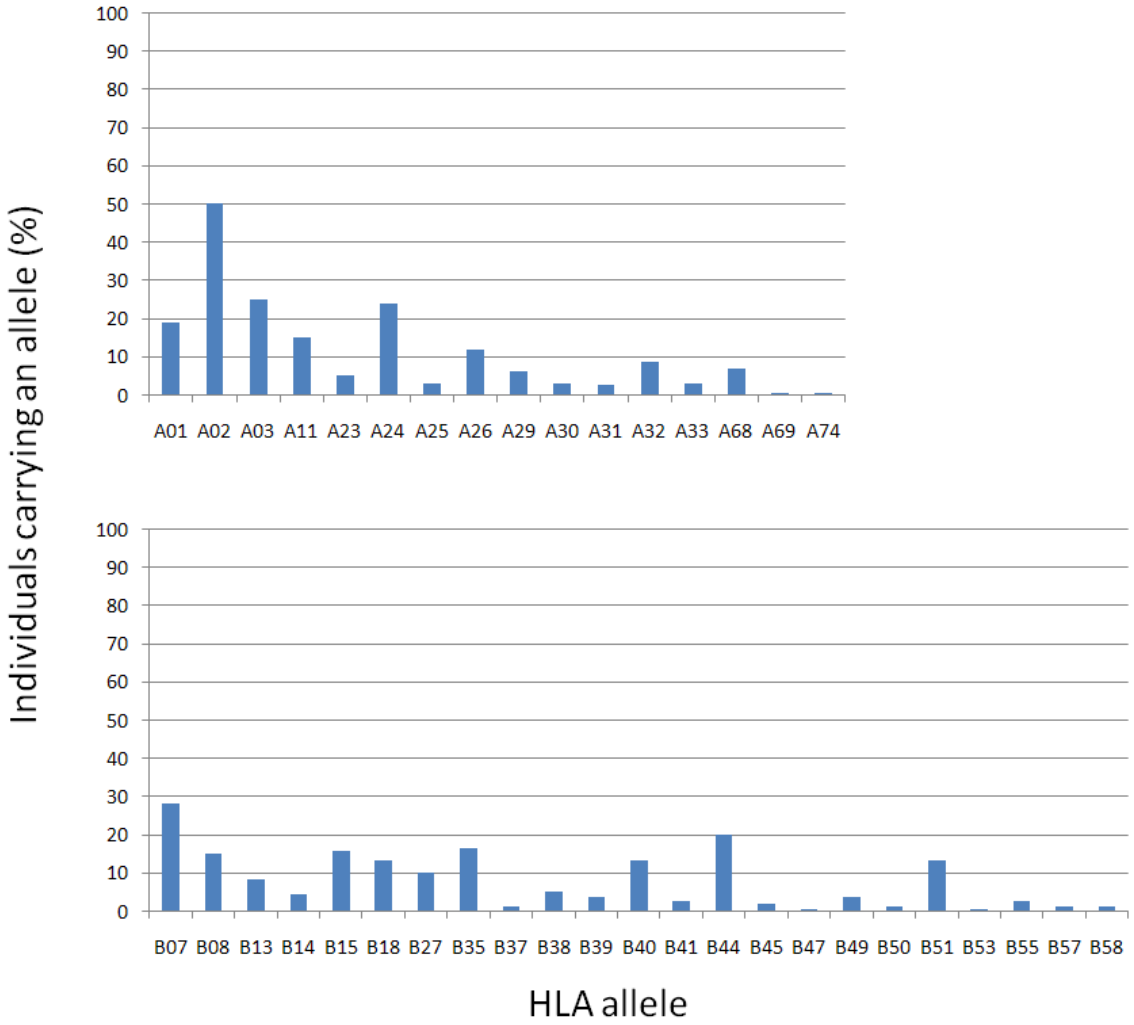

Figure 3.2 HLA distribution in the cohort

3.3.3 HLA-associated sequence polymorphisms in HCV proteins E2, NS3 and NS5B

Genes encoding the HCV proteins E2, NS3 and NS5B were directly sequenced from 159 patients with chronic hepatitis C. Sequences of genotype 1a and 1b isolates were independently aligned by using the population consensus sequence of each genotype (Appendix D). Overall, nine associations of HLA class I alleles with polymorphisms within HCV genotype 1a sequences and six associations with polymorphisms within HCV genotype 1b sequences were identified all with uncorrected *P*-values less than 0.05 (Tables 3.2 and 3.3). Associations were identified for HLA-A alleles with polymorphisms within E2 and NS3 and for HLA-B alleles with polymorphisms within E2, NS3 and NS5B. Only one of these associations had a *P*-value of less than 0.05 after correction for multiple testing.

#	Gene	Amino Acid Position	Polymorphism	HLA type	Absent (n)	Present (n)	OR	Fisher's P-value (unadjusted) Cutoff <0.01	Fisher's P-value (adjusted with FDR)
1	**E2**	466	E2-D466x	B*18	1	8	12.00	0.0096	>0.05
				non-B*18	36	24			
2	**E2**	473	E2-S473x	A*25	0	3		0.0023	>0.05
				non-A*25	59	7			
3	**E2**	475	E2-A475x	A*11	4	6	7.35	0.0077	>0.05
				non-A*11	49	10			
4	**E2**	492	E2-K492x	B*13	1	5	14.69	0.0086	>0.05
				non-B*13	47	16			
5	**E2**	570	E2-V570x	A*24	13	0	0.00	0.0027	>0.05
				non-A*24	32	24			
6	**NS3**	1087	NS3-T1087x	A*11	7	3	24.86	0.0084	>0.05
				non-A*11	58	1			
7	**NS3**	1115	NS3-Q1115x	B*07	15	6	9.20	0.0080	>0.05
				non-B*07	46	2			
8	**NS3**	1211	NS3-T1211x	A*68	0	3		0.0087	>0.05
				non-A*68	54	12			
9	**NS5B**	2690	NS5B-R2690x	B*15	1	2		0.0050	>0.05
				non-B*15	32	0			

Table 3.2 Positive HLA class I associated sequence polymorphisms, genotype 1a

OR = odds ratio

#	Gene	Amino Acid Position	Polymorphism	HLA type	Absent (n)	Present (n)	OR	Fisher's P-value (unadjusted) Cutoff <0.01	Fisher's P-value (adjusted with FDR)
1	**E2**	474	E2-Y474x	A*01	9	4	10.52	0.0085	>0.05
				non-A*01	71	3			
2	**E2**	483	E2-R483x	A*03	16	4		0.0022	>0.05
				non-A*03	67	0			
3	**E2**	492	E2-R492x	A*24	9	16	3.73	0.0086	>0.05
				non-A*24	42	20			
4	**E2**	522	E2-F522x	B*44	9	2	0.12	0.0065	>0.05
				non-B*44	27	49			
5	**E2**	538	E2-L538x	B*44	5	6	8.93	0.0028	>0.05
				non-B*44	67	9			
6	**NS5B**	2758	NS5B-V2758x	B*15	5	4	32.80	0.0023	**0.048**
				non-B*15	41	1			

Table 3.3 Positive HLA class I associated sequence polymorphisms, genotype 1b

OR = odds ratio

Overall, it can be said that HCV genotype 1b is relatively poorly represented in epitope databases and therefore the potential for prediction servers to have large enough datasets for accurate learning is relatively low. In November 2010, the IEDB contained 311 binders and 535 non-binders for HCV genotype 1b. This database explicitly excludes HIV epitopes (Vita et al. 2010). HIV epitopes can be found in the Los Alamos HIV Molecular Immunology Database; there 2398 binders are identified as being derived from HIV-1 subtype B, non-binders are not specifically identified. Both HIV-1 and HCV genomes have

similar lengths (9.2 kb vs. 9.6 kb) and should thus contain, roughly, similar numbers of potential epitopes.

The number of patients in our cohort was small and some of the associations were borderline significant after correction for multiple testing. Therefore an additional method for validating the identified associations was chosen. The web-based prediction servers *NetMHC* (Lundegaard et al. 2008) and *SYFPEITHI* (Rammensee et al. 1999) were used to identify possible HCV CD8 epitopes, because there is a relative lack of experimentally verified HCV epitopes available in public databases. (This contrasts markedly with HIV-1 were many experimentally verified epitopes are available, see Section 2.3.2). The rationale for using *NetMHC* was the excellent performance of this prediction method in a number of comparison studies, and *SYFPEITHI* was chosen because it has been used in a number of previous studies examining escape mutations in HCV. Additionally, the HCV Immunology Database at Los Alamos was examined for known CTL epitopes; this database contains the same HCV data as the Immune Epitope Database (IEDB), but is structured in a more user friendly manner.

Of the 15 predicted epitopes, seven overlapped with previously described epitopes indicating that the epitopes are located in highly immunogenic regions (Tables 3.4 and 3.5). Of the 15 epitopes, seven were identified as potential weak binders by *NetMHC*, and a further epitope had what can be described as an intermediate binding affinity (778 nM). The *SYFPEITHI* prediction server uses a scoring system in which a score of greater than or equal to 20 indicates a high likelihood of being a true epitope. This was the case for five epitopes and in three further epitopes, the associated mutation was located in anchor residue positions.

Overall, eight of the 15 identified polymorphisms have a score of at least 20 using *SYFPEITHI* or an affinity of less at most 500 nM using *NetMHC*. Larger patient datasets are required to add sufficient power to the associations, with the exception of HLA-B*15 RVFTEAMTRY$_{2757\text{-}2766}$ which already showed an adjusted P value of less than 0.5. Additionally, it should be noted again that there is a lack of experimental data identifying epitopes for HCV genotype 1b making the verification of the identified 1b epitopes difficult.

#	Gene	Allele	Amino Acid Position	Amino Acid	SYFPEITHI Prediction	Score	NetMHC Prediction	Affinity (nM)	HCV Immunology Database CTL Epitope	Allele
1	E2	B*18	466	D	DQGWGPISY	17	DQGWGPISY	127 (WB)	RPLTDFDQGW	B53
2	E2	A*25	473	S	-	-	-	-	DFAQGWGPISYANGS	human
3	E2	A*11	475	A	ANGSGPDHR	13	-	-	DFAQGWGPISYANGS	human
4	E2	B*13	492	K	-	-	-	-	YPPKPCGI	B51
5	E2	A*24	570	V	VCGAPPCVI	13	-	-	CVIGGAGNNT	B53
6	NS3	A*11	1087	T	GTRTIASPK	25	GTRTIASPK	94 (WB)	TRTIASPKGPVIQMY	human
7	NS3	B*07	1115	Q	APQGARSLT	21	APQGARSLT	183 (WB)	DLVGWPAPQGSRSLT	human
8	NS3	A*68	1211	T	TTMRSPVFT	13	TTMRSPVFT	84 (WB)	-	-
9	NS5B	B*15	2690	R	NSRGENCGY	11	NSRGENCGY	414 (WB)	-	-

Table 3.4 Putative epitopes associated with sequence polymorphisms identified in the cohort, genotype 1a

The identified polymorphisms are indicated in red.

#	Gene	Allele	Amino Acid Position	Amino Acid	SYFPEITHI Prediction	Score	NetMHC Prediction	Affinity (nM)	HCV Immunology Database CTL Epitope	Allele
1	E2	A*01	474	Y	AQGWGPIT**Y**	17	-	-	-	-
2	E2	A*03	483	R	DQ**R**PYCWHY	12	-	-	-	-
3	E2	A*24	492	R	HYAP**R**PCGI	23	HYAP**R**PCGI	319 (WB)	-	-
4	E2	B*44	522	F	VVVGTTDR**F**	12	-	-	-	-
5	E2	B*44	538	L	GENETDVL**L**	22	GENETDVL**L**	778	-	-
6	NS5B	B*15	2758	V	R**V**FTEAMTRY	23	SLR**V**FTEAM	91 (WB)	-	-

Table 3.5 Putative epitopes associated with sequence polymorphisms identified in the cohort, genotype 1b

The identified polymorphisms are indicated in red.

The statistically most significant association in this study occurred in the NS5B gene at position V2758 and was associated with HLA-B*15. This polymorphism has not been described before. In the analysis of putative epitopes, both *SYFPEITHI* and *NetMHC* predicted an epitope in this region, of a score of 23 and a 91 nM binding affinity respectively. However, the exact position of the predicted epitopes were different with each method (Table 3.5). In the case of *SYFPEITHI*, V2758 is located at a main anchor position for HLA-B*15 (Prilliman et al. 1999), whereas, interestingly, in the case of *NetMHC* the variant is located neither at a known primary nor secondary anchor position. However, escape mutations at non-anchor residue positions have been described before and probably, in these cases, have a greater impact on CTL binding than on MHC binding (Moskophidis and Zinkernagel 1995).

The potential existence of a founder effect in the cohort was also evaluated by performing a phylogenetic analysis of cohort (Figure 3.3). The analysis revealed high phylogenetic distances between different patients. Therefore a shared evolutionary history of the viruses infecting the cohort being responsible for the identified HLA-restricted polymorphisms is unlikely. Also, a further assessment of phylogenetic relatedness (Section 3.2.7) was unable to determine the existence of a founder effect.

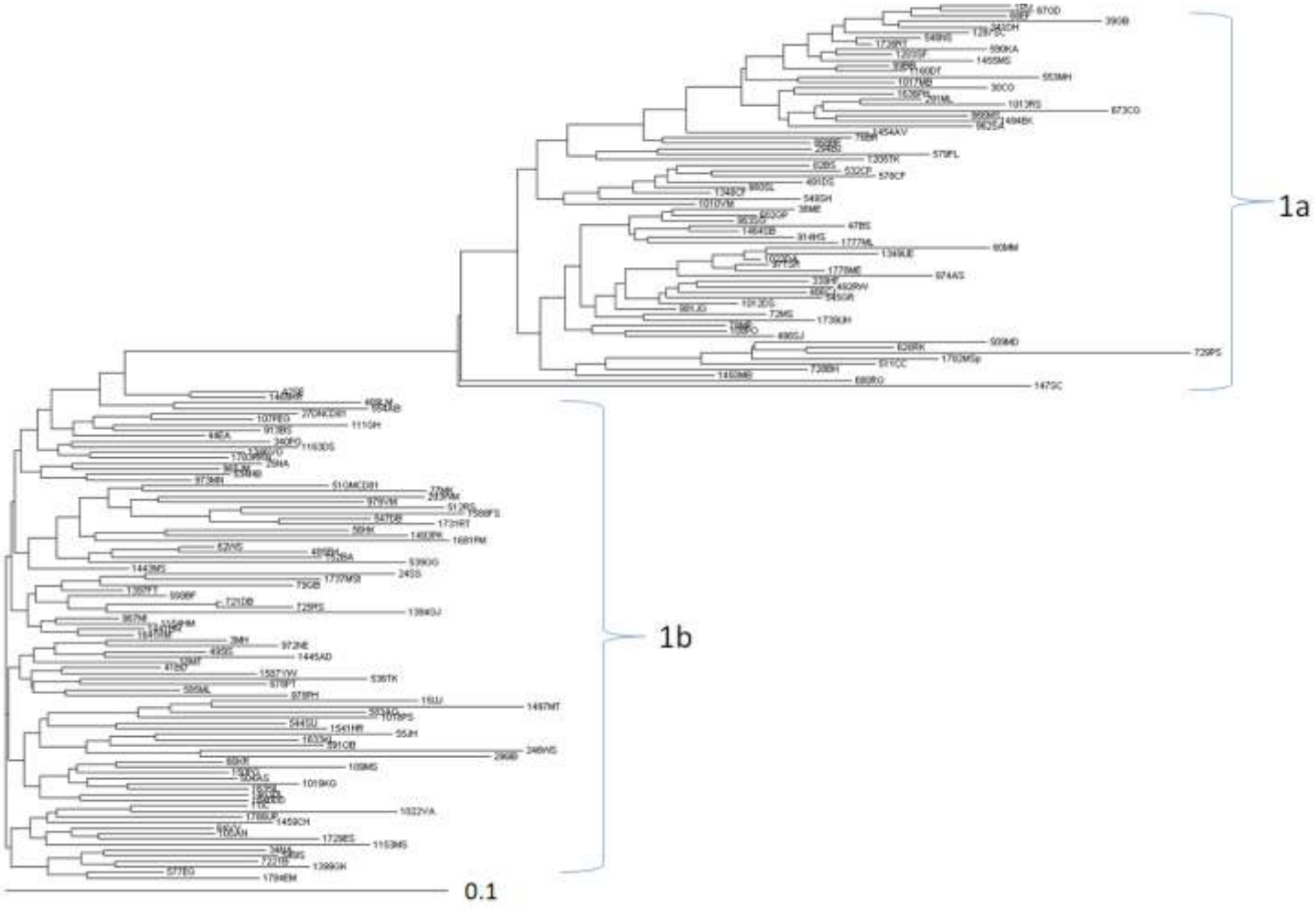

Figure 3.3 Phylogenetic tree of the viral sequences drawn from the cohort

Clusters representing HCV subtypes 1a and 1b are indicated. Lengths of branches represent phylogenetic distances between HCV isolates of different patients. The tree reveals high phylogenetic distances between isolates from different patients.

3.3.4 Association of HLA alleles with outcome of antiviral therapy

The sustained virologic response rate of the cohort was, with 50%, in the expected range of patients infected with HCV genotype 1 under controlled trial conditions. In order to investigate whether particular HLA-A or HLA-B alleles were associated with a higher or lower chance of successful treatment, patient allele types were associated with achieving SVR (Table 3.6). In the case of patients with the allele HLA-B*07 which were infected with HCV genotype 1a, the allele was associated with treatment success (for patients with HLA-B*07, SVR was 66%, for patients negative for HLA-B*07, SVR was only 33%, *P*-value less than 0.05). However, patients carrying the HLA-B*44 allele had a reduced chance of treatment success when infected with HCV genotype 1a (for patients with HLA-B*44, SVR was only 20%, but for patients negative for HLA-B*44, SVR was 55%, *P*-value less than 0.01). Finally, patients infected with HCV genotype 1b carrying the HLA-A*-02 allele had a better treatment response (for patients with HLA-A*02, SVR was 65%, for patients negative for HLA-A*02, SVR was only 41%, *P*-value less than 0.05).

HCV Genotype	Allele	SVR	no SVR	P-value
1a	B*07	14	7	0.0201
	non-B*07	17	31	
1a	B*44	4	16	0.0088
	non-B*44	27	22	
1b	A*02	31	17	0.0299
	non-A*02	16	23	

Table 3.6 Association of HLA class I alleles and sustained virologic response or treatment success

The overall SVR rate was 50%, which is in the expected range for patients infected with HCV genotype 1 under control trail conditions.

The stage and grade of chronic hepatitis C was also associated with patients carrying particular HLA alleles (Tables 3.7 and 3.8). In particular, patients infected with HCV genotype 1a and carrying allele HLA-A*26 had higher levels of both liver inflammation and fibrosis (*P*-value less than 0.05). In contrast, patients infected with HCV genotype 1b, had lower levels of liver fibrosis (HLA-A*02 and HLA-A*68) and lower levels of inflammation (HLA-A*24). HCV RNA viral load or early treatment response was not correlated with the presence of any HLA allele.

HCV Genotype	Allele	Inflamation Stage <= 1	Inflammation Stage > 1	P-value
1a	A*26	2	8	0.0124
	non-A*26	36	19	
1a	B*08	9	1	0.0370
	non-B*08	29	26	
1b	A*24	8	16	0.0299
	non-A*24	34	22	

Table 3.7 HLA class I alleles associated with degree of liver inflammation

HCV Genotype	Allele	Fibrosis Stage <= 1	FIborsis Stage > 1	P-value
1a	A*26	3	7	0.0246
	non-A*26	39	15	
1b	A*02	29	14	0.0239
	non-A*02	15	22	
1b	A*68	7	0	0.0147
	non-A*68	37	36	

Table 3.8 HLA class I alleles are associated with degree of liver fibrosis

3.4 Discussion

In this study, the HLA class I phenotype of patients with chronic HCV genotype 1 infection was shown to be correlated with viral polymorphisms and the outcome of antiviral therapy consisting of pegylated interferon alpha-2b and ribavirin. It was possible to (1) identify several as yet unknown HLA class I-associated viral escape mutations in both HCV genotype 1a and 1b quasispecies located within novel CD8 T cell epitopes. Additionally, (2) it was possible to associate certain HLA class I alleles with successful antiviral treatment.

Overall, we were able to identify 15 HLA-restricted HCV polymorphisms in viral genotypes 1a and 1b. Our cohort size, while larger than almost all previous studies, still suffered from being relatively small with only 69 patients infected with HCV genotype 1a and 87 patients infected with HCV genotype 1b. This led to relatively small statistical power and only one of the polymorphisms withstood correction for multiple testing.

For this reason we performed a detailed epitope analysis using a variety of web-based prediction servers and epitope databases. Of the identified 15 HLA-restricted polymorphisms, 12 either scored highly (greater than or equal to 20) with *SYFPEITHI*, were identified as a weak binder with *NetMHC* (binding affinity between 50-500 nM) or lay within an experimentally identified epitope in the Los Alamos HCV Immunology Database. Also, six polymorphisms were located in putative anchor positions in the HLA peptide groove. It is thought that mutations in epitopes which are located in such primary anchor positions are particularly disruptive to HLA-peptide complex formation and hence substantially reduce the likelihood of effective TCR recognition (Rammensee et al. 1999). To our knowledge, only four of the identified putative HCV CD8 T cell epitopes overlap with previously described epitopes, but these have been described to be associated with other HLA alleles (Thimme et al. 2002; Wertheimer et al. 2003; Lauer et al. 2004; Timm et al. 2007; Neumann-Haefelin et al. 2008b; Thimme et al. 2008). This observation indicates that a high level of immunogenicity originates from these HCV protein regions, as individuals of different HLA phenotypes present only slight different epitopes. This data confirms the need for further immunological studies in larger chronically infected patient cohorts and substantially extended *in vitro* binding affinity studies to identify both HCV peptides which are binders or non-binders, especially in the case of as yet poorly studied HCV genotypes such as 1b.

Recent work has shown that there are substantial differences in the adaptation of HCV genotypes 1a and 3 to immune pressure. In this study, it was possible to independently analyze the quasispecies of HCV genotype 1a and 1b infected patients and show that there are even significant differences at the viral subtype level (Rauch et al. 2009).

The putative epitope in the NSB5 protein of HCV genotype 1b was identified as $RVFTEAMTRY_{2757\text{-}2766}$ by *SYFPEITHI* and $SLRVFTEAM_{2755\text{-}2763}$ by *NetMHC*. Due to the relatively low amount of experimental binding affinity data available for HCV genotype 1b in the publically available databases, it can be assumed that there is currently insufficient data available to train prediction algorithms to perform with high levels of accuracy. In particular, since the polymorphism located within this particular eptiope,V2758X, showed the highest level of significance after correction for multiple testing lends support to this polymorphism being located in a highly immunogenic region. Also, since as of yet no potential epitopes have been described within this region for genotype 1b infected patients, there appear to be immunogenic differences between the virus quasispecies in this region (Kuiken et al. 2005; Yusim et al. 2005; Kuiken et al. 2006).

Prior studies that examined HLA type and treatment success in chronically infected patients concentrated on HLA class II alleles or were performed prior to the widespread availability of pegylated interferon alpha-2b (Jiao and Wang 2005; Patel et al. 2006; Gaudieri et al. 2009). Therefore, this study provides novel information on the role of HLA class I alleles in response to current state-of-the-art therapies. It should be noted however, that the while the differences in SVR with alleles HLA-B*07, HLA-B*44 and HLA-A*02 are striking they are lower than recently identified correlations of treatment success with polymorphisms near the IL-28B gene (Ge et al. 2009; Thomas et al. 2009).

In summary, it was possible to identify several unknown HLA class I-restricted polymorphisms located within putative CD8 T cell epitopes and show substantial differences between differences between the viral quasispecies genotype 1a and 1b. Furthermore, the treatment success with pegylated interferon alpha-2b and ribavirin is significantly increased in patient groups of particular HLA genotypes.

4 ViralDAS[3]

4.1 Introduction to BioSapiens and DAS

BioSapiens was a Network of Excellence funded by the European Union's 6th Framework Program from 2004 to 2009 (http://www.biosapiens.info/page.php). The goal of the network was to provide a large scale, concerted effort to annotate genome data. Included in the network were 25 institutions from 14 countries, including the Max Planck Institute for Informatics, which was involved in Work Package 15 entitled "Thematic Work Package on Infectious Diseases."

The method chosen to make such a large scale annotation possible was through the Distributed Annotation System (DAS, see description below). The annotations are available in public domain and easily accessible through a single portal called the DAS Registry (http://www.dasregistry.org).

DAS is a network protocol which was developed to enable the exchange of biological data (Dowell et al. 2001; Prlic et al. 2007). The protocol was originally targeted to the genome annotation community to solve the problem of "the frustration of integration." Genomic sequence, protein sequence or structure annotations are decentralized and stored by third-party annotators and integrated on an as-needed basis by client-side software. Therefore, there is no dependency on particular database schemas or technologies or client-side technologies, and thus an uncoupling of reference and annotation servers occurs.

DAS consists of three parts: (1) centralized database resources such as Entrez (Sayers et al. 2010), Ensembl (Flicek et al. 2010) and Interpro (Hunter et al. 2009) which perform the function of reference servers and provide genomic and protein sequence as well as structure data, (2) various annotation servers using such software packages as LDAS (2001) and ProServer (Finn et al. 2007) which supply specialized annotation for the data provided by the reference servers and (3) a client viewer such as Dasty (Jones et al. 2005) or SPICE (Prlic et al. 2005) which displays sequence or structure information from the reference servers and various user-selected annotation layers. Communication between the client layer and the servers is defined by the DAS XML specification.

4.2 Implementation of ViralDAS

As there were no reference servers available for virus genomes such as HIV-1 or HCV, a reference server using the GBrowse software package was designed and implemented (Stein et al. 2002): the ViralDAS server for HIV-1 and HCV is the first DAS Server for viral genomes (Marcus 2008). It was specifically adapted from the default DAS server type to meet the requirements posed by a viral genome. The DAS server configuration that has been developed can easily be extended to other viruses which have similar features.

In order to implement ViralDAS, several new feature types had to be created and existing feature types had to be adapted. The current DAS server now supports the feature types transcripts, polyproteins, cleaved proteins and accessory proteins (Figure 4.1). For example, the seven most common transcripts of HIV-1 were annotated (over 30 are known, but many occur only rarely). The three polyproteins of HIV-1 are shown in a separate track.

[3] The work described in this section was performed as part of BioSapiens Network of Excellence (http://www.biosapiens.info/page.php). My own contribution comprises designing and implementing the ViralDAS server. ViralDAS is described in internal reports for BioSapiens and mentioned in a book on collaborative research and resources in bioinformatics (Marcus 2008).

Below the polyprotein track, one can see the track which shows the proteins produced by the cleavage of the polyproteins, as well as other accessory proteins. Over each transcript, polyprotein or protein, the gene name is shown as the three-letter abbreviation e.g. ENV. The protein glyphs are all in "musical staff notation" thus displaying the correct reading frame of each protein. For HIV-1 HXB2 strain numbering has been used (Korber et al. 1998).

The server can be reached via the following URLs:

- http://viraldas.bioinf.mpi-inf.mpg.de
- http://www.dasregistry.org

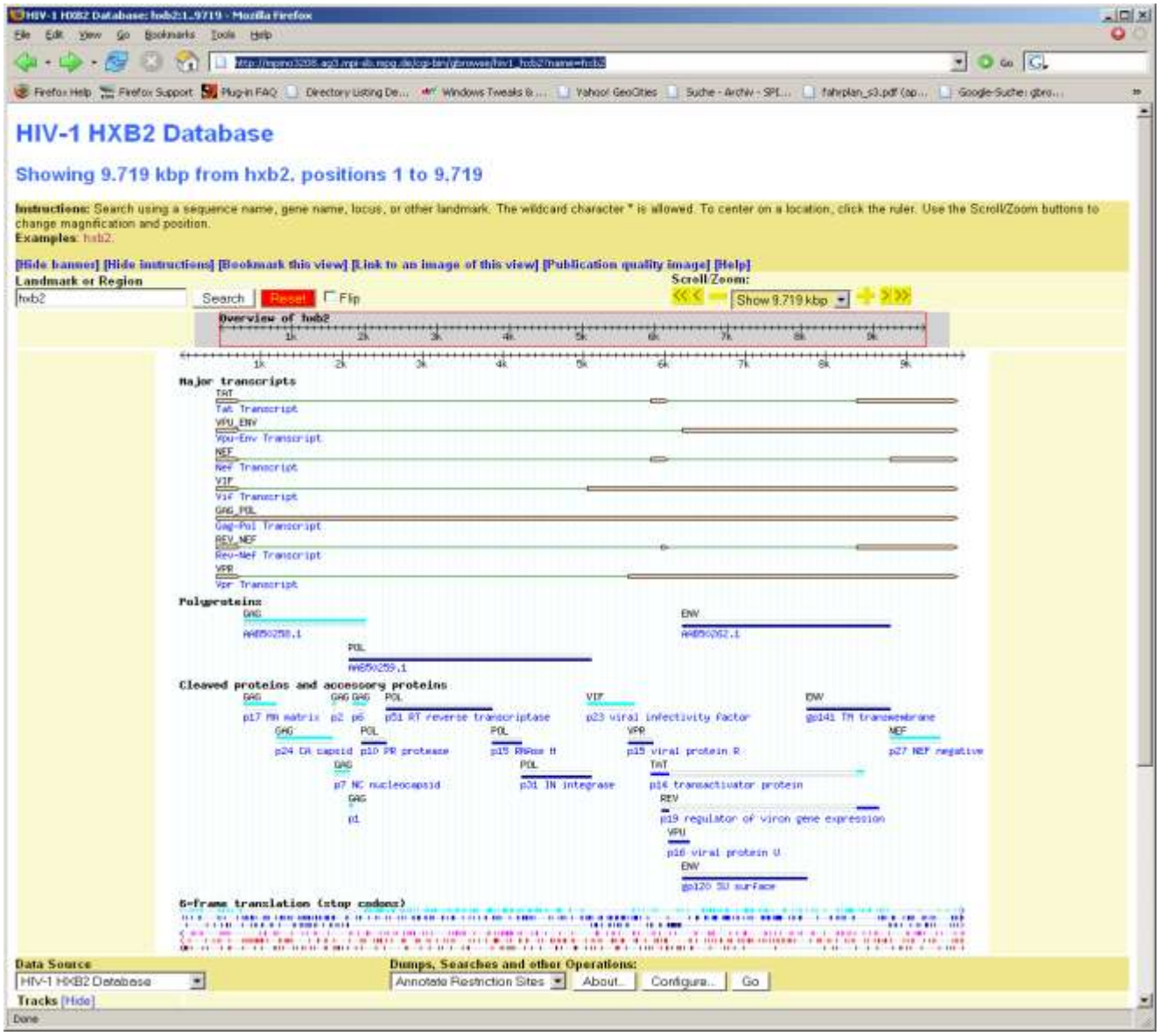

Figure 4.1 Screenshot of the main HIV-1 ViralDAS page

4.3 Outlook

Future possible developments, which would provide more annotation tracks that are specific to the viral genomes, are numerous. HIV-1 develops a large number of resistance mutations which reflect the HAART (combination drug therapy) therapy that the infected individual has undergone. Currently, the drugs which are widely used target several viral

proteins and cause the viral genome to develop resistance mutations in them. Additionally, the virus adapts to the HLA-immune profile of the host and has been shown to produce escape mutations which prevent the host's immune system from recognizing the virus effectively and combating the infection. In fact, particular regions of the virus, known as epitopes, are particularly prone to developing escape mutations depending on the HLA-profile of the host. Finally, there are also a number of HIV-1 viral subtypes and viral recombinant forms, which vary in their prevalence in different regions of the globe. All these aspects could also be implemented for HCV as well.

Linkage to the annotation of the human genome in context with the viral infection still has to be conceptualized and implemented.

5 Predicting MHC Class I Epitopes in Large Datasets[4]

5.1 Introduction

A precise understanding of host immune responses is crucial for basic immunological studies as well as for designing effective disease prevention strategies. Epitope-based analysis methods are effective approaches at assessing immune response, allowing for the quantification of the interaction between a host and pathogen, of vaccine effectiveness or other prevention strategies.

MHC class I molecules deliver peptides from the cytosol and are recognized by CD8 T cells. The binding of antigenic peptides from pathogens to MHC class I molecules is one of the crucial steps in the immunological response against an infectious pathogen (Paul 1998). While not all peptides that bind MHC molecules become epitopes, all T cell epitopes need to bind to MHC molecules. Therefore, deciphering why certain peptides become epitopes and others do not is central to the development of a precise understanding of host immune responses.

The Immune Epitope Database and Analysis Resource (IEDB) (Peters et al. 2005a; Peters et al. 2005b) is a central data repository and service, containing MHC binding data relating to B cell and T cell epitopes from infectious pathogens, experimental pathogens and self-antigens (autoantigens). In most cases, T cell epitopes are defined as peptides that are not only presented to T cell receptors on the cell surface by specific MHC molecules, but that also trigger an immune response. IEDB encompasses patent data from biotechnological and pharmaceutical companies, as well as direct submissions from research programs and partners. Within the database, epitopes are linked with objective and quantifiable measurements with regard to their binding affinity to specific, well defined immune system receptors.

IEDB is not the first database to store such information, as there are a number of databases which include similar information. However, although most of the components of IEDB can be found in other resources, none contains them all. For example, *SYFPEITHI* (Rammensee et al. 1999) contains carefully mapped epitopes or naturally processed peptides, but unlikely IEDB, does not annotate the context in which they are immunogenic. The Los Alamos HIV Molecular Immunology Database (2007), focuses on a restricted dataset. FIMM (Schonbach et al. 2002), is of modest size and solely focuses on cellular immunology and MHCPEP (Brusic et al. 1998), while still widely used, has not been updated since 1998. While MHCBN (Lata et al. 2009) and AntiJen (Blythe et al. 2002; Toseland et al. 2005) contain peptide entries that are not contained in IEDB, IEDB has more entries than any other existing database in this field.

While IEDB is the first epitope database of significant size, the experimental screening of large sets of peptides with respect to their MHC binding capabilities is still very demanding due to the large number of possible peptide sequences and the extensive polymorphism of the MHC proteins. Therefore, there is significant interest in the development of computational methods for predicting the binding capability of peptides to MHC molecules, as a first step towards selecting peptides for actual screening.

[4] The work reported in this section was performed in collaboration with Iris Antes (MPI for Informatics/TU Munich, Germany). It has been published in the journal *BMC Bioinformatics* and appears in this thesis based on the permissions granted by BioMed Central's Open Access Charter (Roomp et al. 2010). My own contribution to the publication comprises selecting the datasets for the analysis, performing the prediction analysis (including robustness and generalizability) resulting in 4 of the 4 published tables (Tables 5.1, 5.2, 5.3 and 5.4) and 3 of 3 published figures (Figures 5.1, 5.2 and 5.3).

Sequence- and structure-based methods, as well as combinations thereof, have been developed and were used for both classification and regression. Classification models aim to distinguish binding from non-binding peptides, whereas regression methods attempt to predict the binding affinity of peptides to MHC molecules. As the quantity of publicly available binding data has been limited until recently, most methods focus on classification. A review of previous methods can be found in Tong et al. (Tong et al. 2007).

Sequence-based methods are computationally more efficient than structure-based methods. However, they are hampered by the need for sufficient experimental data and therefore only achieve high performance on already intensively investigated MHC alleles. Additionally, sequence-based methods do not provide a structural interpretation of their results, which is of importance for designing peptidic vaccines and drug-like molecules. Structure-based methods have the advantage of being independent of the amount of available experimental binding data, but are computationally intensive and therefore not suited for the screening of large datasets.

A recent approach (Antes et al. 2006) performs a combined structure-sequence-based prediction by incorporating structural information obtained from molecular modeling into a sequence-based prediction model. This method therefore not only allows for the fast prediction of MHC class I binders, but also for the efficient construction of docked peptide conformations. This approach is the only prediction method available today, which also allows for the construction of such conformations. We have evaluated this approach for MHC class I alleles of the HLA-A and HLA-B loci for which extensive datasets were available in IEDB and compared it to two sequence-based prediction methods from the literature. These two sequence-based prediction methods are the same as examined in Antes et al. (Antes et al. 2006) and were chosen here as well for comparison reasons. In addition, we have evaluated the prediction server *NetMHC* which has shown to be among the best predictors in recent comparison tests (Lin et al. 2008).

A major focus of this study is on testing the dependency in performance of well established methods on the use of different training and testing datasets. The four methods we have chosen span a representative cross section of available methodology for MHC-peptide binding predictions, from simple binary (*SVMHC*) to rather sophisticated encoding (*DynaPred*^{POS}). The chosen methods include advanced learning strategies such as support vector machines (SVM) (*DynaPred*^{POS} and *SVMHC*) and artificial neural networks (ANN) (*NetMHC*), as well as the more straightforward quantitative matrix based prediction (*YKW*).

Lin et al. (Lin et al. 2008) performed a comparative evaluation of thirty prediction servers developed by 19 groups using an independent dataset. Each server was accessed via the Internet and the predictions were recorded, normalized, and compared. Peters et al. (Peters et al. 2006) performed an extensive analysis of predictors, in which they trained and tested their in-house methods, but did not reimplement any of the external methods used. Instead, web interfaces were used for the external methods. Zhang et al. (Zhang et al. 2009) evaluated five prediction methods using public web interfaces with the default parameters of the methods in question. Three of these methods were in-house and two were external. The authors discarded all peptides used for training their own methods for subsequent testing, but there was some concern that there was some overlap between evaluation and training data sets for one of the external methods. They also trained their own predictors on a dataset of binders and non-binders for a wide variety of alleles, testing on a second set of binders and non-binders which were released at a later point in time. The goal of this analysis was to examine the performance of these predictors on alleles for which little or no data was available (which were described as pan-specific predictors). The alleles for which such pan-specific analyses were performed were not identified and only limited information on the methods performance was available.

Work by other groups which preceded the study by Zhang et al. (Zhang et al. 2009), included a support vector machine based approach (Jacob and Vert 2008), which was trained

and tested on a relatively small datasets, where different predictive models are estimated for different alleles, using training data from 'similar' alleles. The notion of allele similarity is defined by the user therefore requiring human intervention which is not a systematic procedure. A binding energy model (Jojic et al. 2006), which was trained and tested on very small datasets, was used to make pan-specific predictions for only two alleles. A further study (Zhang et al. 2005), which utilized hidden Markov models and artificial neural networks as predictive engines, was again trained on relatively small datasets. The system was used to identify so-called promiscuous peptides, which bind well to a number of diverse alleles.

In this study, we reimplemented the external methods *YKW* and *SVMHC* and trained and tested them along side our in-house method *DynaPred*^{POS} on a wide variety of datasets. This allows for a more objective comparison of the performance of these methods. We also tested *NetMHC* in all the tests where training was not necessary; this server is only available via a web interface and thus could not be reimplemented for this study. In contrast to the previous work described, we perform a detailed analysis of the performance of predictors trained on one allele and their ability to accurately predict other alleles.

5.2 Methods

5.2.1 Datasets with complete peptides

In this section we describe the datasets we use that contain data incorporating full peptides, i.e. information on all nine residues. IEDB was mined for allele/peptide data on May 16th, 2007. Only alleles with a significant number of 9-mer binding and non-binding peptides (the total number being greater than 200) where included in the analysis (Table 5.1). The data was imported into a local relational database to allow for efficient analysis.

Three datasets were generated for each allele: all peptides available in IEDB (the full dataset or Dataset F), all peptides with an available quantitative laboratory test result (IC_{50} or in rare cases EC_{50}), but including only those with a binding affinity between 50 nM and 1000 nM, i.e. including only weak binders and non-binders (the intermediate dataset or Dataset I) and all peptides with an IC_{50}, but excluding those with a binding affinity between 10 nM and 10,000 nM, i.e. including only very strong binders and very clear non-binders (the strong dataset or Dataset S). Alleles with less than 200 peptides in total (binders and non-binders) were excluded from the analysis in all datasets. IC_{50} measures the half maximal (50%) inhibitory concentration (IC) of a radioactive isotope labeled standard peptide to MHC molecules, whereas EC_{50} measures the half maximal effective concentration (EC) of such a reference peptide (Peters et al. 2005a; Peters et al. 2005b). For Dataset F, in cases where a peptide with a particular sequence had more than one entry in IEDB for a particular allele (for example the peptide was tested with same allele by two different laboratories, resulting in two separate IEDB entries), this peptide was included only once. If no binding constant was available, the peptide was also only included once in Dataset F. If a peptide was described as a binder by one laboratory and a non-binder by another laboratory, it was included both as a binder and non-binder in Dataset F. For Datasets I and S, in cases where peptide-allele complexes had duplicate entries in IEDB and the binding affinities differed, resulting in at least one entry with a binding affinity which fell within the ranges used in Datasets I and S, this peptide was included in the respective dataset. If, for a particular allele, there was more than one binding affinity measurement made that fell into one of the ranges used in the analysis, an average binding affinity was calculated and used for that peptide in that particular range. Any peptides annotated in IEDB as binders with IC_{50} values greater than 500 nM, and peptides annotated as non-binders with IC_{50} values less than 500 nM were discarded. We have made the three datasets available:

- http://www.mpi-inf.mpg.de/~roomp/benchmarks/list.htm

We also tested an independent dataset recently published by Lin et al. (Lin et al. 2008), derived from the tumor antigen survivin and the cytomegalovirus internal matrix protein pp65. This data was not used for training any of the four prediction methods in this study and therefore serves as an independent test set.

5.2.2 Prediction methods

Four prediction methods were evaluated with respect to their ability to correctly classify binders and non-binders in the datasets described above. As already described in the introduction, the chosen methods span a representative cross-section of available methodology for MHC binding predictions.

The first prediction method used (*DynaPred*^{POS}) was developed in our laboratory (Antes et al. 2006). The general strategy followed to generate this prediction model, involves as a first step molecular dynamics simulations from which energetic information for all 20 amino acids in each of the nine binding pockets of the binding groove of HLA-A*0201 was extracted. The algorithm is based on the assumption that the total binding affinity of a peptide can be approximated by the sum of the binding affinities of its individual amino acids, neglecting the effect of interactions between neighboring residues. Therefore, each amino acid was simulated individually in each binding pocket; initial conformations were constructed from available crystal structures and, in order to stabilize the peptide conformations, the single so-called pivot amino acid was extended by a glycine residue on both sides (for terminating residues on the non-terminating side only) resulting in pseudo-dimers or -trimers which were used in the simulations. For amino acids with no available experimental structures, existing residues were mutated to the corresponding amino acid using the program SCWRL3 (Canutescu et al. 2003). Subsequently, a binding-free-energy-based scoring matrix (BFESM) was constructed which included important energy terms reflecting the binding properties of the amino acids derived from the simulations. Each entry in the matrix represented one feature (energy term) of a particular amino acid in a particular binding pocket.

The BFESM is used to generate a feature vector for each given peptide in the training dataset; all vectors together produce a feature matrix for model generation and prediction. A local feature matrix is constructed from the BFESM which uses all residue and binding pocket positional information from the scoring matrix. This matrix provides a basis for logistic regression and SVM training (Team 2003) of the final model (*DynaPred*^{POS}).

One feature unique to *DynaPred*^{POS} is the ability to construct bound peptide conformations for all predicted sequences. The bound conformations are generated by connecting the saved residue conformations for the simulation runs and performing a short energy minimization. In a detailed analysis (Antes et al. 2006), the constructed peptide structures were refined within seconds to structures with an average backbone RMSD of 1.53 Å from the corresponding experimental structure.

Additionally, we evaluated two sequence-based prediction methods from the literature. The first method is *SVMHC* from Dönnes et al. (Dönnes and Elofsson 2002), which is based on SVMs and was implemented using the software package SVM-LIGHT (Joachims 1999). For this method, SVM kernels and trade offs were optimized by systematic variation of the parameters and evaluation of prediction performance was made using Matthews Correlation coefficients (Matthews 1975), which were used as the main measure of performance for parameter optimization. The second sequence-based prediction method we evaluated is *YKW* from Yu et al. (Yu et al. 2002), which is based on data-derived matrices. The matrix is generated using logarithmized propensities for occurrence in binding vs.

nonbinding peptides of amino acids at specific positions within the peptide training set to generate an initial matrix. The final matrix was derived by a position dependent weighting of the initial matrix which was derived by an analysis of binding data. The SVMHC and YKW methods were re-implemented for this study using the methodology reported in the original publications.

The fourth and final method we evaluated is an artificial neural network based approach (Nielsen et al. 2003) (Buus et al. 2003), which was developed using ANN which are capable of performing sensitive, quantitative predictions. Such quantitative ANN were shown to be superior to conventional classification ANNs which have been trained to predict binding versus non-binding peptides. *NetMHC* has recently been shown to be among the best predictors in an extensive comparison of prediction servers whose performance was evaluated with 176 peptides derived from the tumor antigen survivin and the cytomegalovirus internal matrix protein pp65 (Lin et al. 2008). *NetMHC* is available via http://www.cbs.dtu.dk/services/NetMHC/. *NetMHC* could not be trained for this study as it was only accessible via with web interface, and was therefore used for testing purposes only. Also, it is probable, that at least some of the data used to train the *NetMHC* server was the same data which was retrieved from IEDB for this study.

For Datasets F, I and S, training and testing of the prediction models for *SVMHC*, *YKW*, and *DynaPred*^{POS} was performed for each HLA-A and HLA-B allele separately. In the case of *DynaPred*^{POS}, the same BFESM generated from the molecular simulations on HLA-A*0201 was used to generate each new feature matrix for each allele separately. For *NetMHC*, the peptide sequences from Dataset F, I and S were submitted to the prediction server and the prediction results were recorded.

The accuracy of the methods was assessed by generating areas under the curve (AUC, see ROC analysis (Fawcett 2004)), which is a widely used non-parametric performance measure. ROC analysis tests the ability of models to separate binders from non-binders without the need of selecting a threshold. The values AUC $\geq$ 0.90 indicate excellent, 0.90 > AUC $\geq$ 0.80 good, 0.80 > AUC $\geq$ 0.70 marginal and 0.70 > AUC poor predictions (Swets 1988).

We used 10-fold cross validation to assess the accuracy of the predictions.

5.2.3 Datasets excluding anchor positions

Additionally, datasets were generated from the complete peptide datasets F, I, and S from several alleles which excluded the primary anchor positions in the 9-mer. In these new datasets, which we called $F_{1,3\text{-}8}$, $I_{1,3\text{-}8}$ and $S_{1,3\text{-}8}$, the primary anchor positions P2 and P9 were replaced by glycine residues. Training and testing of the prediction models *YKW*, *SVMHC* and *DynaPred*^{POS} was performed for each HLA-A and HLA-B allele separately. *NetMHC* was not included in this analysis, because we were only able to access the predictor via the online server and were thus unable to retrain the statistical model.

5.2.4 Robustness

In order to determine how dependent the reproducibility of the results of the prediction methods *YKW*, *SVMHC* and *DynaPred*^{POS} are on the size of the available allele datasets (a phenomenon that we call robustness), we tested the methods' performance with randomly selected balanced datasets of different sizes, selected from all peptides available in IEDB for a particular allele (Dataset F). *NetMHC* was not included in this analysis, because we were unable to retrain the statistical model.

The alleles examined were HLA-A*0201, HLA-A*3101 and HLA-B*0702. The training was performed on each allele separately, followed by testing using 10-fold cross validation. The smallest balanced dataset for each allele consisted of 50 randomly selected binders and 50 randomly selected non-binders and the size of the largest dataset depended on the overall number of binders or non-binders available for the allele. All prediction methods were run on four randomly selected balanced datasets in each size category for each allele.

5.2.5 Generalizability

By this test, we assess the ability of the statistical models (*YKW*, *SVMHC* and *DynaPred*^{POS}), trained for the allele HLA-A*0201, to generalize to other alleles. *NetMHC* was again not included in this analysis, because we were unable to retrain the statistical model. Training was performed on Dataset F of HLA-A*0201, followed by testing on Datasets F of all other alleles. This generalization ability is essential for epitope prediction models as there are many alleles with insufficient data for training an allele-specific model.

5.3 Results

5.3.1 Datasets with complete peptides

In this section we examine the dependency of the performance of the four prediction methods on the selection criteria of the used training dataset.

If all available peptides for an allele are used for the prediction (Table 5.1, Dataset F), *NetMHC* performs particularly well, achieving the highest AUC for all 24 alleles examined. All four methods had a predictive performance of good or excellent for 20 or more of the 24 alleles. *NetMHC* significantly outperforms the three other methods (Wilcoxon Rank Sum Test, P-value < 0.001) and no statistically significant difference between the other three methods could be detected. Therefore, the ranking of the methods can be described as *NetMHC* > (*DynaPred*^{POS}, *SVMHC*, *YKW*).

	Allele Name	Binders	Non-Binders	Total	DynaPredPOS	NetMHC	SVMHC	YKW
		(n)	*(n)*	*(n)*	*AUC*	*AUC*	*AUC*	*AUC*
1	A*0101	163	1316	1479	0.93	0.98	0.95	0.94
2	A*0201	1544	1929	3473	0.93	0.96	0.92	0.91
3	A*0202	723	697	1420	0.88	0.93	0.85	0.85
4	A*0203	732	685	1417	0.88	0.95	0.86	0.84
5	A*0206	633	782	1415	0.88	0.95	0.87	0.86
6	A*0301	637	1618	2255	0.89	0.96	0.88	0.8
7	A*1101	816	1279	2095	0.92	0.96	0.9	0.91
8	A*2402	202	464	666	0.8	0.85	0.78	0.81
9	A*2601	69	885	954	0.84	0.93	0.83	0.84
10	A*3101	510	1480	1990	0.89	0.95	0.89	0.88
11	A*3301	203	994	1197	0.88	0.96	0.89	0.88
12	A*6801	578	620	1198	0.85	0.92	0.82	0.81
13	A*6802	439	980	1419	0.84	0.93	0.85	0.83
14	B*0702	238	1110	1348	0.94	0.98	0.94	0.93
15	B*0801	23	687	710	0.82	0.99	0.78	0.79
16	B*1501	182	836	1018	0.86	0.97	0.89	0.9
17	B*2705	81	917	998	0.93	0.97	0.9	0.94
18	B*3501	273	578	851	0.83	0.93	0.84	0.85
19	B*4001	94	1112	1206	0.9	0.97	0.93	0.91
20	B*4402	76	136	212	0.77	0.84	0.75	0.76
21	B*4403	71	142	213	0.68	0.81	0.65	0.7
22	B*5101	108	249	357	0.82	0.93	0.8	0.81
23	B*5301	127	228	355	0.86	0.95	0.85	0.87
24	B*5801	78	893	971	0.9	0.99	0.93	0.93
	Average AUC				**0.86**	**0.94**	**0.86**	**0.86**

Table 5.1 Prediction accuracies for the full dataset

Alleles included in the study with the number of binders and non-binders available; all available binders and non-binders were included in the analysis irrespective of whether quantitative laboratory test data was available or not (Dataset F). Only unique peptide sequences were included in the counts; all peptides with more than one entry for a particular allele in IEDB were counted once only. The overall performance of the three prediction models on different alleles is shown; AUC = area under the curve (ROC analysis). The average AUC for each method is included at the bottom of each column.

The results are dependent on the size of the datasets: for good results (AUC greater than 0.85), an allele's dataset generally has to contain more than 100 binders and more than 100 non-binders (preferably more than 200 binders and more than 200 non-binders). Also, datasets for which the number of binders and non-binders are relatively balanced produced larger AUCs (i.e. better performance). Unbalanced datasets in IEDB generally have a substantially lower number of binders than non-binders. For *YKW*, *SVMHC* and *DynaPred*^{POS} better results were achieved for HLA-A than HLA-B. This probably is due to the lower number of epitopes which are available for HLA-B in IEDB and, in the case of *DynaPred*^{POS}, due to the fact that the BFESM was generated using HLA-A*0201 simulation results. *NetMHC* however achieved comparable results for both HLA-A than HLA-B.

Intermediate binders (Table 5.2, Dataset I) were difficult to classify. *NetMHC* had the best performance for 10 out of 11 alleles. However, all methods showed at best marginal prediction performance (the largest achieved AUC was 0.79) and in most cases the predictions were poor.

	Allele Name	Binders	Non-Binders	Total	DynaPredPOS	NetMHC	SVMHC	YKW
		(n)	*(n)*	*(n)*	*AUC*	*AUC*	*AUC*	*AUC*
1	A*0201	616	135	751	0.67	0.77	0.65	0.66
2	A*0202	286	87	373	0.53	0.7	0.41	0.54
3	A*0203	261	126	387	0.58	0.79	0.58	0.6
4	A*0206	264	74	338	0.56	0.73	0.57	0.57
5	A*0301	335	106	441	0.58	0.7	0.6	0.59
6	A*1101	374	91	465	0.56	0.69	0.62	0.63
7	A*3101	278	103	381	0.42	0.69	0.48	0.56
8	A*3301	129	72	201	0.62	0.52	0.39	0.63
9	A*6801	273	96	369	0.54	0.68	0.44	0.57
10	A*6802	227	123	350	0.52	0.74	0.5	0.5
11	B*1501	169	33	202	0.59	0.79	0.49	0.59
	Average AUC				**0.56**	**0.71**	**0.52**	**0.59**

Table 5.2 Prediction accuracies for the dataset containing only weak binders and non-binders

Results from Dataset I, in which only weak binders (50 nM to 500 nM binding affinity) and non-binders (500 nM to 1000 nM binding affinity) were included. Alleles included in Dataset F, which had fewer than 200 binders and non-binders in total in Dataset I, were no longer included in the analysis. The average AUC for each method is included at the bottom of each column.

Restricting the datasets to peptides which were either very strong binders or clear non-binders substantially improved the results in most cases (Table 5.3, Dataset S); with thirteen of fourteen alleles the best method, *NetMHC*, achieved AUC equal to or greater than 0.99. With the exception of allele HLA-A*2402, all methods had an excellent predictive performance (AUC greater than 0.90). Despite a substantially lower number of data points in Dataset S, a higher accuracy was found for the best method for all alleles when compared with the Datasets F. A typical ROC plot comparing the performance of the four prediction methods for Dataset S is shown in Figure 5.1.

	Allele Name	Binders	Non-Binders	Total	DynaPredPOS	NetMHC	SVMHC	YKW
		(n)	*(n)*	*(n)*	*AUC*	*AUC*	*AUC*	*AUC*
1	A*0101	34	284	318	0.96	1.00	0.96	0.97
2	A*0201	549	503	1052	0.97	0.99	0.97	0.95
3	A*0202	290	267	557	0.98	1.00	0.97	0.97
4	A*0203	273	255	528	0.97	0.99	0.96	0.96
5	A*0206	216	371	587	0.97	0.99	0.96	0.95
6	A*0301	123	332	455	0.93	1.00	0.94	0.95
7	A*1101	228	269	497	0.95	1.00	0.94	0.97
8	A*2402	69	272	341	0.88	0.84	0.84	0.85
9	A*2601	15	256	271	0.97	1.00	0.94	0.97
10	A*3101	114	349	463	0.96	1.00	0.96	0.97
11	A*3301	36	620	656	0.95	1.00	0.94	0.93
12	A*6801	155	235	390	0.95	1.00	0.94	0.94
13	A*6802	95	440	535	0.96	1.00	0.97	0.94
14	B*0702	45	161	206	0.95	1.00	0.96	0.93
	Average AUC				**0.95**	**0.99**	**0.95**	**0.95**

Table 5.3 Prediction accuracies for the dataset containing only strong binders and clear non-binders

Results from Dataset S, in which only strong binders (less than 10 nM binding affinity) and very clear non-binders (greater than 10,000 nM binding affinity) were included. Alleles included in Dataset F, which had fewer than 200 binders and non-binders in total in Dataset S, were no longer included in the analysis. The average AUC for each method is included at the bottom of each column.

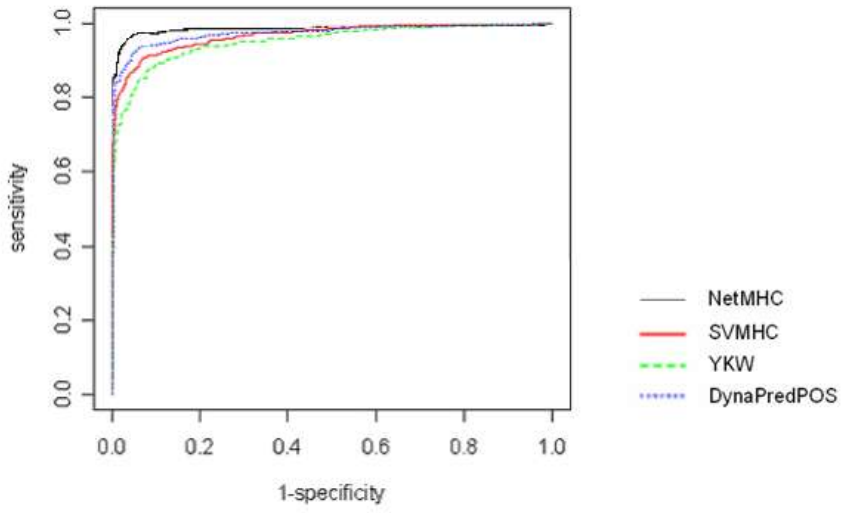

Figure 5.1 Overall performance evaluation

ROC plot for the overall performance evaluation of SVMHC, YKW , DynaPredPOS and *NetMHC* with models that are trained and tested on Dataset S pertaining to allele HLA-A*0201.

Overall, the best performance was achieved in cases where Dataset S was used, the number of binders in the dataset was large (more than 200 binders and more than 200 non-binders), the dataset was relatively well balanced, the *NetMHC* method was used, and the allele was of the HLA-A*02 type.

For the independent dataset of 176 peptides (Table 5.4), while *NetMHC* was the best method for five of the seven alleles tested there was no significant difference in the performance of the methods. For all alleles, with the exception of HLA-A*1101, at least one method had excellent predictive performance (AUC greater than 0.90); generally at least two methods showed excellent predictive performance.

	Allele Name	Binders	Non-Binders	Total	DynaPredPOS	NetMHC	SVMHC	YKW
		(n)	*(n)*	*(n)*	*AUC*	*AUC*	*AUC*	*AUC*
1	A*0201	33	143	176	0.92	0.94	0.82	0.93
2	A*0301	11	165	176	0.77	0.92	0.70	0.85
3	A*1101	17	159	176	0.84	0.89	0.72	0.83
4	A*2402	37	139	176	0.90	0.78	0.90	0.64
5	B*0702	9	167	176	0.86	0.98	0.72	0.68
6	B*0801	10	166	176	0.97	0.92	0.97	0.61
7	B*1501	14	162	176	0.71	0.92	0.71	0.80
	Average AUC				**0.85**	**0.90**	**0.79**	**0.76**

Table 5.4 Prediction accuracies for an independent dataset

A set of 176 novel peptides, generated and tested by Lin et al. (Lin et al. 2008), were used to test the prediction accuracy of the four methods in this study. The average AUC for each method is included at the bottom of each column.

5.3.2 Datasets excluding anchor residues

We examined the possible impact of non-anchor amino acid residues on peptide binding by measuring the prediction accuracy of *YKW*, *SVMHC* and *DynaPred*^{POS} on datasets in which the primary anchor position amino acids P2 and P9 of the peptides were removed and replaced by glycine residues. The best AUC of the performing method is listed for each dataset Table 5.5. If the anchor positions were excluded from the analysis, it was found that the AUCs from Dataset $S_{1,3\text{-}8}$ were generally largest. This is consistent with the results obtained from the datasets containing complete peptides. The predictive performance results were generally good for HLA-A alleles and excellent for HLA-A*02 type alleles, but marginal to poor for Dataset $F_{1,3\text{-}8}$ and Dataset $I_{1,3\text{-}8}$.

Allele Name	Dataset	Binders	Non-Binders	Total	AUC	Best Method(s)
		(n)	*(n)*	*(n)*		
A*0101	$F_{1,3-8}$	161	1312	1473	0.67	SVMHC
	$I_{1,3-8}$	106	35	141	0.66	SVMHC
	$S_{1,3-8}$	34	281	315	0.80	SVMHC
A*0201	$F_{1,3-8}$	1475	1888	3363	0.80	SVMHC
	$I_{1,3-8}$	601	134	735	0.56	YKW
	$S_{1,3-8}$	540	494	1034	0.93	SVMHC
A*0203	$F_{1,3-8}$	701	657	1358	0.81	SVMHC
	$I_{1,3-8}$	252	105	357	0.56	YKW
	$S_{1,3-8}$	270	252	522	0.92	SVMHC
A*1101	$F_{1,3-8}$	841	1275	2116	0.73	SVMHC/YKW
	$I_{1,3-8}$	374	91	465	0.61	SVMHC
	$S_{1,3-8}$	226	267	493	0.82	SVMHC
A*3101	$F_{1,3-8}$	509	1471	1980	0.75	SVMHC
	$I_{1,3-8}$	278	103	381	0.49	YKW
	$S_{1,3-8}$	114	342	456	0.85	SVMHC
B*0702	$F_{1,3-8}$	230	1093	1323	0.73	SVMHC
	$I_{1,3-8}$	122	44	166	0.44	SVMHC
	$S_{1,3-8}$	43	160	203	0.68	YKW
B*3501	$F_{1,3-8}$	257	566	823	0.70	YKW
	$I_{1,3-8}$	131	39	170	0.61	SVMHC
	$S_{1,3-8}$	38	131	169	0.73	YKW

Table 5.5 Prediction accuracies for datasets which excluded anchor positions

The prediction accuracy on datasets F1,3-8, M1,3-8, C1,3-8 and S1,3-8 where the primary anchor-position amino acids P2 and P9 of the peptides were replaced by glycine residues.

5.3.3 Robustness

In this test we examined the dependency of the quality of the obtained prediction models (*YKW*, *SVMHC* and *DynaPred*^{POS}) on the size of the training sets used. *NetMHC* was not included in this analysis as the predictor is only available online and therefore could not be trained by the authors. We found that in most cases the AUCs stabilized at or close to their maximum level, when the size of the randomly selected balanced dataset consisted of more than 200 binders and 200 non-binders (Figure 5.2). This effect was observed with all three prediction methods and for all three alleles included in the study. *SVMHC* performance was less stable for small datasets, which might be due to its simple encoding method and is a significant drawback of this method.

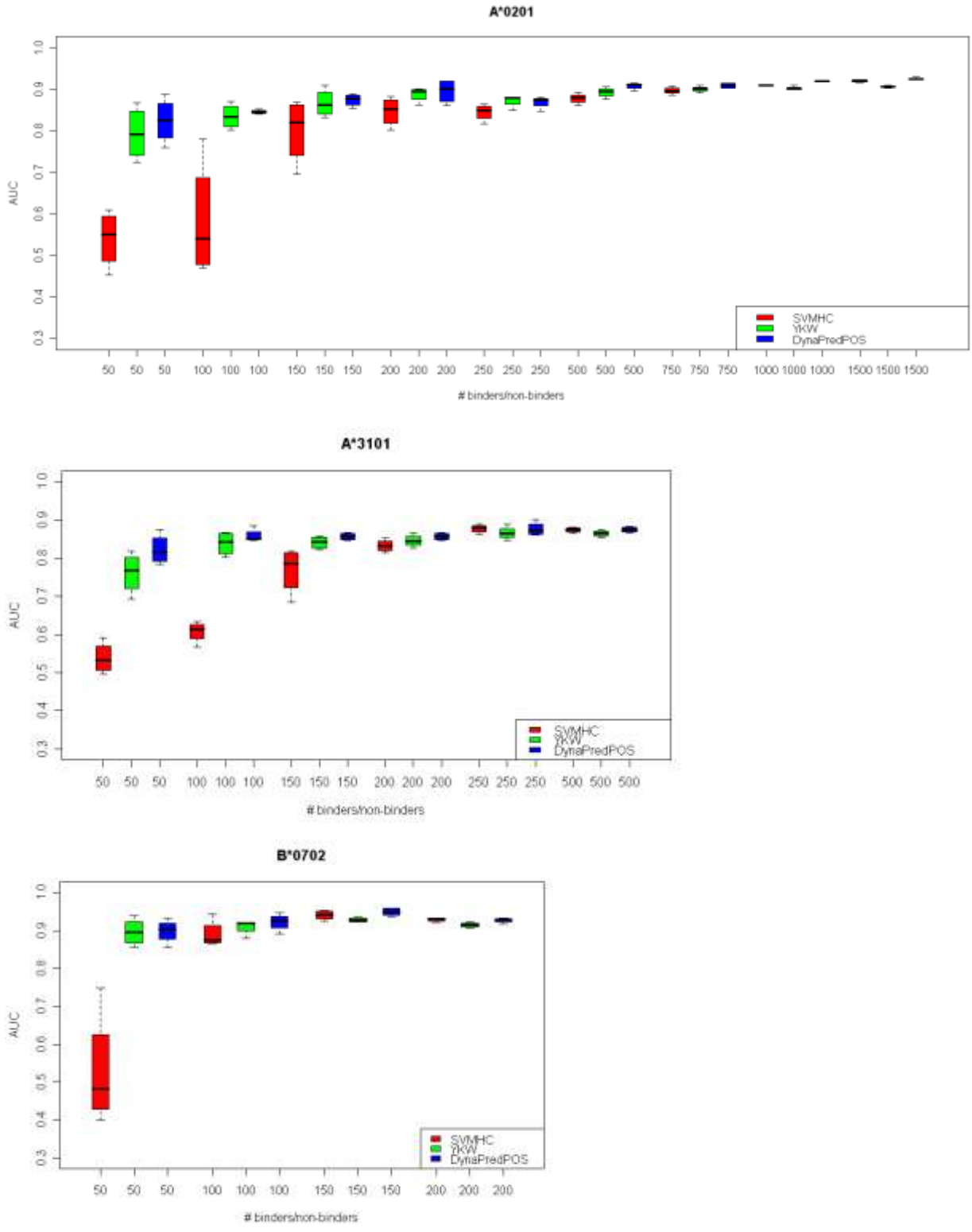

Figure 5.2 Robustness analysis

The reproducibility of the results of the prediction methods and their dependence on the size of the available dataset was examined in selected alleles. Box plots of randomly selected balanced sets of binders and non-binders from Dataset F for the alleles HLA-A*0201, HLA-A*3101, and HLA-B*0702 are shown. The smallest dataset for each allele consisted of 50 binders and 50 non-binders. The size of the largest dataset for each allele depends on the total number of binders or non-binders available for that particular allele. *NetMHC* was not included in this analysis as the predictor is only available online and could therefore not be trained by the authors.

5.3.4 Generalizability

Last, we evaluated the generalizability of *YKW*, *SVMHC* and *DynaPred*^{POS} on all HLA-A and HLA-B alleles for Datasets F. *NetMHC* was again not included in this test because the model could not be trained by the authors. In Figure 5.3 the AUCs for the models trained on the HLA-A*0201 dataset is given for different alleles. It can be seen that *DynaPred*^{POS} outperforms the other models for alleles of the HLA-A*02 type, but for the other alleles the performance of the three methods is very similar. The prediction capabilities are good to marginal for some alleles implying that cross-allele prediction is feasible in some cases.

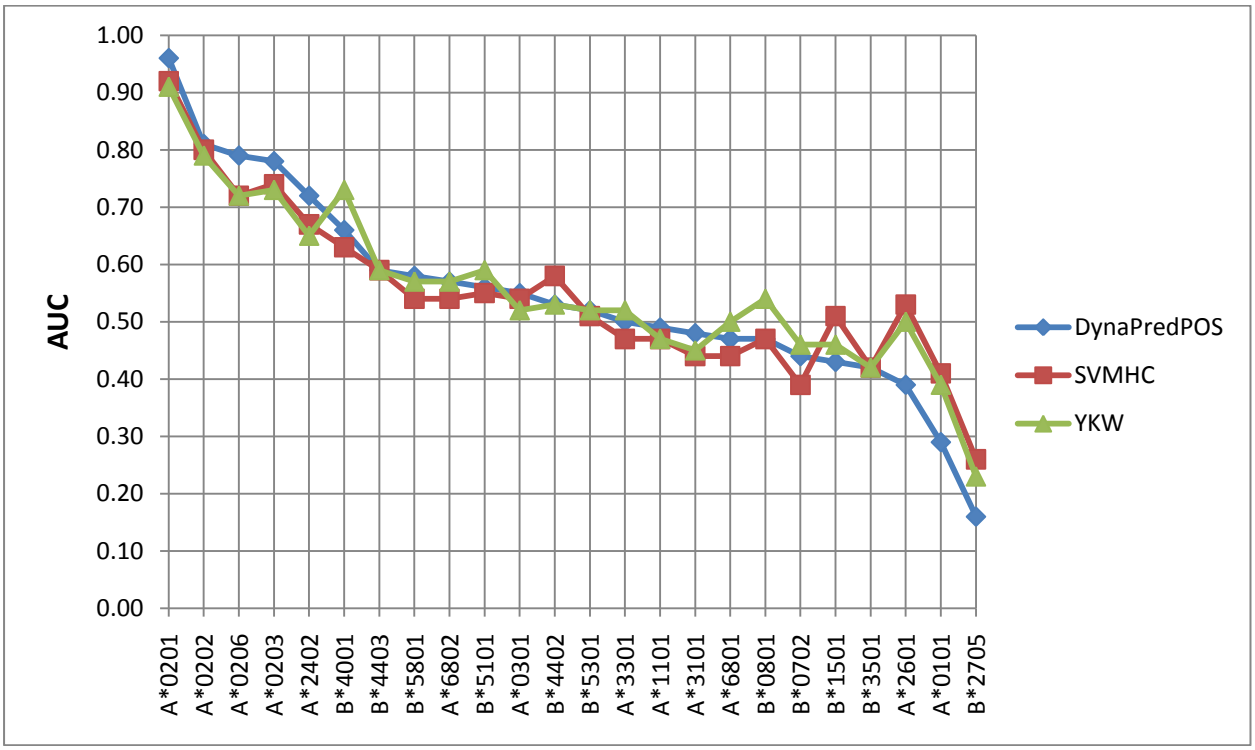

Figure 5.3 Performance Comparison on Dataset F

The performance of the three prediction models, trained on Dataset F of HLA-A*0201 and tested on Dataset F of all alleles (AUCs). *NetMHC* was not included in this analysis as the predictor is only available online and therefore could not be trained by the authors.

The MHC supertype classifications schemes generate clustered sets of molecules with largely overlapping peptide repertoires (Reche and Reinherz 2007; Sidney et al. 2008). These classification schemes generally depend on features such as published motifs, binding data and the analysis of shared repertoires of binding peptides, etc. There has been interest in the development of pan-specific algorithms that can predict peptide binding to alleles for which limited or even no experimental data is available. This would, in contrast to the typical supertype classification scheme which depends on the availability to such data, allow for the prediction of binding in cases where no such data is available.

In recent work by Zhang et al. (Zhang et al. 2009), their predictor was trained on a dataset of binders and non-binders for a wide variety of alleles, and tested on a second set of datasets which consisted of binders and non-binders which were released at later point in time. While the study claims to analyze performance on alleles for which no or only limited data is available, these alleles are never identified and only very limited information on the

results is given. Also, the performance on alleles for which no data was available for training was poor. In contrast, we have performed a detailed analysis of the performance of predictors trained on one allele, and their ability to accurately predict other alleles.

There have been several papers defining supertypes using a number of different approaches (Doytchinova et al. 2004; Sidney et al. 2008). Generally HLA-A*02 alleles and HLA-A*24 alleles are not clustered in the same supertype. Our generalizability work was however able to make reasonable predictions for HLA-A*2402 (the only allele in our study of this supertype), using a predictor trained on allele HLA-A*0201.

5.4 Discussion

Until the creation of IEDB, the available peptide sequence datasets were small and spread over many separate efforts. In addition the datasets consisted predominantly of binding sequences, so that most prediction models based on these data used random non-binding data for training purposes. Through the IEDB database a sufficiently large number of experimentally verified non-binders have become available for learning for the first time. Therefore, the prediction models evaluated here could not only be tested on a substantially larger number of binders, but in addition experimentally verified non-binders could be included in the training datasets and alleles that have previously not been analyzed due to insufficient dataset sizes could be included in the study.

As expected, Dataset S, which consisted only of peptides for which a quantitative laboratory result was available in IEDB and which were either strong binders or clear non-binders, performed better than Dataset F. As the binding affinity at which a binder becomes a non-binder has a threshold of 500 nM in IEDB, removing all peptides from the dataset which we described as so-called intermediate binders (50 nM to 1000 nM) improved the performance of the methods (results not shown), as did using a subset of data containing only the very strongest binders and clearest non-binders. Due to the error involved in experimental binding affinity analysis (Kessler et al. 2003), we suggest using a cutoff of 500 nM may incorrectly categorize a weak binder as a non-binder or vice versa. Perhaps adding a category containing such intermediate binders, in addition to the already existing categories binder and non-binder, would be a useful addition to IEDB.

The excellent performance of *NetMHC*, on Dataset S in particular where it performs with an AUC of 1 for many alleles, may be in part due to the fact that this method could not be trained for the purposes of this study as *NetMHC* was only accessible via a web interface. It should also be noted that some of the data used to train *NetMHC* was probably identical to that extracted from IEDB for this study. This conjecture is also supported by the prediction results of the methods on the independent, novel dataset, which showed no statistically significant results between the methods (Table 5.4).

Peters et al. (Peters et al. 2006) performed an extensive analysis in which they trained and tested their own in-house methods, but external methods were not reimplemented by the authors. Instead, the available web interfaces for external methods were used with the default settings. In contrast to this, we reimplemented the external methods *YKW* and *SVMHC* to allow for both training and testing.

The analysis of the reproducibility of the results of the examined prediction models and their dependence on the size of the available dataset (robustness) showed that all methods require a sufficient number of data points for reproducible results (Figure 5.2). Overall, most alleles appear to require a minimum of 200 binders and 200 non-binders in Dataset F before the AUCs stabilize at or close to their maximum level. We suggest that performing the analysis on alleles with too few data points (which is still unfortunately the case with many HLA-B alleles) can lead to unreliable results.

The analysis of the methods' generalizability showed that the prediction capabilities are good to marginal for some alleles, implying that cross-allele prediction is feasible in some cases. In other cases, the AUCs were very low. Having trained with HLA-A*0201 and then tested for generalizability on HLA-A and HLA-B alleles in Dataset F, a possible reason for certain alleles to give rise to such low AUCs may be that a particular subset of binders that bind well to HLA-A*0201 may be very clear non-binders for the alleles in question. Conversely, clear non-binders for HLA-A*0201 may be binders for the other alleles leading to low AUCs.

In contrast to the former study (Antes et al. 2006), which included only binding sequences (as sufficient numbers of experimentally verified non-binding sequences were not available at that time) in testing generalizability, we observed a much improved level of cross-allele prediction with these newer datasets.

6 Insights into T Cell Receptor Binding[5]

6.1 Introduction

T cell immune responses are dependent on the appropriate and effective processing of peptides from a protein source, the stable binding of the peptide to the MHC molecule and recognition of this complex by the T cell receptor. Of these three steps, the binding of the peptide to the MHC is the most selective event (Yewdell and Bennink 1999). There is a large body of work from the last 20 years, which examines MHC-peptide binding (Lafuente and Reche 2009). Comparisons of methods which predict MHC-peptide binding have shown that the best methods have high prediction accuracies for alleles for which sufficient binding data are available (Lin et al. 2008; Roomp et al. 2010).

However, there is still a significant proportion of stable MHC-peptide complexes which elicit no TCR response. Comparatively little work has been performed on predicting which TCR will bind to MHC-peptide complexes. Simplified representations of amino acids, represented as a string of numbers or bits (where each number or bit represents an amino acid), have been used to study negative selection and TCR cross-reactivity (Detours et al. 1999; Detours and Perelson 1999; Chao et al. 2005). A more recent approach has been developed in which the problem was studied numerically but still in a highly simplified form (Kosmrlj et al. 2008; Kosmrlj et al. 2009); a single short string of amino acids represents the highly variable CDR3 which interacts with the peptide. Hence, interactions between the TCR and MHC are not explicitly considered (Kosmrlj et al. 2010). A comparison of the *in silico* results with those of 18 unspecified crystallized TCR-MHC-peptide complexes from the Protein Data Bank (PDB) (Berman et al. 2000) showed qualitative agreement: the theoretical predictions showed that weakly interacting amino acids on the TCR are enriched, whereas strongly interacting amino acids are present in relative low frequency. Overall it can be said that the current *in silico* approaches for predicting which TCR residues are likely to interact with the MHC-peptide antigen are still highly simplified, but the models provide important insights in to the population dynamics of the T cell repertoire and its interaction with viruses such as HIV-1 (Kosmrlj et al. 2010).

We were interested in gaining a better understanding of the interaction modes of human TCR V regions, based on detailed analyses of the crystal structures of TCRs bound to MHC-peptide antigen complexes. We therefore have (1) studied the interaction of all available TCRs bound to HLA-A*0201 peptide complexes. The rationale for choosing complexes which included HLA-A*0201 stems from the fact that HLA-A2 is the most common HLA phenotype in many ethnic populations and HLA-A*0201 is the most common allele (Player et al. 1996; Shieh et al. 1996). There are more instances in PDB of TCR-MHC-peptide complexes that include HLA-A*0201 than any other allele. Also, finding TCRs which interact with one specific HLA allele reduces the variability of one of the components of the analysis, allowing for more reliable comparisons of the interaction patterns between different complexes. Based on these results (2), we have developed the first rule-based classification methods for predicting the residue interactions of the TCR to the MHC-peptide antigen complex based on interactions found in actual structures of the TCR-MHC-peptide

[5] The work reported in this section was performed in collaboration with Francisco Domingues (MPI for Informatics, Germany). It has been published in *Molecular Immunology* and appears with the journal's permission (Roomp and Domingues 2011). My own contribution to the publication comprises the crystal structure selection, structural alignments, extraction of non-covalent interactions, development and testing of the prediction methods resulting in 4 of the 4 published tables (Tables 6.1, 6.2, 6.3 and 6.4) and 5 of the 5 published figures (Figures 6.1, 6.2, 6.3, 6.4 and 6.5).

complexes. Additionally (3), we have performed a comprehensive evaluation of TCRs known to interact with HLA-A*0201 in cell assay studies, as this type of data is more commonly available than crystal structure data. We were interested in identifying possible biases in the different datasets, i.e., in determining if the available crystal structures are broadly representative of the known TCR types.

6.2 Materials and Methods

6.2.1 Crystal structure selection

One serious obstacle to the study of TCR-MHC-peptide interactions is the small number of solved crystal structures available. In 2005, 24 such structures for both human and mouse were recorded (Rudolph et al. 2006) and mid-2010 just 31 structures were available. It should be noted that this statistic does not contain structures that were superseded by redetermination at higher resolution. However, MHC and TCR complexes with other molecules, such as superantigens or antibodies, were included.

The relatively low number of available structures, when compared to the many TCR V and J genes which exist in both human and mouse, is due to the many technical challenges faced when generating such complexes (Rudolph et al. 2006). Additionally, the type of structures represented are thought to be a biased sample because some types of TCR are difficult to crystallize (Marrack et al. 2008). The TCRα chains appear to show more variability in terms of the structures available than the TCRβ chains.

We initially selected the complexes for this study based on the following criteria: each complex had to consist of the MHC molecule HLA-A*0201, a 9 or 10mer peptide and a T cell receptor. Subsequently, the sequences of the CDR1, CDR2, and CDR3 in both TCR subunits were aligned and only seven structures that differed in the CDR1, CDR2, and CDR3 sequences were retained (Table 6.1). The PDB entries included in the study were 1OGA (Stewart-Jones et al. 2003), 1AO7 (Garboczi et al. 1996), 1BD2 (Ding et al. 1998), 2BNQ (Chen et al. 2005), 1LP9 (Buslepp et al. 2003), 3HG1 (Cole et al. 2009) and 3GSN (Gras et al. 2009). In PDB entries with two co-crystallized complexes, one complex was chosen for the analysis.

PDB ID	Resol-ution	Peptide	Affinity (nM)[a]	TRAV[b]	TRAJ[b]	TRBV[b]	TRBJ[b]
1OGA	1.40	GILGFVFTL	17 (SB)	**27*01** (100.00%)	**42*01** (100.00%)	**19*01** (100.00%)	**2-7*01** (100.00%)
1AO7	2.60	LLFGYPVYV	3 (SB)	**12-2*02** (100.00%)	**24*02** (100.00%)	**6-5*01** (100.00%)	**2-7*01** (100.00%)
1BD2	2.50	LLFGYPVYV	3 (SB)	**29/DV5*01** (98.90%)	**54*01** (89.50%)	**6-5*01** (100.00%)	**2-7*01** (100.00%)
2BNQ	1.70	SLLMWITQV	5 (SB)	**21*01** (100.00%)	**6*01** (100.00%)	**6-5*01** (100.00%)	**2-2*01** (100.00%)
1LP9	2.00	ALWGFFPVL	8 (SB)	**12D-2*01** (100.00%)(mouse)	**50*01** (100.00%)(mouse)	**13-3*01** (100.00%)(mouse)	**2-7*01** (100.00%)(mouse) **2-7*02** (100.00%)(mouse)
3HG1	3.00	ELAGIGILTV	417 (WB)	**12-2*01** (98.90%)	**27*01** (100.00%)	**30*01** (100.00%)	**2-2*01** (100.00%)
3GSN	2.80	NLVPMVATV	35 (SB)	**24*01** (100.00%)	**49*01** (100.00%)	**6-5*01** (100.00%)	**1-2*01** (100.00%)

Table 6.1 The PDB crystal structures selected for the study

[a]The binding affinities were predicted using *NetMHC*: SB is a strong binder, WB is a weak binder. [b]The TCRα and TCRβ subunits V -domains were classified using the IMGT nomenclature (Kaas et al. 2004). The percentages indicate the level of similarity between the structures' V- and C-domains and those of IMGT reference protein sequences.

All peptides have been experimentally verified to bind to HLA-A*0201 and a comprehensive listing of these experiments can be found in the Immune Epitope Database and Analysis Resource (Vita et al. 2010). A variety of assay types have been used in these experiments and the results are often qualitative not quantitative. To allow for a better comparisons of the peptides in the study, we calculated their binding affinity to HLA-A*0201 *in silico* by using *NetMHC* (Buus et al. 2003; Nielsen et al. 2003; Lundegaard et al. 2008), which was found to be among the best predictors for MHC peptide binding (Zhang et al. 2009). Six of seven peptides were found to be strong binders (binding affinity <35 nM), while the single 10mer peptide was a weak binder (417 nM). A number of structures were excluded from this analysis: 2VLR (TCR has same amino acid sequence as 1OGA), 2F54 (TCR is same as 2BNQ, except for a few amino acids in TCRα subunit which are outside of the CDRs), 3D39 (TCR same as 1AO7, with slightly different peptide), 3D3V (TCR same as 1AO7, with slightly different peptide), 3H9S (TCR same as 1AO7, with slightly different peptide), 3GJF (has a TCR-like antibody), 3E3Q (is a high affinity mutant), and 2PYE (*in vitro* enhanced mutant).

To summarize the similarities of the CDRs between the chosen structures: CDR1α, 1AO7 and 3HG1 have the same sequence; CDR2α, all are different; CDR3α, all are different; CDR1β, 3GSN, 1BD2, 2BNQ and 1AO7 have the same sequence; CDR2β, 3GSN, 1BD2, 2BNQ and 1AO7 have the same sequence; CDR3β, all are different; peptides, 1AO7 and 1BD2 have the same sequence.

6.2.2 Structural alignments and the identification of CDRs

Structural alignments of all seven TCRα and TCRβ subunits were generated with MultiProt (Shatsky et al. 2004) using default parameters. These results were confirmed using Matt (Menke et al. 2008). Additionally, HLA-A*0201 amino acid positions were confirmed to be the same in each of the structures to ensure that all comparisons were valid.

For CDR1 and CDR2 these regions were identified and numbered to be consistent with previous studies (Arden et al. 1995; Marrack et al. 2008). For CDR3 IMGT classification was used (Kaas et al. 2004) and the numbering of the amino acids was kept consistent with CDR1 and CDR2.

6.2.3 Interactions analysis

Hydrogen atoms were placed on the experimental structural models using Reduce (Word et al. 1999b) , the non-covalent interactions were identified using Probe (Word et al. 1999a). Interactions between the TCRα chain and the MHC-peptide complex were extracted, as were interactions between the TCRβ chain and the MHC-peptide complex. Hydrogen bonds and van der Waals interactions between (1) two amino acid side chain atoms, (2) two amino acid backbone atoms or (3) an amino acid side chain atom and a backbone atom were treated equally in subsequent analyses.

Alternative representations of the residue interaction data were produced. We generated a table with a detailed listing the contacts and a heat map to highlight the most important residue interactions using the statistical computing software R (Team 2003). We also generated a 2D residue interaction network representation using Cytoscape (Shannon et al. 2003) showing all interactions between TCRα/TCRβ and MHC-peptide. Furthermore, we represented the interactions between residues that occurred in the majority of structures in a 3D molecular graphics visualization using VMD (Humphrey et al. 1996).

6.2.4 Predicting interactions between TCR and MHC-peptide

We have developed two new rule-based classification algorithms to predict interactions between the T cell receptor and the MHC-peptide antigen. The methods are heuristic and consist of hierarchical sets of rules. Each amino acid position in TCR's CDR loops is examined for an interaction (contact) with either the MHC or peptide and a prediction is made based as to whether an interaction has been observed at that position or not. Hydrogen bonds and van der Waals interactions were weighted equally for the purposes of the method. The accuracy of the two methods was assessed by generating areas under the curve (AUC), see ROC analysis (Fawcett 2004). AUC is a widely used non-parametric performance measure. In particular, the performance analysis tested the ability of the methods to accurately separate positions which have no interactions from positions which have interactions. The values AUC $\geq$ 0.90 indicate excellent and 0.90 > AUC $\geq$ 0.80 good prediction performance, with a random predictor having an AUC of 0.5 (Swets 1988). We used leave-one-out cross validation to assess the accuracy of the predictions. The algorithms and accuracy evaluation were implemented in R.

These two methods calculate a score in the range of 0 to 1 at each amino acid position *i* in the TCR's CDR1, CDR2, and CDR3 loops based on the rules described in Figures 6.1 and 6.2. A score of 0 indicates no interaction to the MHC-peptide is predicted at that position, while a score of 1 predicts an interaction. In the cross validation procedure a test structure is selected from the seven structures in the study and compared to the other six structures (called training structures).

The methods were developed in an intuitive manner, based on examining the existing structural data. The methods were kept as simple as possible, parameter optimization was not performed, and overfitting was avoided. In fact the only adjustment made to the original methods, which were constructed prior to actually testing the methods' performance, was to include a threshold for the percentage of contacts in a position, which is described below. The decision to include the threshold was based on the recognition that in many CDR amino acid positions, where the majority of structures show an interaction, diverse amino acids are found. These diverse amino acids have highly differing side chain polarities and charges (acid - base properties) and showed themselves to be impossible to group based on such physical and chemical properties. In the future, when more HLA-MHC-peptide complexes become available i.e. the datasets which can be studied become sufficiently large, statistical learning methods could well be applied to this problem.

Overall the first method (Method I) can be described as follows:

- If a gap exists in the test TCR at position *i*, the amino acid is skipped
- If the overall fraction of training structures with contacts at this position is more than 50%, a score of 1 is given
- If the overall fraction of training structures with contacts at this position is less than or equal to 50%, and if the test TCR has an amino acid at the position of interest (i.e. no gap), then the score is calculated as follows:
    - If the test TCR's amino acid type (T_i) exists in the training structures and at least one such training structure's amino acid has a contact, variable *A* is 1. Otherwise *A* is 0.
    - *A* is multiplied by variable *B*. *B* is the proportion of the amino acids of type T_i in the training structures that have contacts.
    - A penalty is given when only a small number of training structures with the amino acid T_i type have contacts: penalty score *C* of 0.25 is given, where fewer than 25% of amino acids T_i have contact. Otherwise *C* is 0.
    - Score = Max (0, $A*B - C$).

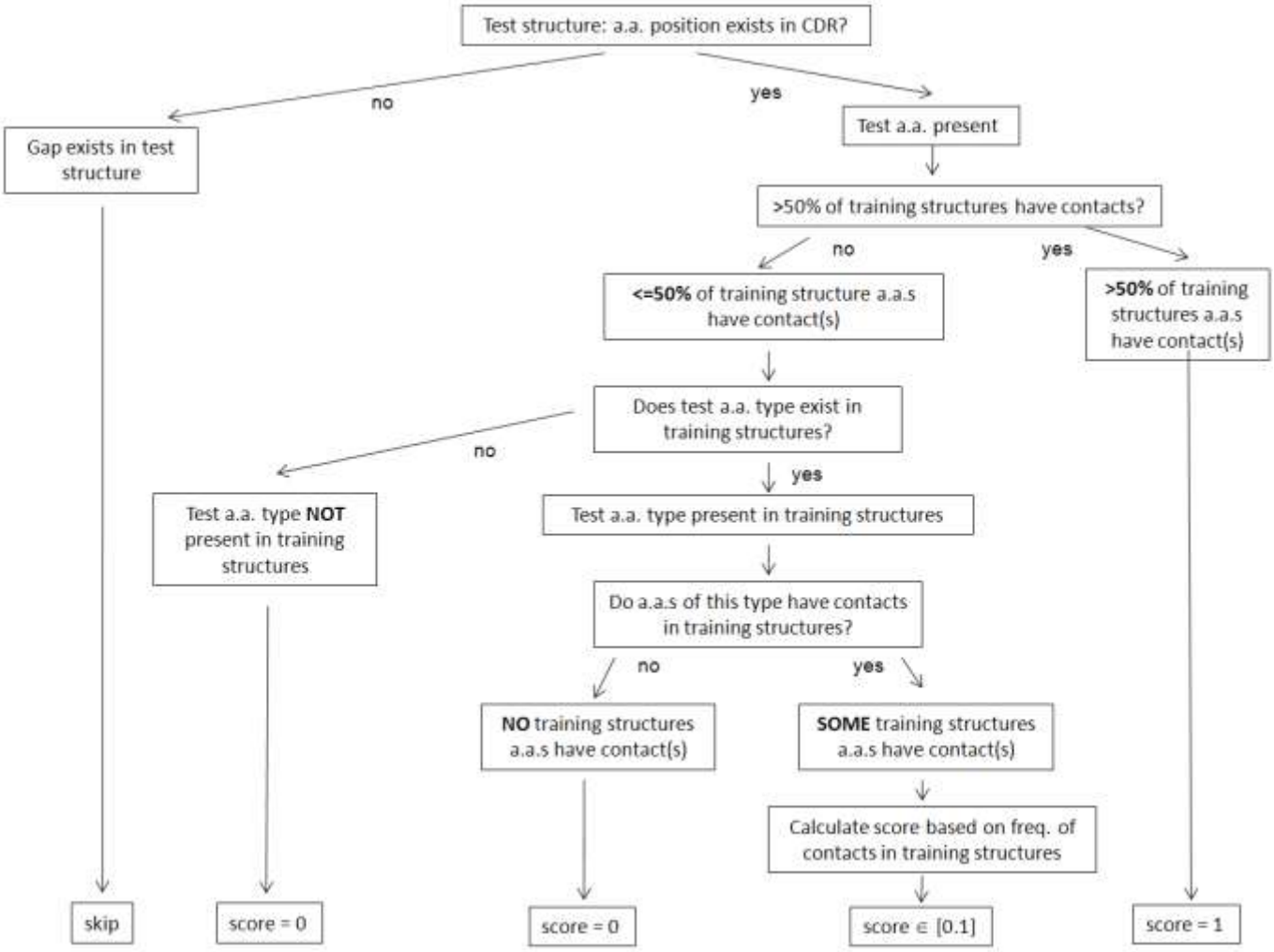

Figure 6.1 First rule-based classification algorithm (Method I) for predicting contacts between the TCR and the MHC-peptide antigen using crystal structures

The second approach (Method II) is a simplification of the first approach, in which the type of amino acid is disregarded (Figure 6.2):

- If a gap exists in the test TCR, the position *i* in the amino acid sequence is skipped
- If the overall number of training structures with contacts at position *i* is more than 50%, a score of 1 is given
- If the overall number of training structures with contacts at position *i* is less than or equal to 50%, and if the test TCR has no gap in *i* then the score was calculated in the following manner:
    - Variable *A* is 1 if some training structures have contacts, else A is 0.
    - *A* is multiplied by variable *B*. *B* is the proportion of amino acids of any type that have contacts.
    - There is a penalty when only a very small number of structures have contacts: penalty score *C* of 0.25 is given, where fewer than 25% of amino acids have a contact. Otherwise *C* is 0.
    - Score = Max(0, *A***B* – *C*).

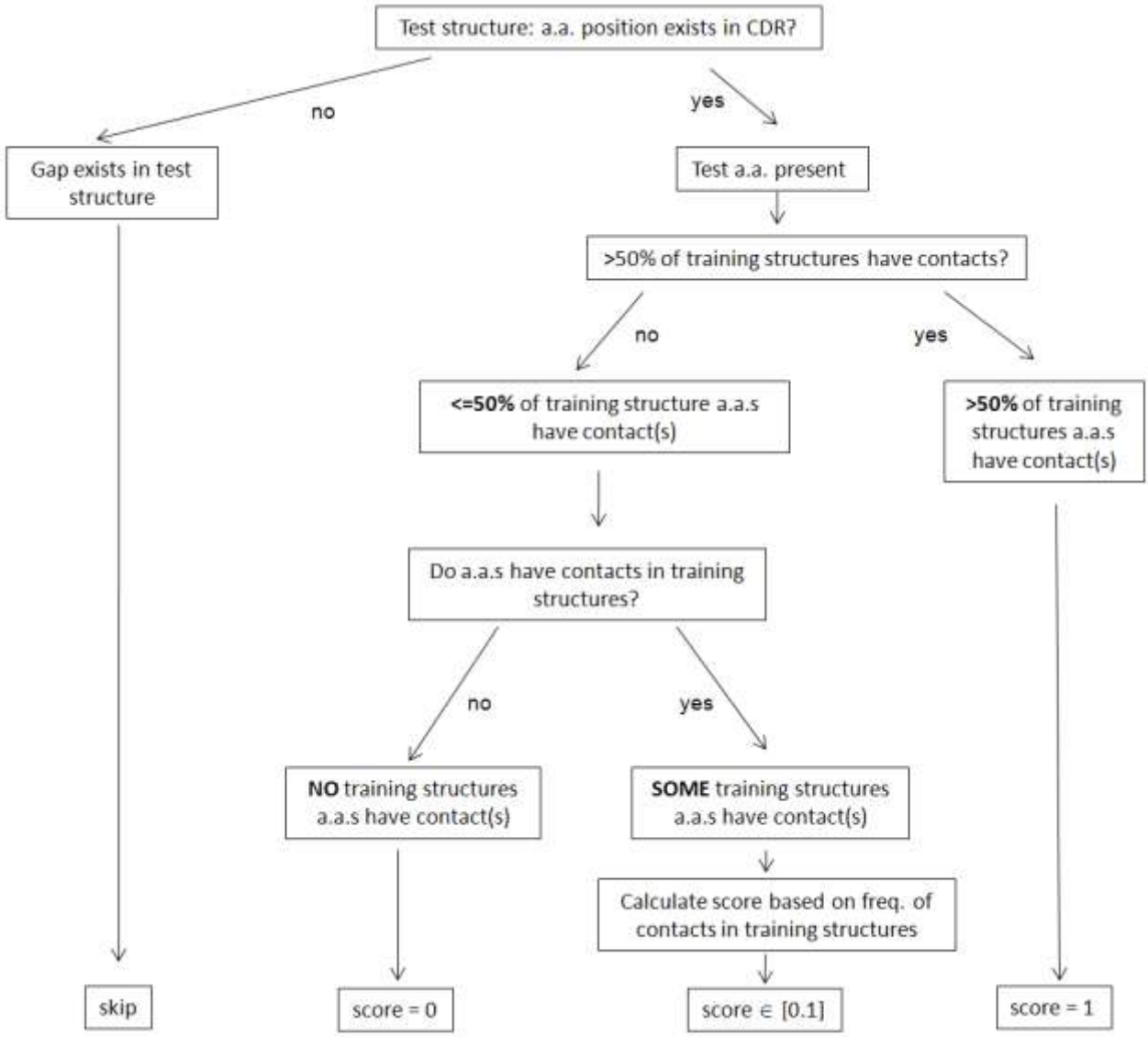

Figure 6.2 Second rule-based classification algorithm (Method II) for predicting contacts between the TCR and the MHC-peptide antigen using crystal structures

6.2.5 Experimentally verified TCRs

Due to the small number of solved crystal structures of a TCR binding to HLA-A*0201 we were interested in determining if the available crystal structures were broadly representative of TCRs identified by other experiments to bind to HLA-A*0201; we searched for sequence data on TCRs which were verified in the laboratory to bind to HLA-A*0201 with high avidity. The T cell clone SK22, specific for the FMNL1-derived peptide PP2, was identified as being such a TCR (Schuster et al. 2007). A further study by the same group (Liang et al. 2010) identified four allo-HLA-A2-restricted T cell clones expressing TCRs with specificity for a peptide derived from the tumor associated antigen (HER2/neu) called HER2-dervied peptide 369 ($HER2_{369}$). Additionally, the clone JG-9 was isolated which recognizes the endogenously processed Ag pp65 (Schub et al. 2009). A further study identified a CTL clone named R6C12 expressing a TCR highly avid for a GP100 derived peptide in an HLA-A201-restricted manner (Morgan et al. 2003). Finally, a study identified a number of CTL clones which bind to an influenza matrix protein antigen in complex with HLA-A2 (Moss et al. 1991).

6.3 Results

6.3.1 Interactions in TCR-MHC-peptide complexes

Six different types of interactions were distinguished in this analysis: (1) a hydrogen bond between two backbone atoms, (2) a hydrogen bond between backbone and side-chain atoms, (3) a hydrogen bond between two side-chain atoms, (4) a van der Waals contact between two backbone atoms, (5) a van der Waals contact between backbone and side-chain atoms, and (6) a van der Waals contact between two side-chain atoms. Different types of interactions are usually observed at each CDR position (Table 6.2). In the case of interactions occurring between the TCR and the peptide this is not unexpected as the peptide sequence varies among the selected structures with the exception of the structures with PDB identifiers 1AO7 and 1BD2.

The only pairs of interacting positions which occur in the majority of structures (greater than or equal to four of seven structures) and where all interactions were of the same type comprised: (1) position 93 in CDR3α interacts with position 5 in the peptide by a van der Waals contact between side chain atoms (five of seven structures), (2) position 95 in CDR3α interacts with position 5 in the peptide by a backbone to side chain van der Waals contact (five of seven structures; structure 3HG1 also forms additional interactions at this position), (3) position 28 in CDR1β interacts with position 8 in the peptide by van der Waals contact between side chains (four of seven structures), and (4) position 46 in CDR2β interacts with position 65 in the MHC by van der Waals contact between side chain atoms (four of seven structures). Interestingly, although these frequently observed interactions were consistently of the same type, there was great variability in amino acid types of the TCR and peptide at these positions. There are a number of further interactions which occur in four or more structures, where the types of interactions are not consistent (Table 6.2).

In the heatmap analysis (Figure 6.3) the highest number of structures which shared an interaction between two residue positions (independent of the interaction type) was five out of seven; this occurred at three positions between the CDR3α loop and the peptide. A number of interactions were also identified which were shared by four structures and these could be found between CDR1α and both the MHC and the peptide, between CDR1β and the peptide, between CDR2β and the MHC, between CDR3α and the peptide, and between CDR3β and both the MHC and peptide. Only CDR2α did not have an interaction with either the MHC or peptide which was shared by a majority of the seven structures analyzed.

TCR aa[a]	HLA or Peptide aa[b]	Total No of Contacts[c]	Type of Contact[d]						
			cnt: mc_mc	cnt: mc_sc	cnt: sc_sc	H-bond: mc_mc	H-bond: mc_sc	H-bond: sc_sc	>=2 cnt[e]
CDR1a27	W167	**3**		1	1				1
CDR1a28	W167	**4**		2	2				
CDR1a30	Pep1	**2**			2				
CDR1a30	Pep2	**2**		1			1		
CDR1a30	Pep4	**3**		1	1		1		
CDR1a30	Pep5	**4**			1			1	2
CDR1a30	Q155	**3**			2				1
CDR1a30	T163	**3**			3				
CDR1a31	Pep5	**3**			2			1	
CDR2a50	H151	**3**			3				
CDR2a50	E154	**3**			2				1
CDR2a50	Q155	**3**			2				1
CDR2a50	A158	**3**			1				2
CDR2a51	A158	**2**			1				1
CDR3a93	Pep5	**5**			5				
CDR3a93	Q155	**2**			2				
CDR3a94	R65	**3**					1	1	1
CDR3a95	Pep4	**5**	2	1		1			1
CDR3a95	Pep5	**5**		4					1
CDR3a95	R65	**2**					1	1	
CDR3a95	K66	**3**		2	1				
CDR3a95a	Pep4	**4**	1	1		2			
CDR3a95a	Pep5	**2**							2
CDR3a96	Pep4	**3**	1				1		1
CDR3a96	R65	**2**					1		1
CDR3a96	K66	**3**		2	1				
CDR3a96a	R65	**2**		1	1				
CDR3a97	Pep4	**2**				1	1		
CDR3a97	Pep5	**4**		2	1				1
CDR3a97	R65	**2**		2					
CDR3a97	A69	**2**			1		1		
CDR1b28	Pep8	**4**						4	
CDR1b28	Q72	**2**						1	1
CDR2b46	R65	**4**			4				
CDR2b48	K68	**2**			1				1
CDR2b48	A69	**2**		1					1
CDR2b48	Q72	**2**			1				1
CDR2b49	Q72	**4**		2	1				1
CDR2b50	Q72	**2**		1					1
CDR2b52	Q72	**3**			3				
CDR2b52	R75	**2**			2				
CDR2b54	R65	**3**			2			1	
CDR3b97	Pep8	**2**			2				
CDR3b98	Pep5	**3**			1				2
CDR3b98	Pep6	**2**		1		1			
CDR3b98	Pep7	**2**		1	1				
CDR3b98	K146	**2**			1				1
CDR3b98	A150	**2**			1		1		
CDR3b98	V152	**2**			1				1
CDR3b99	Pep6	**4**	2	1					1
CDR3b99	Pep7	**3**	2						1
CDR3b99	Pep8	**2**				2			
CDR3b99	T73	**2**		1	1				
CDR3b99a	Pep5	**2**	1	1					
CDR3b99a	Pep7	**4**		2		1			1
CDR3b99b	Pep5	**3**	1	2					
CDR3b99b	Pep7	**4**	1	2			1		
CDR3b99b	A150	**4**		2	1				1
CDR3b99c	Pep5	**3**		2	1				
CDR3b99c	Q155	**4**			1		1	1	1
CDR3b100	A149	**2**		1			1		
CDR3b100	A150	**4**		3					1
CDR3b100	H151	**2**			2				

Table 6.2 Interaction analysis where two or more structures have interactions between residues at the same position

[a]TCRα or TCRβ residues of the CDR1, CDR2 or CDR3. The table is ordered with respect to the amino acid position in the TCR. [b]HLA or peptide residues. [c]Total number of contacts between the TCR and HLA/peptide. [d]Type of contacts. *cnt* is a van der Waals bond contact, *H-bond* is a hydrogen bond, *mc_mc* is a residue interaction between backbone atoms, *mc_sc* is an interaction between a backbone atom and the side-chain atom of another residue, *sc_sc* is an interaction between the side chain atoms in two residues. [e]Cases where there are multiple interactions between two residues of different types. Interactions between the TCR and peptide are highlighted in grey

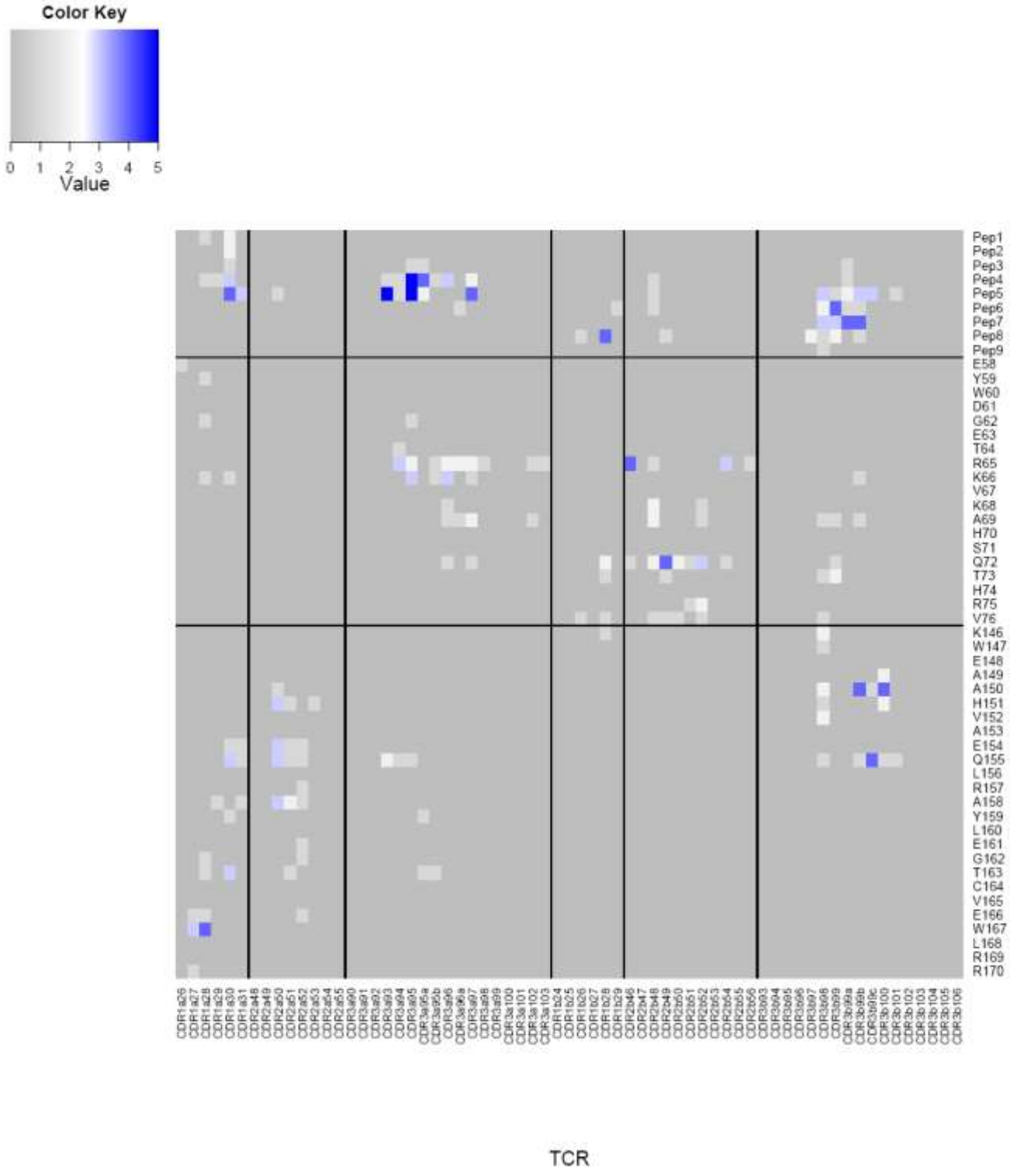

Figure 6.3 Heatmap analysis

Interactions between CDR1, CDR2 and CDR3 of both TCRα and TCRβ (columns) and HLA-A*0201 and peptide (rows) are indicated with regard to their frequency. Black lines indicate discontinuous regions. No interactions were identified in the majority of positions (gray), up to a maximum of 5 interactions (dark blue).

The most common interactions were between the TCR and peptide residue positions 4 and 5. All nine peptide residues were found to be capable of interacting with the TCR in at least one of the structures (Figure 6.4).

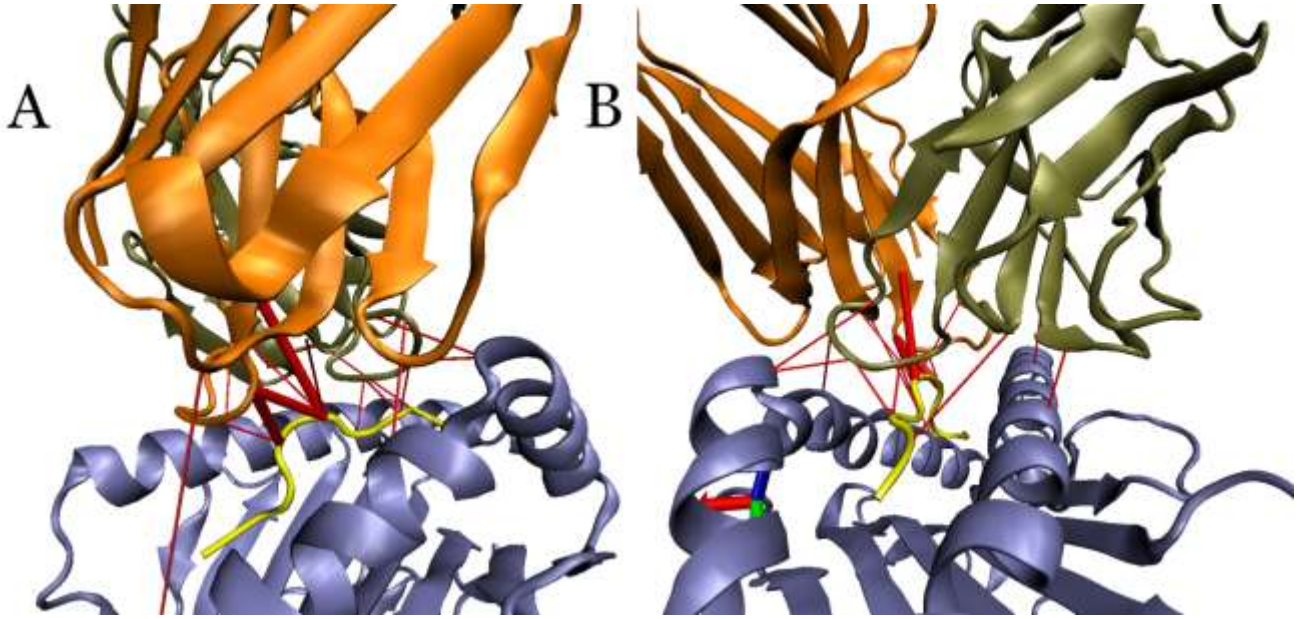

Figure 6.4 Common interactions between TCR and the MHC and peptide

(A) A view of the common interactions where the MHC α-helices and peptide go from left to right, (B) a view after a 90º rotation around the horizontal axis. TCRα (orange), TCRβ (dark yellow), peptide (yellow), MHC (blue), interactions occurring in five of seven structures (thick red cylinders), interactions occurring in four of seven structures (thin red cylinders).

When all interactions between TCRα, TCRβ, the peptide and the MHC are examined (Figure 6.5), the broad variety of interactions between the TCR and MHC-peptide becomes clear. Most residues in all chains are capable of a broad range of interactions, as would be expected considering that the overall interaction between the TCR and the MHC-peptide is a relatively weak one and that cross reactivity is common.

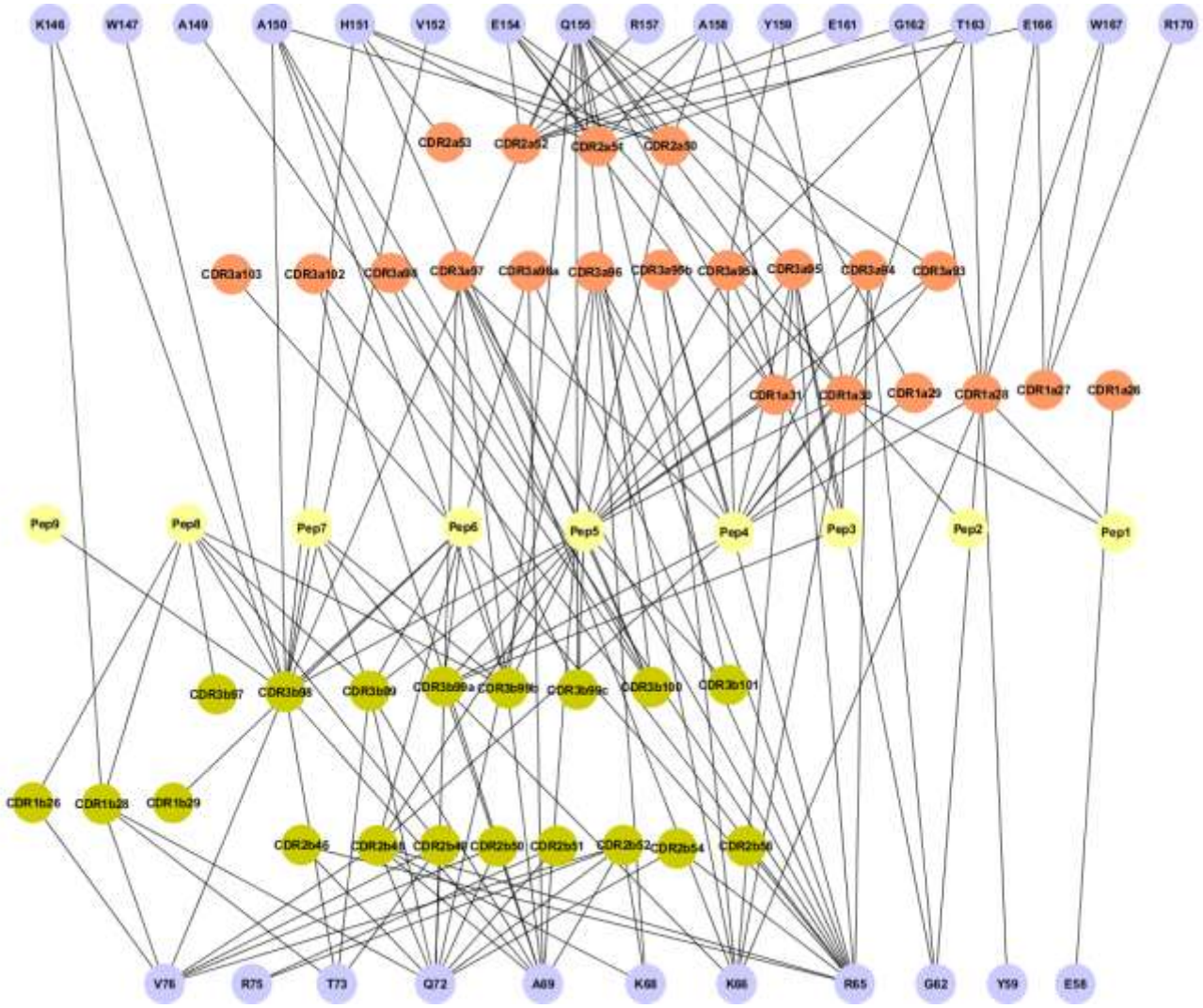

Figure 6.5 All interactions occurring in the structures with PDB identifiers 1OGA, 1AO7, 1BD2, 1LP9, 2BNQ, 3HG1 and 3GSN

All identified interactions (black edges) between TCRα (orange), TCRβ (dark yellow), peptide (yellow) and MHC (blue). CDR1, CDR2 and CDR3 are grouped together. Amino acids which did not participate in interactions are not shown. All edges have the same width, there is no indication of frequency at which each interaction occurs.

6.3.2 Prediction results

Using Method I or Method II, models were generated with six complexes and tested with a seventh complex not included in the training set (leave-one-out cross validation). Area under the curve (AUC, ROC analysis) values were computed for both methods for each case (Table 6.3).

Method II classified more true positives correctly, but also had a somewhat increased rate of generating false positives. Method I was able to make good to excellent predictions in 6 or 7 cases (only one AUC was less than 0.80), while the Method II was able to make good to excellent predictions in all cases. No statistically significant difference between the AUCs produced by the two methods could be found based on the two sample Wilcoxon test, although the second method produced a higher AUC in 6 of 7 cases.

PDB ID	Method I (AUC)[a]	Method II (AUC)[a]
1ao7	0.84	0.86
1bd2	0.76	0.83
1lp9	0.82	0.90
1oga	0.85	0.90
2bnq	0.83	0.87
3gsn	0.92	0.94
3hg1	0.89	0.88
Average AUC	**0.84**	**0.88**

Table 6.3 Assessment of prediction accuracies of the rules-based algorithms

[a]AUC = area under the curve (ROC analysis). The average AUC for each method is included at the bottom of each column

6.3.3 TCRs which are known to bind to HLA-A*0201

In addition to the solved crystal structures selected for this study, we have identified studies describing further TCRs which have been shown to experimentally bind to HLA-A*0201 with high avidity (Table 6.4). Solved crystal structures show a bias towards TRBV6-5, TCRs binding to influenza matrix peptide show a bias towards TRAV27/TRBV19 and the remaining TCRs show a bias towards TRBV12.

PDB ID or [clone name]	Peptide	Affinity (nM)[a]	TRAV[b]	TRAJ[b]	TRBV[b]	TRBJ[b]	Reference
1OGA	GILGFVFTL	17 (SB)	27*01	42*01	19*01	2-7*01	(Stewart-Jones et al., 2003)
1AO7	LLFGYPVYV	3 (SB)	12-2*02	24*02	6-5*01	2-7*01	(Garboczi et al., 1996)
1BD2	LLFGYPVYV	3 (SB)	29/DV5*01	54*01	6-5*01	2-7*01	(Ding et al, 1998)
2BNQ	SLLMWITQV	5 (SB)	21*01	6*01	6-5*01	2-2*01	(Chen et al., 2005)
1LP9	ALWGFFPVL	8 (SB)	12D-2*01	50*01	13-3*01	2-7*01/2-7*02	(Buslepp et al., 2003)
3HG1	ELAGIGILTV	417 (WB)	12-2*01	27*01	30*01/30*02	2-2*01	(Cole et al., 2009)
3GSN	NLVPMVATV	35 (SB)	24*01	49*01	6-5*01	1-2*01	(Gras et al., 2009)
[HER2-1]	KIFGSLAFL	24 (SB)	19*01	24*02	12-3*01	2-3*01	(Liang et al., 2010)
[HER2-2]	KIFGSLAFL	24 (SB)	27*01	20*01	12-3*01	2-7*01	(Liang et al., 2010)
[HER2-3]	KIFGSLAFL	24 (SB)	38-1*01	28*01	7-8*01	2-7*01	(Liang et al., 2010)
[HER2-4]	KIFGSLAFL	24 (SB)	21*01	20*01	12-3*01	2-1*01	(Liang et al., 2010)
[R6C12]	ITDQVPFSV	128 (WB)	41*01	54*01	12-3*01	2-1*01	(Morgan et al., 2003)
[JG-9]	NLVPMVATV	35 (SB)	35*02	50*01	12-4*01	1-2*01	(Schub et al., 2009)
[SK22]	RLPERMTTL	85 (WB)	38-2/DV8*01	41*01	27*01	2-5*01	(Schuster et al., 2007)
[1a11b]	GILGFVFTL	17 (SB)	27*01	n/a	19*01	1-1*01	(Moss et al, 1991)
[1a11a]	GILGFVFTL	17 (SB)	27*01	n/a	19*01	1-1*01	(Moss et al, 1991)
[1a8]	GILGFVFTL	17 (SB)	27*01	n/a	19*01	2-7*01	(Moss et al, 1991)
[1a6]	GILGFVFTL	17 (SB)	6*01	n/a	19*01	2-1*01	(Moss et al, 1991)
[B1a]	GILGFVFTL	17 (SB)	27*01	n/a	19*01	2-5*01	(Moss et al, 1991)
[B1b]	GILGFVFTL	17 (SB)	13-1*01	n/a	19*01	2-7*01	(Moss et al, 1991)
[B1c]	GILGFVFTL	17 (SB)	13-1*01	n/a	19*01	2-2*01	(Moss et al, 1991)
[B1d]	GILGFVFTL	17 (SB)	14/DV4*01	n/a	19*01	2-3*01	(Moss et al, 1991)
[B1e]	GILGFVFTL	17 (SB)	27*01	n/a	28*01	2-7*01	(Moss et al, 1991)

Table 6.4 Variable and joining region family affiliation of HLA-A*0201 restricted TCRs

[a]Binding affinities were predicted using *NetMHC*: SB is a strong binder, WB is a weak binder. [b]TCRα and TCRβ subunits V- and J-domains were classified using the IMGT nomenclature (Kaas et al., 2004)

6.4 Discussion

We present two rule-based methods for predicting interactions between TCRs' CDR1, CDR2 and CDR3 loops and the MHC-peptide antigen, which are simple, transparent, and effective. Both methods show good prediction accuracy with an average AUC of greater than 0.80 (0.84 for Method 1, 0.88 for Method 2).

There were no other algorithms to which we could compare our method directly. The only other related approach we could find was by developed by Kosmrlj et al. (Kosmrlj et al. 2008; Kosmrlj et al. 2009; Kosmrlj et al. 2010). In this previous work the problem was studied in a highly simplified from. The TCR and MHC-peptide complex are modeled numerically as strings of amino acids which are typically 10 amino acids in length. The strings indicate the amino acids at the interface between the TCR and MHC-peptide complex and the model assumes that only one site on the TCR interacts with a corresponding site which consists solely of the bound peptide. The analysis is restricted to the TCR's CDR3 and interactions between the TCR and MHC are not explicitly considered. This method therefore cannot be used to make predictions concerning the individual non-covalent interactions in the interface between the TCR CDR loops and the MHC-peptide antigen.

The non-covalent interactions at protein-protein interaction sites are comprised of van der Waals interactions, hydrogen bonds, and salt bridges (Reichmann et al. 2007; Garcia et al. 2009). The interaction surfaces of both the TCR and MHC-peptide complex are large: approximately thousand square angstroms are buried when an interaction event occurs and most of the contacts are mediated by van der Waals interactions (Rudolph et al. 2006). Although van der Waals interactions are weak, a large number of interactions over a large area add up to substantial binding energies. Both hydrogen bonds and salt bridges, which are comparatively rare in TCR and MHC-peptide protein-protein interaction interfaces, add specificity because of their dependence on the stereochemistry and geometry of the bond (Garcia et al. 2009).

The most common interactions between the TCR and MHC-peptide complex were found between the TCRα CDR3 and amino acid positions 4 and 5 of the peptide (Figure 6.4), where 5 of 7 structures were found to have such interactions. But common interactions between both TCR subunits and the MHC, as well as TCRβ and the peptide, where 4 of 7 structures had such interactions, were also found. Also, all peptide residue positions were capable of forming contacts with the TCR.

These results are different from those of a previous detailed study (Rudolph et al. 2006), where the interaction between the peptide component of the MHC-peptide antigen and the TCR was examined. In that study, the number of contacts per peptide residue with the TCR were evaluated in a number of structures where the MHC was HLA-A*0201. The most common interactions in complexes which contained HLA-A*0201 were observed between the TCR and peptide residue positions 5, 7 and 8. Additionally, peptide residue positions 2 and 9 had no observed contacts with the TCR at all. In contrast, our analysis showed that interactions between the TCR and peptide position 4 are also very common (five of seven structures). And while Rudolph et al. identified no interactions between TCR and peptide residue positions 2 and 9, we identified such interactions. In fact, we show that all nine peptide residue positions are capable of interacting with the TCR in at least one of the structures. It remains unknown whether the same applies to 10mers which bind to HLA-A*0201 as there is currently insufficient data to make such an analysis.

Key positions, also called anchor residues, with which the peptide binds HLA-A*0201, are located at positions 2 and the C-terminus (position 9) of the peptide and tend to be hydrophobic (Falk et al. 1991). Furthermore, secondary anchor residues have been found to make an important contribution to the binding between the peptide and HLA-A*0201 and occur at positions 1, 3 and 7 (Ruppert et al. 1993). Typically, primary and secondary anchor

residues are located within binding pockets of the HLA-A*0201 molecule. Therefore, the prominent role of peptide residues 4 and 5 in the interaction with the TCR is consistent with the occurrence of primary and secondary anchor residues at other locations within the peptide which are devoted to MHC-peptide binding. Put another way, the peptide residues which are important for interaction with the MHC are located in MHC binding pockets and therefore less available to interact with the TCR. Therefore other non-anchor peptide residues, such as residues 4 and 5, should be more available to interact with the TCR, which appears to be the case.

Our analysis of complexes consisting of TCR bound to HLA-A*0201 also produced quite different results from those proposed evolutionarily conserved interactions found in a previous study (Marrack et al. 2008): (1) Marrack et al.'s proposed evolutionarily conserved contacts between CDR1α amino acids positions 29 and 31 and MHC class I α2 domain's amino acid in position 158 were only observed in one of seven structures examined in our study, (2) the proposed evolutionarily conserved contacts between CDR2α positions 50 and 51 and MHC class I α2 domain's in position 158 were only observed in three and two structures of seven structures, respectively, (3) the proposed evolutionarily conserved contacts between CDR1β positions 28 and 29 and MHC class I α1/α in position 69 were not seen in any structure in our study and (4) the proposed evolutionarily conserved contacts between CDR2β positions 46 and 48 and MHC class I α1/α position 69 were observed in zero and two of seven structures, respectively, in our study.

We were interested in determining if the available crystal structures are broadly representative of the allorestricted TCRs identified by other experiment types. In functional cell assays, CTLs expressing TCRs which are HLA-A-02 restricted were observed to have a high expression of the specific TRBV12 chains in CD4 and CD8 populations (Liang et al. 2010). The four clones generated for the study by Liang et al. all possessed CDR3 regions differing both in sequence and length, despite being selected to bind to the $HER2_{369}$ peptide. Random combinations of TCRα and TCRβ chains were also tested and one TCRα chain was identified (HER2-2α) which was capable of peptide recognition when paired with any of the TRBV12-derived TCRβ chains (which also possessed different CDR3 regions). The authors suspect a dominant role of the TCRα chain in the selection of the preimmune TCR repertoire and HLA-A*0201 specificity. The CDR3 regions of the TCRβ chains are still thought to play an important role in peptide recognition, determining the dimensions of functional avidity, CD8 dependency, and tumor reactivity. It is not known how common such dominant TCRα chains are and it is not clear whether TCRβ chains are also capable of such behavior (Liang et al. 2010). Two further studies (Morgan et al. 2003; Schub et al. 2009) identified TCRs composed of TCRβ chains containing TRBV12, where the bound peptides were considerably different than the peptide used in the $HER2_{369}$ studies by Liang et al. Previous work with the influenza matrix protein showed that this antigen selects a very restricted TCR repertoire, with both TRAV27 and TRBV19 (Vα10.2 and Vβ17 respectively in the old nomenclature) being overrepresented (Moss et al. 1991). The immunodominance in TCR chain bias has been found to predominantly rely on TCRβ chains containing TRBV19 in HLA-A*0201 positive adults. Four water molecules which are completely buried in the protein-protein interaction interface of structure 1OGA, which is of the same type as those identified by Moss et al, are thought to strongly contribute to the overall shape complimentarily (Rudolph et al. 2006).

In contrast, the solved crystal structures are heavily biased in favor of TCRβ chains using TRBV6-5 (Vβ13 in the old nomenclature) in humans (Marrack et al. 2008). The TRAV regions show less bias. It is currently unknown what the reasons for identified biases are i.e. why certain TCR chains are more common in crystallographic studies. It has been speculated that specific TCRs containing some regions are easier to crystallize (Marrack et al. 2008). Interestingly, no TRBV12-3*01 have been crystallized so far, although they were found to be common in cell assay studies. Therefore, while solved crystal structures show a bias towards

TRBV6-5, TCRs binding to influenza matrix peptide show a bias towards TRAV27/TRBV19 and the remaining TCRs show a bias towards TRBV12; a crystal structure of the TRAV27/TRBV19 is available, but a TCR containing TRBV12 has not yet been crystallized.

Our prediction approach was developed using the most diverse set of TCR-MHC complexes available, restricted to HLA-A*0201. We found that the TCR-MHC crystal structures generally are representative of TCRs which have been described to bind HLA-A*0201 in cell assays (Table 6.4). However, each dataset tends to show a particular bias, as described above. While it is not possible to predict which HLA-A*0201 restricted TCRs may be isolated experimentally in the future, we believe we have been able to develop a method which is capable of reasonable predictions for likely future candidates.

Overall, it can be said that TCR amino acids will frequently interact with a large range of diverse amino acids both in the peptide and HLA-A*0201. Additionally, all CDR loops are capable of interacting with the MHC and the peptide, and all peptide residues are capable of interacting with the TCR, although the most frequently found interactions occur at positions 4 and 5 in the peptide. In contrast to previous generalizations of this binding event, where CDR1 and CDR2 loops were described as primarily interacting with the MHC molecule and the peptide having a preponderance for contacts with the CDR3 loops, we have observed a more complex picture which includes important interactions between the CDR1 and CDR2 loops and the peptide, as well as the CDR3 loop and the MHC.

7 Summary

In this thesis, we have presented work which helps to provide a better understanding of virus-host interactions with a particular focus on the human leukocyte antigen system, which has been shown repeatedly to be strongly associated with the ability of a human host to mount an effective immune response during viral infection. T cell immune responses are dependent on (1) the appropriate and effective processing of peptides from a protein source, (2) the stable binding of the peptide to the HLA molecule and (3) the recognition of this complex by the T cell receptor. We have examined aspects of each of these components in this thesis.

We have performed comprehensive analysis of the immune-driven evolution of clade B HIV-1 viruses in a large patient cohort treated at a single hospital in Germany and examined its implications for antiretroviral therapy. We analyzed associations of HLA-A, HLA-B and HLA-DRB1 alleles with the emergence of mutations in the complete protease gene and the first 330 amino acids of the reverse transcriptase gene of HIV-1, studying their distribution and persistence, and their impact on antiviral drug therapy. The clinical data of 179 HIV-1-infected patients, the results of HLA genotyping and virus sequences were analyzed using a variety of statistical approaches. We describe new HLA-associated mutations in both viral protease and RT, several of which are associated with HLA-DRB1. The mutations reported are remarkably persistent within our cohort, with a minority of patients developing them more slowly. Interestingly, several HLA-associated mutations occur at the same positions as drug resistance mutations in patients, where the viral sequence was acquired before exposure to these drugs. The influence of HLA on thymidine analogue mutation pathways was not observed. We were able to confirm immune-driven selection pressure by MHC class I and II alleles through the identification of HLA-associated mutation. HLA-B alleles were involved in more associations (68%) than either HLA-A (23%) or HLA-DRB1 (9%). As several of the HLA-associated mutations lie at positions associated with drug resistance, our results indicate possible negative effects of HLA genotypes on the development of HIV-1 drug resistance.

The adaptive immune response against HCV is also significantly shaped by the host's composition of HLA alleles. Therefore, as with HIV-1, the HLA phenotype is a critical determinant of viral evolution during infection. In order to identify novel associations of HLA class I alleles with polymorphisms of HCV escape variants, the genes encoding the HCV proteins E3, NS3 and NS5B in a cohort of 159 patients with chronic HCV genotype 1 infection were sequenced. Several HLA class I-restricted escape mutations were identified within novel putative CD8 T cell epitopes. As many of these epitopes overlap with already published epitopes and/or are predicted to lie within epitopes using *NetMHC* and *SYFPEITHI*, they appear to indicate that a region of high immunogenicity exists within the accordant protein region in the virus. A strictly separate analysis of patients infected with either HCV subtype 1a or 1b was able to show for the first time that there are strongly differing patterns of such escape mutations, implying that there is divergent adaptation to host immune pressure at the HCV subtype level. Furthermore, a number of clinical parameters are associated with the HLA class I phenotype; these consisted of liver fibrosis, inflammation and sustained viral response after therapy with pegylated interferon alpha-2b and ribavirin.

We have also developed a DAS server which serves as the first reference and annotation server for HIV-1 and HCV targeted at the genome annotation community. It is specifically adapted from the default DAS server type to meet the requirements posed by a viral genome.

Experimental screening of large sets of peptides with respect to their MHC binding capabilities is still very demanding due to the large number of possible peptide sequences and

the extensive polymorphism of the MHC proteins. Therefore, there is significant interest in the development of computational methods for predicting the binding capability of peptides to MHC molecules, as a first step towards selecting peptides for actual screening. We have examined the performance of four diverse MHC Class I prediction methods on comparatively large HLA-A and HLA-B allele peptide binding datasets extracted from the Immune Epitope Database and Analysis resource. The chosen methods span a representative cross-section of available methodology for MHC binding predictions. Until the development of IEDB, such an analysis was not possible, as the available peptide sequence datasets were small and spread out over many separate efforts. We tested three datasets which differ in the IC_{50} cutoff criteria used to select the binders and non-binders. The best performance was achieved when predictions were performed on the dataset consisting only of strong binders (IC_{50} less than 10 nM) and clear non-binders (IC_{50} greater than 10,000 nM). In addition, robustness of the predictions was only achieved for alleles that were represented with a sufficiently large (greater than 200), balanced set of binders and non-binders. All four methods show good to excellent performance on the comprehensive datasets, with the method based on artificial neural networks outperforming the other methods. However, all methods show pronounced difficulties in correctly categorizing intermediate binders.

Finally, we have examined conserved interactions between T cell receptors and MHC proteins with bound peptide antigens, which are still not well understood. In order to gain a better understanding of the interaction modes of human TCR V regions, we have performed a structural analysis of the TCRs bound to their MHC-peptide ligands in human, using the available structural models determined by X-ray crystallography. We identified important differences to previous studies in which such interactions were evaluated. Based on the interactions found in the actual experimental structures, we developed the first rule-based approach for predicting the ability of TCR residues in the CDR1, CDR2, and CDR3 loops to interact with the MHC-peptide antigen complex. Two relatively simple algorithms show good performance under cross validation.

8 Outlook

The host's genetic profile plays an important role in the ability to fight viral infections, other pathogens and cancer, and therefore needs to be taking into account during the development of immune therapy in its various forms. A number of interesting therapeutic approaches have been developed over the last few years which are continuing to be optimized, involving epitopes bound to HLA molecules and their recognition by T cell receptors. Also, much basic research is being done to help elucidate the immune response mechanisms, hopefully then leading to further novel therapeutic approaches.

For HCV, a HLA-A*0201-restricted T cell receptor has recently been identified which has a high affinity for a peptide epitope derived from NS3 protein. The TCR was isolated from a patient with chronic HCV infection. The authors were able to transduce both CD4 and CD8 T cells which successfully recognized protein-loaded targets. Thus, T cells transduced with an HCV-specific TCR may be promising for the treatment of patients with chronic HCV (Zhang et al. 2010).

Epitope mutation presents a significant barrier to the development of immunotherapy for HIV-1. A recent approach to dealing with this problem was to develop an HIV-specific TCR which was modified to have a broader coverage of epitope variants. A TCR which recognized an HLA-A*0201-restricted epitope in the HIV-1 Gag protein was engineered to have various combinations of four naturally occurring polymorphisms in the CDR3 loop. Higher avidity TCR mutants were shown to have a broad recognition of variant epitopes. This more physiologic approach may minimize deleterious TCR reactivities which were observed in previous studies which relied on random mutagenesis (Bennett et al. 2010).

Work which has identified HLA-A*0201-restricted TCRs that could potentially be used in the treatment of HCV and HIV-1 is a relatively recent development. The development and use of such methods has a much longer history in the area of cancer treatment and has recently been able to demonstrate significant success in patients in specific scenarios. In order to better understand the therapeutic approaches and their potential advantages, as well as the hurdles encountered with this type of treatment, it is useful to examine how this particular type of cancer treatment has developed and continues to develop.

Adoptive cell therapy (ACT), frequently also called adoptive T-cell immunotherapy or cellular immunotherapy, consists of a cancer patient undergoing a preparative regimen, consisting of total body irradiation or the administration of cytotoxic drugs, which depletes the patient's lymphoid compartment. Additionally, a patient's own (autologous) lymphocytes with anti-tumor activity are cultured from an excised tumor and expanded *in vitro*. Donor (allogeneic) lymphocytes may also be used. These T lymphocytes are then reinfused into the cancer patient, often in combination with appropriate growth factors to stimulate their survival and expansion *in vivo* (Rosenberg et al. 2008; Grupp and June 2010).

This treatment method offers substantial theoretical and practical advantages over other immunotherapy methods (e.g. non-specific immunomodulation with interleukin 2 (IL2) and active immunization approaches with, for example, cancer vaccines). These advantages include the fact that only a small number of anti-tumor cells need to be isolated, as they can be selected and expanded *ex vivo*. They can also be specifically activated to exhibit the required anti-tumor effector functions and such high avidity CTLs have superior antitumor efficacy *in vitro* and *in vivo.* Importantly, this method also allows for the manipulation of the host, allowing an optimal environment to be created into which the cells can be transferred (Rosenberg et al. 2008).

ACT has emerged as the most effective treatment for metastatic melanoma; recent studies have shown that in patients suffering from this type of cancer between 49% and 72% will achieve an objective clinical response (defined as a 30% reduction in the sum of the

longest diameters of measurable lesions when comparing post- and pre-treatment measurements). A significant proportion of patients will also have a complete response; in one of the most recent trials where the patients received maximum immunodepletion, 28% of the cohort achieved a complete response. Such complete responses are often durable and are seen at all organ sites including the brain (Dudley et al. 2005; Rosenberg and Dudley 2009).

However, despite the encouraging results in the case of metastatic melanoma, applications to other types of common cancer (breast, prostate and ovarian cancers) have shown relatively poor results using this methodology, as it can often be difficult to identify tumor reactive lymphocytes (Morgan et al. 2006; Berry et al. 2009).

Further work has been done to improve the adoptive T cell therapy of cancer: genetically engineered normal human lymphocytes have been generated which recognize cancer antigens and thus mediate cancer regression *in vivo*. In one ground-breaking study, patients were treated with autologous peripheral blood lymphocytes that had been transduced with genes encoding a TCR which specifically recognizes the MART1 melanoma antigen presented by HLA-A*0201 (Morgan et al. 2006). Thus, the laborious task of isolating and expanding specific T cells from individual patients has been bypassed.

Additional important components in an effective strategy for treating a wide variety of cancers are currently being developed. TCRs have been identified which target common epithelial cancers, thus giving hope for extending ACT to a wider variety of cancer types; (Rosenberg et al. 2008). Work is ongoing with the goal of generating high-affinity TCR specific to different major HLAs, so that these can be selected and employed in an "off-the-shelf" manner (Leisegang et al. 2010). The results of a pilot study by the National Cancer Institute has also recently been published. The aim of the study was to prioritize cancer antigens, generating a well vetted and priority-ranked list of cancer vaccine target antigens using objective criteria defined by different panels of experts. Of the 75 antigens selected, none had all the characteristics of an ideal cancer antigen as defined by the study authors. However, 46 antigens were immunogenic in clinical trials and 20 are suggestive of clinical efficacy in the study's therapeutic function category (Cheever et al. 2009). These tumor-associated antigens may also be suitable targets for genetically engineered TCRs in a variety of cancer types, as they represent mutant, overexpressed or abnormally expressed proteins in cancer cells, or viral proteins associated with virus-associated malignancies (Leisegang et al. 2010).

It is hoped that the knowledge gained from the ACT as well as the development of genetically engineered human lymphocytes for cancer, which express a specific TCR, can be used in the treatment of viral diseases such as HIV-1 and HCV, where work is currently only at the *in vitro* stage.

A mouse model has been recently developed which will have an important impact on the ability to easily identify high-avidity pathogenic and therapeutic human TCRs. Previously, it has been difficult to identify T cells which target self (e.g. tumors) because of the innate tolerance mechanisms of the host. The new mouse model was constructed to express the human TCR repertoire: the entire human TCRα and TCRβ gene loci were inserted into the mouse genome (1.1 and 0.7 Mb respectively). Additionally, the HLA class I transgene was also inserted, which increases the generation of CD8 T cells expressing human TCRs relative to mouse. This system makes it possible to examine the unskewed TCR repertoire against human self antigens. A successful induction of TCRs recognizing Melan-A melanoma antigen was preformed and the isolated T cell clones were found to be similar to those from patients with autoimmune vitiligo or melanoma (Li et al. 2010).

A further area of research which involves host genetic HLA profiles and the treatment of disease involves peptide-based vaccines. Numerous clinical vaccination trails have been carried out or are ongoing using HLA-presented peptides linked to a wide variety of diseases with some encouraging results, but a final breakthrough is still lacking. Examples of peptides recently evaluated in humans are a vaccine containing a Her2/neu peptide which was tested in

a clinical trial of breast cancer patients expressing HLA-A2 or HLA-A3 (Patil et al. 2010), clinical trials of multi-epitope peptide vaccines for HIV-1 (Spearman et al. 2009; Graham et al. 2010) and immunogenicity and safety studies in humans for an HCV peptide vaccine (Firbas et al. 2010). Theoretical work identifying an ideal vaccine to target *Toxoplasma gondii* in HLA-A*0201 restricted patients is ongoing (Cong et al. 2010).

Heteroclitic peptides are peptides which have been altered at their anchor residue positions and commonly used to enhance peptide vaccines. A recent study compared TCR binding to natural and anchor-modified heteroclitic peptides using surface plasmon resonance and the analysis peptide-MHC tetramer binding at the cell surface. It was possible, for the first time, to distinguish between T cells that have a strong preference for either type of antigen. In fact, substantial differences in TCR binding were identified that were, according to the authors, unpredictable. Also, the T cells primed by either group of peptide expressed different TCRs. Therefore, if a peptide vaccine is generated using heteroclitic peptides, the stimulated T cells may exhibit suboptimal recognition of the intended natural target antigen *in vivo* and such vaccines should therefore be carefully evaluated in a clinical setting to ensure efficacy (Cole et al. 2010).

While immune manipulations involving a single modality have shown to be successful in individual therapeutic settings, the interactions between tumors, viruses or other pathogens and the host's immune systems is highly complex. Thus, combination therapies may prove to generate superior adaptive immune response and thus efficacy of treatment. Without a doubt, the genetic profile of the host plays an important role in the ability to fight cancer, viral infections and other pathogens and thus needs to been taken into consideration in the development of immune therapy in its various forms.

Appendices

Appendix A: HLA Genotyping and HIV-1 Sequencing Methods

HLA genotyping of patients

HLA genotyping was performed using the INNO-LiPA line probe assays from Innogenetics for HLA-A, HLA-B (MHC Class I loci) and HLA-DRB1 (MHC class II locus) according to the manufacturer's instructions. These are assays based on the reverse hybridization principle (De Vreese et al. 2004). Ambiguous results were further resolved by strand specific PCR-SSP (One Lambda, Canoga Park, CA, USA).

Sequencing of HIV-1 proteins

For HIV-1 genotyping, a fragment of the *pol* gene containing the complete protease and the first 650-750 nucleotides of the RT were analyzed by direct sequencing of PCR products as described in Balduin et al. (Balduin et al. 2005).

Appendix B: HIV-1 Population Consensus Sequences

Protease protein (subtype B)

```
>CON_B
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTV
LVGPTPVNIIGRNLLTQIGCTLNF
```

Reverse transcriptase protein (subtype B)

```
>CON_B
PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLV
DFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLP
QGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGFTTPDKKHQKEP
PFLWMGYELHPDKWTVQPIVLPEKDSWTVNDIQKLVGKLNWASQIYAGIKVKQLCKLLRGTKALTEVIPLTEEAE
LELAENREILKEPVHGVYYDPSKDLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMRGAHTNDVKQLTEAVQKI
ATESIVIWGKTPKFKLPIQKETWEAWWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRET
KLGKAGYVTDRGRQKVVSLTDTTNQKTELQAIHLALQDSGLEVNIVTDSQYALGIIQAQPDKSESELVSQIIEQL
IKKEKVYLAWVPAHKGIGGNEQVDKLVSAGIRKVL
```

Appendix C: HLA Genotyping, HCV Genotyping and Sequencing Methods

HLA genotyping of patients

EDTA blood for HLA genotyping was collected before initiation of antiviral therapy. HLA genotyping was performed by strand-specific reverse hybridization with the INNO-LiPA line probe assays for HLA-A and HLA-B (Innogenetics) according to the manufacturer's instructions (Ahlenstiel et al. 2007). Ambiguous results were further resolved by strand-specific PCR-SSP (One Lambda, Canoga Park, CA).

HCV RNA detection and quantification, HCV genotyping

Quantitative reverse transcription polymerase chain reaction (Roche AMPLICOR HCV Monitor version 2.0; Roche Diagnostics, Basel, Switzerland) was used to quantify HCV RNA levels at baseline and during antiviral therapy. Qualitative HCV RNA detection was performed at week 4, 8, 12, 24, and 48 by bDNA assay (Versant 3.0; formerly Bayer Diagnostics, Leverkusen, Germany [now provided by Siemens]; detection limit, 615 IU/ml). In patients who had HCV RNA levels less than 1,000 IU/ml by bDNA testing, qualitative HCV RNA testing was performed with the more sensitive transcription-mediated amplification assay (TMA; Versant qualitative HCV RNA; formerly Bayer Diagnostics, Leverkusen, Germany; detection limit 5.3 IU/ml; now provided by Siemens). The cut-off of 1,000 IU/ml instead of 615 IU/ml was chosen to improve the specificity of the bDNA assay. SVR was defined as a negative qualitative HCV RNA test 24 weeks after treatment completion. Treatment was discontinued in patients with HCV RNA levels greater than 1,000 IU/ml by bDNA assay after 24 weeks of treatment. HCV genotyping was performed by using the INNO-LiPA reverse hybridization assay (INNO-LiPA HCV; Innogenetics, Gent, Belgium). All subtypes that remained unclear were subsequently analyzed by comparing their sequences of E2, NS3 and NS5B with reference sequences from the Los Alamos HCV Sequence Database (Kuiken et al. 2005).

Sequencing of HCV proteins

HCV RNA was extracted from 100 µl serum of 159 patients. Complementary DNA to HCV RNA was generated by using random hexamer oligonucleotides. HCV proteins were amplified by nested PCR. The E2 protein was amplified from amino acid positions 464 to 576 (112 amino acids), the NS3 protease from amino acid positions 1026 to 1221 (195 amino acids) and the NS5B RNA-dependent RNA polymerase from amino acid positions 2670 to 2987 (317 amino acids). After initial denaturation at 95°C for 2 min, 35 cycles of 95°C for 60 s, 47°C for 60 s, and 72°C for 60 s were performed for the first and second rounds of PCR in a PE9700 thermocycler (Perkin-Elmer Applied Biosystems, Weiterstadt, Germany). For amplification, the Expand Long Template PCR System (Roche Diagnostics, Mannheim, Germany) and Taq DNA Polymerase (Roche Diagnostics, Mannheim, Germany) were used. The amplification product was analyzed on a 2% agarose gel stained with ethidium bromide. Primers for nested PCR were: external sense, E2-1s (TGC AAT GAC TCC CTC CAC ACT GG), NS3-1s (GGC GTG TGG GGA CAT CAT C), NS5B-1s (TAT GAC ACC CGC TGC TTT GAC TC); external antisense, E2-1a (CTT CCG GAA GCA GTC CGT GGG GC), NS3-1a (GGT GGA GTA CGT GAT GGG GC), NS5B-1a (AGT GKT TAG CTC CCC GTT CA); internal sense, E2-2s (CCA GAG CGC ATG GCC AGC TG), NS5B-2s (GTC ACT GAG AAT GAC ATC CG); internal antisense, E2-2a (CAA GGT GTT GTT GCC GAC CCC CCC), NS3-2a (CAT ATA CGC TCC AAA GCC CA), NS5B-2a (TAG CTC CCC

GTT CAY CGG TTG). For sequencing of E2, NS3, and NS5B genes, 40 µl of the second round PCR product were purified with QIAquick PCR Purification Kit (Qiagen GmbH, Hilden, Germany). PCR products were labeled (BigDyeTM Terminator Cycle Sequencing Ready Reaction Kit, Applied Biosystems, Darmstadt, Germany), purified (DyEx 2.0 Spin Kit, Qiagen GmbH, Hilden, Germany), and sequenced (Applied Biosystems 3100 Genetic Analyzer, Darmstadt, Germany).

Appendix D: HCV Population Consensus Sequences

Source: HCV Sequence Database in Los Alamos (Kuiken et al. 2008).

E2 protein

```
>CON_1A_E2 (Genotype 1a)
DFDQGWGPISYANGSGPDHRPYCWHYPPKPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSGAPTYSWGENDTDVFV
LNNTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVG

>CON_1B_E2 (Genotype 1b)
KFAQGWGPITYAEPNSSDQRPYCWHYAPRPCGIVPASQVCGPVYCFTPSPVVVGTTDRFGVPTYSWGENETDVLL
LNNTRPPQGNWFGCTWMNSTGFTKTCGGPPCNIGGVG
```

NS3 protein

Note: first 5 amino acids ("GWRLL") in the consensus sequences were added for the analysis, because the patient sequences have them. This added sequence belongs to the 3' end of NS2.

```
>CON_1A_NS3 (Genotype 1a)
GWRLLAPITAYAQQTRGLLGCIITSLTGRDKNQVEGEVQIVSTAAQTFLATCINGVCWTVYHGAGTRTIASPKGP
VIQMYTNVDQDLVGWPAPQGARSLTPCTCGSSDLYLVTRHADVIPVRRRGDSRGSLLSPRPISYLKGSSGGPLLC
PAGHAVGIFRAAVCTRGVAKAVDFIPVENLETTMRSPVFTDNSSPPAVPQ

>CON_1B_NS3 (Genotype 1b)
GWRLLAPITAYSQQTRGLLGCIITSLTGRDKNQVEGEVQVVSTATQSFLATCVNGVCWTVYHGAGSKTLAGPKGP
ITQMYTNVDQDLVGWQAPPGARSLTPCTCGSSDLYLVTRHADVIPVRRRGDSRGSLLSPRPVSYLKGSSGGPLLC
PSGHAVGIFRAAVCTRGVAKAVDFVPVESMETTMRSPVFTDNSSPPAVPQ
```

NS5B protein

```
>CON_1A_NS5B (Genotype 1a)
RVAIKSLTERLYVGGPLTNSRGENCGYRRCRASGVLTTSCGNTLTCYIKARAACRAAGLQDCTMLVCGDDLVVIC
ESAGVQEDAASLRAFTEAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDGAGKRVYYLTRDPTTPLARAAWET
ARHTPVNSWLGNIIMFAPTLWARMILMTHFFSVLIARDQLEQALDCEIYGACYSIEPLDLPPIIQRLHGLSAFSL
HSYSPGEINRVAACLRKLGVPPLRAWRHRARSVRARLLSRGGRAAICGKYLFNWAVRTKLKLTPIAAAGQLDLSG
WFTAGYSGGDIYHSVSH

>CON_1B_NS5B (Genotype 1b)
RQAIRSLTERLYIGGPLTNSKGQNCGYRRCRASGVLTTSCGNTLTCYLKASAACRAAKLQDCTMLVNGDDLVVIC
ESAGTQEDAASLRVFTEAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDASGKRVYYLTRDPTTPLARAAWET
ARHTPVNSWLGNIIMYAPTLWARMILMTHFFSILLAQEQLEKALDCQIYGACYSIEPLDLPQIIERLHGLSAFSL
HSYSPGEINRVASCLRKLGVPPLRVWRHRARSVRAKLLSQGGRAATCGKYLFNWAVRTKLKLTPIPAASQLDLSG
WFVAGYSGGDIYHSLSR
```

Appendix E: HCV Alignments against Reference Sequences

The HCV viral strain H77 (GenBank accession numbers M67463 (nucleotide) and AAA45534.1 (protein)) was used as a reference strain in this study, as in the HCV Sequence Database at Los Alamos (Kuiken et al. 2008). The position of the E2, NS3, and NS5B population consensus sequence locations relative to the sequence of the H77 protein are indicated.

E2 protein alignment

```
CON_1A_E2       ------------------------------------------------------------
AAA45534.1      YFSMVGNWAKVLVVLLLFAGVDAETHVTGGNAGRTTAGLVGLLTPGAKQNIQLINTNGSW 420
CON_1B_E2       ------------------------------------------------------------


CON_1A_E2       -------------------------------------------DFDQGWGPISYANGSGP 17
AAA45534.1      HINSTALNCNESLNTGWLAGLFYQHKFNSSGCPERLASCRRLTDFAQGWGPISYANGSGL 480
CON_1B_E2       -------------------------------------------KFAQGWGPITYAEPNSS 17
                                                          .* ******:**: ..

CON_1A_E2       DHRPYCWHYPPKPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSGAPTYSWGENDTDVFVLN 77
AAA45534.1      DERPYCWHYPPRPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSGAPTYSWGANDTDVFVLN 540
CON_1B_E2       DQRPYCWHYAPRPCGIVPASQVCGPVYCFTPSPVVVGTTDRFGVPTYSWGENETDVLLLN 77
                *.*******.*:*******..******************** *.****** *:***::**

CON_1A_E2       NTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVG-------------------------- 112
AAA45534.1      NTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVGNNTLLCPTDCFRKYPEATYSRCGSG 600
CON_1B_E2       NTRPPQGNWFGCTWMNSTGFTKTCGGPPCNIGGVG-------------------------- 112
                ***** ****************.**.*** *****
```

NS3 protein alignment

Note: first 5 amino acids ("GWRLL") in the consensus sequences were added for the analysis, because the patient sequences have them. This added sequence belongs to the 3' end of NS2.

```
CON_1A_NS3      ------------------------------------------------------------
AAA45534.1      HNGLRDLAVAVEPVVFSRMETKLITWGADTAACGDIINGLPVSARRGQEILLGPADGMVS 1020
CON_1B_NS3      ------------------------------------------------------------


CON_1A_NS3      -GWRLLAPITAYAQQTRGLLGCIITSLTGRDKNQVEGEVQIVSTAAQTFLATCINGVCWT 59
AAA45534.1      KGWRLLAPITAYAQQTRGLLGCIITSLTGRDKNQVEGEVQIVSTATQTFLATCINGVCWT 1080
CON_1B_NS3      -GWRLLAPITAYSQQTRGLLGCIITSLTGRDKNQVEGEVQVVSTATQSFLATCVNGVCWT 59
                 ***********:****************************:****:*:*****:******

CON_1A_NS3      VYHGAGTRTIASPKGPVIQMYTNVDQDLVGWPAPQGARSLTPCTCGSSDLYLVTRHADVI 119
AAA45534.1      VYHGAGTRTIASPKGPVIQTYTNVDQDLVGWPAPQGSRSLTPCTCGSSDLYLVTRHADVI 1140
CON_1B_NS3      VYHGAGSKTLAGPKGPITQMYTNVDQDLVGWQAPPGARSLTPCTCGSSDLYLVTRHADVI 119
                ******::*:*.****: * *********** ** *:**********************

CON_1A_NS3      PVRRRGDSRGSLLSPRPISYLKGSSGGPLLCPAGHAVGIFRAAVCTRGVAKAVDFIPVEN 179
AAA45534.1      PVRRRGDSRGSLLSPRPISYLKGSSGGPLLCPTGHAVGLFRAAVCTRGVAKAVDFIPVEN 1200
CON_1B_NS3      PVRRRGDSRGSLLSPRPVSYLKGSSGGPLLCPSGHAVGIFRAAVCTRGVAKAVDFVPVES 179
                *****************:**************:*****:****************:***.

CON_1A_NS3      LETTMRSPVFTDNSSPPAVPQ-------------------------------------- 200
AAA45534.1      LETTMRSPVFTDNSSPPAVPQSFQVAHLHAPTGSGKSTKVPAAYAAKGYKVLVLNPSVAA 1260
CON_1B_NS3      METTMRSPVFTDNSSPPAVPQ-------------------------------------- 200
                :*******************
```

NS5B protein alignment

```
CON_1A_NS5B      ------------------------------RVAIKSLTERLYVGGPLTNSRGENCGYRRCR 31
AAA45534.1       TRCFDSTVTESDIRTEEAIYQCCDLDPQARVAIKSLTERLYVGGPLTNSRGENCGYRRCR 2700
CON_1B_NS5B      ------------------------------RQAIRSLTERLYIGGPLTNSKGQNCGYRRCR 31
                                               * **:*******:*******:*:********

CON_1A_NS5B      ASGVLTTSCGNTLTCYIKARAACRAAGLQDCTMLVCGDDLVVICESAGVQEDAASLRAFT 91
AAA45534.1       ASRVLTTSCGNTLTRYIKARAACRAAGLQDCTMLVCGDDLVVICESAGVQEDAASLRAFT 2760
CON_1B_NS5B      ASGVLTTSCGNTLTCYLKASAACRAAKLQDCTMLVNGDDLVVICESAGTQEDAASLRVFT 91
                 ** *********** *:** ****** ******** ************.********.**

CON_1A_NS5B      EAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDGAGKRVYYLTRDPTTPLARAAWETA 151
AAA45534.1       EAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDGAGKRVYYLTRDPTTPLARAAWETA 2820
CON_1B_NS5B      EAMTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDASGKRVYYLTRDPTTPLARAAWETA 151
                 **********************************.:**********************

CON_1A_NS5B      RHTPVNSWLGNIIMFAPTLWARMILMTHFFSVLIARDQLEQALDCEIYGACYSIEPLDLP 211
AAA45534.1       RHTPVNSWLGNIIMFAPTLWARMILMTHFFSVLIARDQLEQALNCEIYGACYSIEPLDLP 2880
CON_1B_NS5B      RHTPVNSWLGNIIMYAPTLWARMILMTHFFSILLAQEQLEKALDCQIYGACYSIEPLDLP 211
                 **************:****************:*:*::***:**:*:**************

CON_1A_NS5B      PIIQRLHGLSAFSLHSYSPGEINRVAACLRKLGVPPLRAWRHRARSVRARLLSRGGRAAI 271
AAA45534.1       PIIQRLHGLSAFSLHSYSPGEINRVAACLRKLGVPPLRAWRHRAWSVRARLLARGGKAAI 2940
CON_1B_NS5B      QIIERLHGLSAFSLHSYSPGEINRVASCLRKLGVPPLRVWRHRARSVRAKLLSQGGRAAT 271
                  **:**********************:***********.***** ****:**::**:**

CON_1A_NS5B      CGKYLFNWAVRTKLKLTPIAAAGQLDLSGWFTAGYSGGDIYHSVSH-------------- 317
AAA45534.1       CGKYLFNWAVRTKLKLTPITAAGRLDLSGWFTAGYSGGDIYHSVSHARPRWFWFCLLLLA 3000
CON_1B_NS5B      CGKYLFNWAVRTKLKLTPIPAASQLDLSGWFVAGYSGGDIYHSLSR-------------- 317
                 ******************.**.:*******.***********:*:

CON_1A_NS5B      -----------
AAA45534.1       AGVGIYLLPNR 3011
CON_1B_NS5B      -----------
```

APPENDICES

List of Publications

Papers in peer-reviewed journals (first author or co-first author)

Ahlenstiel G, **Roomp K**, Däumer M, Nattermann J, Vogel M, Rockstroh JK, Beerenwinkel N, Kaiser R, Nischalke HD, Sauerbruch T, Lengauer T, Spengler U; Kompetenznetz HIV/AIDS. Selective pressures of HLA genotypes and antiviral therapy on human immunodeficiency virus type 1 sequence mutation at a population level. *Clinical and Vaccine Immunology*: 14 (10), 1266-1273, 2007.

Roomp K, Antes I, Lengauer T. Predicting MHC Class I Epitopes in Large Datasets. *BMC Bioinformatics*: 11 (90), 2010.

Roomp K, Domingues F. Predicting interactions between T cell receptors and MHC-peptide complexes. *Molecular Immunology*, 48(4), 553-62, 2011.

Papers in peer-reviewed journals (contributing author)

Beerenwinkel N, Sing T, Lengauer T, Rahnenführer J, **Roomp K**, Savenkov I, Fischer R, Hoffmann D, Selbig J, Korn K, Walter H, Berg T, Braun P, Fätkenheuer G, Oette M, Rockstroh J, Kupfer B, Kaiser R, Däumer M. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics:* 21 (21), 3943-3950, 2005.

Däumer M, Awerkiew S, Sierra S, Kartashev V, Poplavskaja T, Klein R, Sichtig N, Thiele B, Lengauer T, **Roomp K**, Pfister H, Kaiser R. Selection of Thymidine Analogue Resistance Mutational Patterns in Children Infected from a Common HIV-1 Subtype G Source. *AIDS Research and Human Retroviruses*: 26 (3), 275-278, 2010.

Lange CM, **Roomp K**, Dragan A, Nattermann J, Michalk M, Spengler U, Weich V, Berg T, Lengauer T, Zeuzem S, Sarrazin C. HLA Class I Allele Associations with HCV Polymorphisms and Outcome of Antiviral Therapy in Patients with Chronic Hepatitis C. *Journal of Hepatology,* 53 (6),1022-1028, 2010.

Papers in conference proceedings, books, etc.

Roomp K, Beerenwinkel N, Sing T, Schülter E, Büch J, Sierra-Aragon S, Däumer M, Hoffmann D, Kaiser R, Lengauer T, Selbig J. Arevir: A Secure Platform for Designing Personalized Antiretroviral Therapies Against HIV. *Lecture Notes in Computer Science*: 4075,185-194, 2006.

Beerenwinkel N, **Roomp K**, Däumer M. Evolution of Drug Resistance in HIV. In: *Bioinformatics - From Genomes to Therapies 3*: 1457-1496, 2007.

Lengauer T, McHardy A, Büch J, **Roomp K**. The Current Outbreak of the Novel H1N1 Influenza: MPII Provides the Portal for Accessing the Relevant Viral Sequence Data, Spotlight 2009 (Featured Research Story of the Max Planck Institute for Informatics). Max Planck Institute for Informatics, Saarbrücken.

Bibliography

(2005). HIV Molecular Immunology 2005. B. T. M. Korber, C. Brander, B. F. Hayneset al. Los Alamos, New Mexico, Los Alamos National Laboratory, Theoretical Biology and Biophysics.

(2006). Hepatitis C Viruses - Genomes and Molecular Biology. Norfolk, UK, Horizon Bioscience.

(2007). HIV Molecular Immunology 2006/2007. Los Alamos, New Mexico, Los Alamos National Laboratory, Theoretical Biology and Biophysics.

(2009). Chronic Viral Hepatitis: Diagnosis and Therapeutics (Clinical Gastroenterology), Humana Press, Springer.

Abascal, F., R. Zardoya and D. Posada (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics **21**(9): 2104-2105.

Ahlenstiel, G., K. Roomp, M. Daumer, J. Nattermann, M. Vogel, J. K. Rockstroh, N. Beerenwinkel, R. Kaiser, H. D. Nischalke, T. Sauerbruch, T. Lengauer and U. Spengler (2007). "Selective pressures of HLA genotypes and antiviral therapy on human immunodeficiency virus type 1 sequence mutation at a population level." Clin Vaccine Immunol **14**(10): 1266-1273.

Allen, T. M., M. Altfeld, X. G. Yu, K. M. O'Sullivan, M. Lichterfeld, S. Le Gall, M. John, B. R. Mothe, P. K. Lee, E. T. Kalife, D. E. Cohen, K. A. Freedberg, D. A. Strick, M. N. Johnston, A. Sette, E. S. Rosenberg, S. A. Mallal, P. J. Goulder, C. Brander and B. D. Walker (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.

Annan, K. (2000). "We the Peoples: The Role of the United Nations in the 21st Century." UN Report.

Antes, I., S. W. Siu and T. Lengauer (2006). "DynaPred: a structure and sequence based method for the prediction of MHC class I binding peptide sequences and conformations." Bioinformatics **22**(14): e16-24.

Arden, B., S. P. Clark, D. Kabelitz and T. W. Mak (1995). "Human T-cell receptor variable gene segment families." Immunogenetics **42**(6): 455-500.

Balduin, M., S. Sierra, M. P. Daumer, J. K. Rockstroh, M. Oette, G. Fatkenheuer, B. Kupfer, N. Beerenwinkel, D. Hoffmann, J. Selbig, H. J. Pfister and R. Kaiser (2005). "Evolution of HIV resistance during treatment interruption in experienced patients and after restarting a new therapy." J Clin Virol **34**(4): 277-287.

Barouch, D. H., J. Kunstman, M. J. Kuroda, J. E. Schmitz, S. Santra, F. W. Peyerl, G. R. Krivulka, K. Beaudry, M. A. Lifton, D. A. Gorgone, D. C. Montefiori, M. G. Lewis, S. M. Wolinsky and N. L. Letvin (2002). "Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes." Nature **415**(6869): 335-339.

Beerenwinkel, N. (2004). Computational Analysis of HIV Drug resistance Data. Aachen, Germany, Shaker.

Beerenwinkel, N., J. Rahnenfuhrer, R. Kaiser, D. Hoffmann, J. Selbig and T. Lengauer (2005). "Mtreemix: a software package for learning and using mixture models of mutagenetic trees." Bioinformatics **21**(9): 2106-2107.

Beerenwinkel, N., K. Roomp and M. Däumer (2007). Evolution of Drug Resistance in HIV. Bioinformatics - From Genomes to Therapies. T. Lengauer, Wiley.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society **57**: 289-300.

Bennett, M. S., A. Joseph, H. L. Ng, H. Goldstein and O. O. Yang (2010). "Fine-tuning of T-cell receptor avidity to increase HIV epitope variant recognition by cytotoxic T lymphocytes." Aids **24**(17): 2619-2628.

Berg, T., V. Weich, G. Teuber, H. Klinker, B. Moller, J. Rasenack, H. Hinrichsen, T. Gerlach, U. Spengler, P. Buggisch, H. Balk, M. Zankel, K. Neumann, C. Sarrazin and S. Zeuzem (2009). "Individualized treatment strategy according to early viral kinetics in hepatitis C virus type 1-infected patients." Hepatology **50**(2): 369-377.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Berry, L. J., M. Moeller and P. K. Darcy (2009). "Adoptive immunotherapy for cancer: the next generation of gene-engineered immune cells." Tissue Antigens **74**(4): 277-289.

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

Blom, B., M. C. Verschuren, M. H. Heemskerk, A. Q. Bakker, E. J. van Gastel-Mol, I. L. Wolvers-Tettero, J. J. van Dongen and H. Spits (1999). "TCR gene rearrangements and expression of the pre-T cell receptor complex during human T-cell differentiation." Blood **93**(9): 3033-3043.

Blythe, M. J., I. A. Doytchinova and D. R. Flower (2002). "JenPep: a database of quantitative functional peptide data for immunology." Bioinformatics **18**(3): 434-439.

Borbulevych, O. Y., K. H. Piepenbrink, B. E. Gloor, D. R. Scott, R. F. Sommese, D. K. Cole, A. K. Sewell and B. M. Baker (2009). "T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility." Immunity **31**(6): 885-896.

Borrow, P., H. Lewicki, X. Wei, M. S. Horwitz, N. Peffer, H. Meyers, J. A. Nelson, J. E. Gairin, B. H. Hahn, M. B. Oldstone and G. M. Shaw (1997). "Antiviral pressure

exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus." Nat Med **3**(2): 205-211.

Bostan, N. and T. Mahmood (2010). "An overview about hepatitis C: a devastating virus." Crit Rev Microbiol **36**(2): 91-133.

Brenner, M. B., J. McLean, D. P. Dialynas, J. L. Strominger, J. A. Smith, F. L. Owen, J. G. Seidman, S. Ip, F. Rosen and M. S. Krangel (1986). "Identification of a putative second T-cell receptor." Nature **322**(6075): 145-149.

Briggs, J. A., T. Wilk, R. Welker, H. G. Krausslich and S. D. Fuller (2003). "Structural organization of authentic, mature HIV-1 virions and cores." EMBO J **22**(7): 1707-1715.

Brown, S. A., T. D. Lockey, C. Slaughter, K. S. Slobod, S. Surman, A. Zirkel, A. Mishra, V. R. Pagala, C. Coleclough, P. C. Doherty and J. L. Hurwitz (2005). "T cell epitope "hotspots" on the HIV Type 1 gp120 envelope protein overlap with tryptic fragments displayed by mass spectrometry." AIDS Res Hum Retroviruses **21**(2): 165-170.

Brusic, V., G. Rudy and L. C. Harrison (1998). "MHCPEP, a database of MHC-binding peptides: update 1997." Nucleic Acids Res **26**(1): 368-371.

Bryant, D. and V. Moulton (2004). "Neighbor-net: an agglomerative method for the construction of phylogenetic networks." Mol Biol Evol **21**(2): 255-265.

Burrows, S. R., R. A. Elkington, J. J. Miles, K. J. Green, S. Walker, S. M. Haryana, D. J. Moss, H. Dunckley, J. M. Burrows and R. Khanna (2003). "Promiscuous CTL recognition of viral epitopes on multiple human leukocyte antigens: biological validation of the proposed HLA A24 supertype." J Immunol **171**(3): 1407-1412.

Buslepp, J., H. Wang, W. E. Biddison, E. Appella and E. J. Collins (2003). "A correlation between TCR Valpha docking on MHC and CD8 dependence: implications for T cell selection." Immunity **19**(4): 595-606.

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

Canutescu, A. A., A. A. Shelenkov and R. L. Dunbrack, Jr. (2003). "A graph-theory algorithm for rapid protein side-chain prediction." Protein Sci **12**(9): 2001-2014.

Carreno, L. J., P. A. Gonzalez and A. M. Kalergis (2006). "Modulation of T cell function by TCR/pMHC binding kinetics." Immunobiology **211**(1-2): 47-64.

CDC (1981a). "Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California." MMWR Morb Mortal Wkly Rep **30**(25): 305-308.

CDC (1981b). "Pneumocystis pneumonia--Los Angeles." MMWR Morb Mortal Wkly Rep **30**(21): 250-252.

CDC (1982a). "Possible transfusion-associated acquired immune deficiency syndrome (AIDS) - California." MMWR Morb Mortal Wkly Rep **31**(48): 652-654.

CDC (1982b). "Unexplained immunodeficiency and opportunistic infections in infants--New York, New Jersey, California." MMWR Morb Mortal Wkly Rep **31**(49): 665-667.

CDC (1982c). "Update on acquired immune deficiency syndrome (AIDS)--United States." MMWR Morb Mortal Wkly Rep **31**(37): 507-508, 513-504.

CDC (1986). "Classification system for human T-lymphotropic virus type III/lymphadenopathy-associated virus infections." MMWR Morb Mortal Wkly Rep **35**(20): 334-339.

CDC (1992). "1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults." MMWR Recomm Rep **41**(RR-17): 1-19.

Chang, K. M., B. Rehermann, J. G. McHutchison, C. Pasquinelli, S. Southwood, A. Sette and F. V. Chisari (1997). "Immunological significance of cytotoxic T lymphocyte epitope variants in patients chronically infected by the hepatitis C virus." J Clin Invest **100**(9): 2376-2385.

Chao, D. L., M. P. Davenport, S. Forrest and A. S. Perelson (2005). "The effects of thymic selection on the range of T cell cross-reactivity." Eur J Immunol **35**(12): 3452-3459.

Cheever, M. A., J. P. Allison, A. S. Ferris, O. J. Finn, B. M. Hastings, T. T. Hecht, I. Mellman, S. A. Prindiville, J. L. Viner, L. M. Weiner and L. M. Matrisian (2009). "The prioritization of cancer antigens: a national cancer institute pilot project for the acceleration of translational research." Clin Cancer Res **15**(17): 5323-5337.

Chen, J. L., G. Stewart-Jones, G. Bossi, N. M. Lissin, L. Wooldridge, E. M. Choi, G. Held, P. R. Dunbar, R. M. Esnouf, M. Sami, J. M. Boulter, P. Rizkallah, C. Renner, A. Sewell, P. A. van der Merwe, B. K. Jakobsen, G. Griffiths, E. Y. Jones and V. Cerundolo (2005). "Structural and kinetic basis for heightened immunogenicity of T cell vaccines." J Exp Med **201**(8): 1243-1255.

Choo, Q. L., G. Kuo, A. J. Weiner, L. R. Overby, D. W. Bradley and M. Houghton (1989). "Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome." Science **244**(4902): 359-362.

Clevers, H., B. Alarcon, T. Wileman and C. Terhorst (1988). "The T cell receptor/CD3 complex: a dynamic protein ensemble." Annu Rev Immunol **6**: 629-662.

Cohen, M. S., N. Hellmann, J. A. Levy, K. DeCock and J. Lange (2008). "The spread, treatment, and prevention of HIV-1: evolution of a global pandemic." J Clin Invest **118**(4): 1244-1254.

Cole, D. K., E. S. Edwards, K. K. Wynn, M. Clement, J. J. Miles, K. Ladell, J. Ekeruche, E. Gostick, K. J. Adams, A. Skowera, M. Peakman, L. Wooldridge, D. A. Price and A. K. Sewell (2010). "Modification of MHC anchor residues generates heteroclitic peptides that alter TCR binding and T cell recognition." J Immunol **185**(4): 2600-2610.

Cole, D. K., F. Yuan, P. J. Rizkallah, J. J. Miles, E. Gostick, D. A. Price, G. F. Gao, B. K. Jakobsen and A. K. Sewell (2009). "Germ line-governed recognition of a cancer epitope by an immunodominant human T-cell receptor." J Biol Chem **284**(40): 27281-27289.

Colf, L. A., A. J. Bankovich, N. A. Hanick, N. A. Bowerman, L. L. Jones, D. M. Kranz and K. C. Garcia (2007). "How a single T cell receptor recognizes both self and foreign MHC." Cell **129**(1): 135-146.

Collins, A., D. R. Littman and I. Taniuchi (2009). "RUNX proteins in transcription factor networks that regulate T-cell lineage choice." Nat Rev Immunol **9**(2): 106-115.

Cong, H., E. J. Mui, W. H. Witola, J. Sidney, J. Alexander, A. Sette, A. Maewal and R. McLeod (2010). "Towards an immunosense vaccine to prevent toxoplasmosis: Protective Toxoplasma gondii epitopes restricted by HLA-A*0201." Vaccine.

Cox, A. L., T. Mosbruger, Q. Mao, Z. Liu, X. H. Wang, H. C. Yang, J. Sidney, A. Sette, D. Pardoll, D. L. Thomas and S. C. Ray (2005). "Cellular immune selection with hepatitis C virus persistence in humans." J Exp Med **201**(11): 1741-1752.

Dai, S., E. S. Huseby, K. Rubtsova, J. Scott-Browne, F. Crawford, W. A. Macdonald, P. Marrack and J. W. Kappler (2008). "Crossreactive T Cells spotlight the germline rules for alphabeta T cell-receptor interactions with MHC molecules." Immunity **28**(3): 324-334.

De Vreese, K., R. Barylski, F. Pughe, M. Blaser, C. Evans, J. Norton, G. Semana, R. Holman, P. Loiseau, D. Masson, M. Gielis, A. De Brauwer, I. De Canck, G. Verpooten and F. Hulstaert (2004). "Performance characteristics of updated INNO-LiPA assays for molecular typing of human leukocyte antigen A (HLA-A), HLA-B, and HLA-DQB1 alleles." Clin Diagn Lab Immunol **11**(2): 430-432.

DeFranco, A. L., R. M. Locksley and M. Robertson (2007). Immunity - The Immune Response in Infectious and Inflammatory Disease. London, UK, New Science Press Ltd.

Delves, P. J., S. J. Martin, D. R. Burton and I. M. Roitt (2006). Roitt's Essential Immunology, Blackwell Publishing.

Detours, V., R. Mehr and A. S. Perelson (1999). "A quantitative theory of affinity-driven T cell repertoire selection." J Theor Biol **200**(4): 389-403.

Detours, V. and A. S. Perelson (1999). "Explaining high alloreactivity as a quantitative consequence of affinity-driven thymocyte selection." Proc Natl Acad Sci U S A **96**(9): 5153-5158.

Ding, Y. H., K. J. Smith, D. N. Garboczi, U. Utz, W. E. Biddison and D. C. Wiley (1998). "Two human T cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids." Immunity **8**(4): 403-411.

Dönnes, P. and A. Elofsson (2002). "Prediction of MHC class I binding peptides, using SVMHC." BMC Bioinformatics **3**: 25.

Douek, D. C. (2003). "Disrupting T-cell homeostasis: how HIV-1 infection causes disease." AIDS Rev **5**(3): 172-177.

Douek, D. C., L. J. Picker and R. A. Koup (2003). "T cell dynamics in HIV-1 infection." Annu Rev Immunol **21**: 265-304.

Dowell, R. D., R. M. Jokerst, A. Day, S. R. Eddy and L. Stein (2001). "The distributed annotation system." BMC Bioinformatics **2**: 7.

Doytchinova, I. A., P. Guan and D. R. Flower (2004). "Identifiying human MHC supertypes using bioinformatic methods." J Immunol **172**(7): 4314-4323.

Draenert, R., S. Le Gall, K. J. Pfafferott, A. J. Leslie, P. Chetty, C. Brander, E. C. Holmes, S. C. Chang, M. E. Feeney, M. M. Addo, L. Ruiz, D. Ramduth, P. Jeena, M. Altfeld, S. Thomas, Y. Tang, C. L. Verrill, C. Dixon, J. G. Prado, P. Kiepiela, et al. (2004). "Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection." J Exp Med **199**(7): 905-915.

Drummond, A. and K. Strimmer (2001). "PAL: an object-oriented programming library for molecular evolution and phylogenetics." Bioinformatics **17**(7): 662-663.

Dudley, E. C., M. Girardi, M. J. Owen and A. C. Hayday (1995). "Alpha beta and gamma delta T cells can share a late common precursor." Curr Biol **5**(6): 659-669.

Dudley, M. E., J. R. Wunderlich, J. C. Yang, R. M. Sherry, S. L. Topalian, N. P. Restifo, R. E. Royal, U. Kammula, D. E. White, S. A. Mavroukakis, L. J. Rogers, G. J. Gracia, S. A. Jones, D. P. Mangiameli, M. M. Pelletier, J. Gea-Banacloche, M. R. Robinson, D. M. Berman, A. C. Filie, A. Abati, et al. (2005). "Adoptive cell transfer therapy following non-myeloablative but lymphodepleting chemotherapy for the treatment of patients with refractory metastatic melanoma." J Clin Oncol **23**(10): 2346-2357.

Ebert, P. J., L. I. Ehrlich and M. M. Davis (2008). "Low ligand requirement for deletion and lack of synapses in positive selection enforce the gauntlet of thymic T cell maturation." Immunity **29**(5): 734-745.

Egerton, M., R. Scollay and K. Shortman (1990). "Kinetics of mature T-cell development in the thymus." Proc Natl Acad Sci U S A **87**(7): 2579-2582.

Erickson, A. L., Y. Kimura, S. Igarashi, J. Eichelberger, M. Houghton, J. Sidney, D. McKinney, A. Sette, A. L. Hughes and C. M. Walker (2001). "The outcome of hepatitis C virus infection is predicted by escape mutations in epitopes targeted by cytotoxic T lymphocytes." Immunity **15**(6): 883-895.

Este, J. A. and T. Cihlar (2010). "Current status and challenges of antiretroviral research and therapy." Antiviral Res **85**(1): 25-33.

Falk, K., O. Rotzschke, S. Stevanovic, G. Jung and H. G. Rammensee (1991). "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules." Nature **351**(6324): 290-296.

Fawcett, T. (2004). "ROC graphs: notes and practical considerations for researchers." Technical Report HPL-2003-4.

Feng, D., C. J. Bond, L. K. Ely, J. Maynard and K. C. Garcia (2007). "Structural evidence for a germline-encoded T cell receptor-major histocompatibility complex interaction 'codon'." Nat Immunol **8**(9): 975-983.

Ferre, A. L., P. W. Hunt, D. H. McConnell, M. M. Morris, J. C. Garcia, R. B. Pollard, H. F. Yee, Jr., J. N. Martin, S. G. Deeks and B. L. Shacklett (2010a). "HIV Controllers with HLA-DRB1*13 and HLA-DQB1*06 Alleles Have Strong, Polyfunctional Mucosal CD4+ T-Cell Responses." J Virol **84**(21): 11020-11029.

Ferre, A. L., D. Lemongello, P. W. Hunt, M. M. Morris, J. C. Garcia, R. B. Pollard, H. F. Yee, Jr., J. N. Martin, S. G. Deeks and B. L. Shacklett (2010b). "Immunodominant HIV-specific CD8+ T-cell responses are common to blood and gastrointestinal mucosa, and Gag-specific responses dominate in rectal mucosa of HIV controllers." J Virol **84**(19): 10354-10365.

Finn, R. D., J. W. Stalker, D. K. Jackson, E. Kulesha, J. Clements and R. Pettett (2007). "ProServer: a simple, extensible Perl DAS server." Bioinformatics **23**(12): 1568-1570.

Firbas, C., T. Boehm, V. Buerger, E. Schuller, N. Sabarth, B. Jilma and C. S. Klade (2010). "Immunogenicity and safety of different injection routes and schedules of IC41, a Hepatitis C virus (HCV) peptide vaccine." Vaccine **28**(12): 2397-2407.

Flicek, P., B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, et al. (2010). "Ensembl's 10th year." Nucleic Acids Res **38**(Database issue): D557-562.

Fooksman, D. R., S. Vardhana, G. Vasiliver-Shamis, J. Liese, D. A. Blair, J. Waite, C. Sacristan, G. D. Victora, A. Zanin-Zhorov and M. L. Dustin (2010). "Functional anatomy of T cell activation and synapse formation." Annu Rev Immunol **28**: 79-105.

Foster, T. L., T. Belyaeva, N. J. Stonehouse, A. R. Pearson and M. Harris (2010). "All three domains of the hepatitis C virus nonstructural NS5A protein contribute to RNA binding." J Virol **84**(18): 9267-9277.

Frahm, N., K. Yusim, T. J. Suscovich, S. Adams, J. Sidney, P. Hraber, H. S. Hewitt, C. H. Linde, D. G. Kavanagh, T. Woodberry, L. M. Henry, K. Faircloth, J. Listgarten, C. Kadie, N. Jojic, K. Sango, N. V. Brown, E. Pae, M. T. Zaman, F. Bihl, et al. (2007). "Extensive HLA class I allele promiscuity among viral CTL epitopes." Eur J Immunol **37**(9): 2419-2433.

Garboczi, D. N., P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison and D. C. Wiley (1996). "Structure of the complex between human T-cell receptor, viral peptide and HLA-A2." Nature **384**(6605): 134-141.

Garcia, K. C., J. J. Adams, D. Feng and L. K. Ely (2009). "The molecular basis of TCR germline bias for MHC is surprisingly simple." Nat Immunol **10**(2): 143-147.

Garcia, K. C., M. Degano, R. L. Stanfield, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton and I. A. Wilson (1996). "An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex." Science **274**(5285): 209-219.

Gaudieri, S., A. Rauch, K. Pfafferott, E. Barnes, W. Cheng, G. McCaughan, N. Shackel, G. P. Jeffrey, L. Mollison, R. Baker, H. Furrer, H. F. Gunthard, E. Freitas, I. Humphreys, P. Klenerman, S. Mallal, I. James, S. Roberts, D. Nolan and M. Lucas (2009). "Hepatitis C virus drug resistance and immune-driven adaptations: relevance to new antiviral therapy." Hepatology **49**(4): 1069-1082.

Ge, D., J. Fellay, A. J. Thompson, J. S. Simon, K. V. Shianna, T. J. Urban, E. L. Heinzen, P. Qiu, A. H. Bertelsen, A. J. Muir, M. Sulkowski, J. G. McHutchison and D. B. Goldstein (2009). "Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance." Nature **461**(7262): 399-401.

Gellert, M. (2002). "V(D)J recombination: RAG proteins, repair factors, and regulation." Annu Rev Biochem **71**: 101-132.

Ghany, M. G., D. B. Strader, D. L. Thomas and L. B. Seeff (2009). "Diagnosis, management, and treatment of hepatitis C: an update." Hepatology **49**(4): 1335-1374.

Giudicelli, V., D. Chaume and M. P. Lefranc (2005). "IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes." Nucleic Acids Res **33**(Database issue): D256-261.

Goldrath, A. W. and M. J. Bevan (1999). "Selecting and maintaining a diverse T-cell repertoire." Nature **402**(6759): 255-262.

Goulder, P. J., R. E. Phillips, R. A. Colbert, S. McAdam, G. Ogg, M. A. Nowak, P. Giangrande, G. Luzzi, B. Morgan, A. Edwards, A. J. McMichael and S. Rowland-Jones (1997). "Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS." Nat Med **3**(2): 212-217.

Gouttenoire, J., F. Penin and D. Moradpour (2010). "Hepatitis C virus nonstructural protein 4B: a journey into unexplored territory." Rev Med Virol **20**(2): 117-129.

Graham, B. S., M. J. McElrath, M. C. Keefer, K. Rybczyk, D. Berger, K. J. Weinhold, J. Ottinger, G. Ferarri, D. C. Montefiori, D. Stablein, C. Smith, R. Ginsberg, J. Eldridge, A. Duerr, P. Fast and B. F. Haynes (2010). "Immunization with cocktail of HIV-derived peptides in montanide ISA-51 is immunogenic, but causes sterile abscesses and unacceptable reactogenicity." PLoS One **5**(8): e11995.

Gras, S., X. Saulquin, J. B. Reiser, E. Debeaupuis, K. Echasserieau, A. Kissenpfennig, F. Legoux, A. Chouquet, M. Le Gorrec, P. Machillot, B. Neveu, N. Thielens, B. Malissen, M. Bonneville and D. Housset (2009). "Structural bases for the affinity-driven selection of a public TCR against a dominant human cytomegalovirus epitope." J Immunol **183**(1): 430-437.

Grupp, S. A. and C. H. June (2010). "Adoptive Cellular Therapy." Curr Top Microbiol Immunol.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.

Haks, M. C., J. M. Lefebvre, J. P. Lauritsen, M. Carleton, M. Rhodes, T. Miyazaki, D. J. Kappes and D. L. Wiest (2005). "Attenuation of gammadeltaTCR signaling efficiently diverts thymocytes to the alphabeta lineage." Immunity **22**(5): 595-606.

Hale, J. S. and P. J. Fink (2010). "T-cell receptor revision: friend or foe?" Immunology **129**(4): 467-473.

Hanna, G. J., V. A. Johnson, D. R. Kuritzkes, D. D. Richman, A. J. Brown, A. V. Savara, J. D. Hazelwood and R. T. D'Aquila (2000). "Patterns of resistance mutations selected by treatment of human immunodeficiency virus type 1 infection with zidovudine, didanosine, and nevirapine." J Infect Dis **181**(3): 904-911.

Harcourt, G. C., S. Garrard, M. P. Davenport, A. Edwards and R. E. Phillips (1998). "HIV-1 variation diminishes CD4 T lymphocyte recognition." J Exp Med **188**(10): 1785-1793.

Harkiolaki, M., S. L. Holmes, P. Svendsen, J. W. Gregersen, L. T. Jensen, R. McMahon, M. A. Friese, G. van Boxel, R. Etzensperger, J. S. Tzartos, K. Kranc, S. Sainsbury, K. Harlos, E. D. Mellins, J. Palace, M. M. Esiri, P. A. van der Merwe, E. Y. Jones and L. Fugger (2009). "T cell-mediated autoimmune disease due to low-affinity crossreactivity to common microbial peptides." Immunity **30**(3): 348-357.

Hastie, T., R. Tibshirani and J. Friedman (2001). The elements of statistical learning: data mining, inference, and prediction, Springer-Verlag.

Hayday, A. C. (2000). "[gamma][delta] cells: a right time and a right place for a conserved third way of protection." Annu Rev Immunol **18**: 975-1026.

Hayes, S. M., L. Li and P. E. Love (2005). "TCR signal strength influences alphabeta/gammadelta lineage fate." Immunity **22**(5): 583-593.

Hetherington, S., A. R. Hughes, M. Mosteller, D. Shortino, K. L. Baker, W. Spreen, E. Lai, K. Davies, A. Handley, D. J. Dow, M. E. Fling, M. Stocum, C. Bowman, L. M. Thurmond and A. D. Roses (2002). "Genetic variations in HLA-B region and hypersensitivity reactions to abacavir." Lancet **359**(9312): 1121-1122.

Hogquist, K. A., T. A. Baldwin and S. C. Jameson (2005). "Central tolerance: learning self-control in the thymus." Nat Rev Immunol **5**(10): 772-782.

Horton, R., L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, Jr., M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler and S. Beck (2004). "Gene map of the extended human MHC." Nat Rev Genet **5**(12): 889-899.

Houghton, M. (2009). "The long and winding road leading to the identification of the hepatitis C virus." J Hepatol **51**(5): 939-948.

Hraber, P., C. Kuiken and K. Yusim (2007). "Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection." Hepatology **46**(6): 1713-1721.

Humphrey, W., A. Dalke and K. Schulten (1996). "VMD: visual molecular dynamics." J Mol Graph **14**(1): 33-38, 27-38.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, et al. (2009). "InterPro: the integrative protein signature database." Nucleic Acids Res **37**(Database issue): D211-215.

Huseby, E. S., F. Crawford, J. White, P. Marrack and J. W. Kappler (2006). "Interface-disrupting amino acids establish specificity between T cell receptors and complexes of major histocompatibility complex and peptide." Nat Immunol **7**(11): 1191-1199.

Huseby, E. S., J. White, F. Crawford, T. Vass, D. Becker, C. Pinilla, P. Marrack and J. W. Kappler (2005). "How the T cell repertoire becomes peptide and MHC specific." Cell **122**(2): 247-260.

Huson, D. H. and D. Bryant (2006). "Application of phylogenetic networks in evolutionary studies." Mol Biol Evol **23**(2): 254-267.

Hymes, K. B., T. Cheung, J. B. Greene, N. S. Prose, A. Marcus, H. Ballard, D. C. William and L. J. Laubenstein (1981). "Kaposi's sarcoma in homosexual men-a report of eight cases." Lancet **2**(8247): 598-600.

Jacob, L. and J. P. Vert (2008). "Efficient peptide-MHC-I binding prediction for alleles with few known binders." Bioinformatics **24**(3): 358-366.

Jerne, N. K. (1971). "The somatic generation of immune recognition." Eur J Immunol **1**(1): 1-9.

Jiao, J. and J. B. Wang (2005). "Hepatitis C virus genotypes, HLA-DRB alleles and their response to interferon-alpha and ribavirin in patients with chronic hepatitis C." Hepatobiliary Pancreat Dis Int **4**(1): 80-83.

Joachims, M. L., J. L. Chain, S. W. Hooker, C. J. Knott-Craig and L. F. Thompson (2006). "Human alpha beta and gamma delta thymocyte development: TCR gene rearrangements, intracellular TCR beta expression, and gamma delta developmental potential--differences between men and mice." J Immunol **176**(3): 1543-1552.

Joachims, T. (1999). Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Machines. B. Scholkopf, C. Burges and A. Smola. Cambridge, MA, MIT Press.

John, M., C. B. Moore, I. R. James and S. A. Mallal (2005). "Interactive selective pressures of HLA-restricted immune responses and antiretroviral drugs on HIV-1." Antivir Ther **10**(4): 551-555.

Johnson, V. A., F. Brun-Vezinet, B. Clotet, B. Conway, D. R. Kuritzkes, D. Pillay, J. M. Schapiro, A. Telenti and D. D. Richman (2005). "Update of the drug resistance mutations in HIV-1: Fall 2005." Top HIV Med **13**(4): 125-131.

Jojic, N., M. Reyes-Gomez, D. Heckerman, C. Kadie and O. Schueler-Furman (2006). "Learning MHC I--peptide binding." Bioinformatics **22**(14): e227-235.

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci **8**(3): 275-282.

Jones, N. A., X. Wei, D. R. Flower, M. Wong, F. Michor, M. S. Saag, B. H. Hahn, M. A. Nowak, G. M. Shaw and P. Borrow (2004). "Determinants of human immunodeficiency virus type 1 escape from the primary CD8+ cytotoxic T lymphocyte response." J Exp Med **200**(10): 1243-1256.

Jones, P., N. Vinod, T. Down, A. Hackmann, A. Kahari, E. Kretschmann, A. Quinn, D. Wieser, H. Hermjakob and R. Apweiler (2005). "Dasty and UniProt DAS: a perfect pair for protein feature visualization." Bioinformatics **21**(14): 3198-3199.

Kaas, Q., M. Ruiz and M. P. Lefranc (2004). "IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data." Nucleic Acids Res **32**(Database issue): D208-210.

Kaufman, L. and P. J. Rousseeuw (1990). Finding groups in data : an introduction to cluster analysis. New York, Wiley.

Kaur, G. and N. Mehra (2009). "Genetic determinants of HIV-1 infection and progression to AIDS: immune response genes." Tissue Antigens **74**(5): 373-385.

Kawashima, Y., K. Pfafferott, J. Frater, P. Matthews, R. Payne, M. Addo, H. Gatanaga, M. Fujiwara, A. Hachiya, H. Koizumi, N. Kuse, S. Oka, A. Duda, A. Prendergast, H. Crawford, A. Leslie, Z. Brumme, C. Brumme, T. Allen, C. Brander, et al. (2009). "Adaptation of HIV-1 to human leukocyte antigen class I." Nature **458**(7238): 641-645.

Kersh, G. J., F. F. Hooshmand and S. M. Hedrick (1995). "Efficient maturation of alpha beta lineage thymocytes to the CD4+CD8+ stage in the absence of TCR-beta rearrangement." J Immunol **154**(11): 5706-5714.

Kessler, J. H., B. Mommaas, T. Mutis, I. Huijbers, D. Vissers, W. E. Benckhuijsen, G. M. Schreuder, R. Offringa, E. Goulmy, C. J. Melief, S. H. van der Burg and J. W. Drijfhout (2003). "Competition-based cellular peptide binding assays for 13 prevalent HLA class I alleles using fluorescein-labeled synthetic peptides." Hum Immunol **64**(2): 245-255.

Kiepiela, P., A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, S. Chetty, P. Rathnavalu, C. Moore, K. J. Pfafferott, L. Hilton, P. Zimbwa, S. Moore, T. Allen, C. Brander, M. M. Addo, M. Altfeld, I. James, S. Mallal, M. Bunce, L. D. Barber, et al. (2004). "Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA." Nature **432**(7018): 769-775.

Klein, L., M. Hinterberger, G. Wirnsberger and B. Kyewski (2009). "Antigen presentation in the thymus for positive selection and central tolerance induction." Nat Rev Immunol **9**(12): 833-844.

Koenig, S., A. J. Conley, Y. A. Brewah, G. M. Jones, S. Leath, L. J. Boots, V. Davey, G. Pantaleo, J. F. Demarest, C. Carter and et al. (1995). "Transfer of HIV-1-specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression." Nat Med **1**(4): 330-336.

Korber, B. T., B. T. Foley, C. L. Kuiken, S. K. Pillai and J. G. Sodroski (1998). "Numbering Positions in HIV Relative to HXB2CG." Los Alamos National Laboratory, Los Alamos: 102–111.

Kosmrlj, A., A. K. Chakraborty, M. Kardar and E. I. Shakhnovich (2009). "Thymic selection of T-cell receptors as an extreme value problem." Phys Rev Lett **103**(6): 068103.

Kosmrlj, A., A. K. Jha, E. S. Huseby, M. Kardar and A. K. Chakraborty (2008). "How the thymus designs antigen-specific and self-tolerant T cell receptor sequences." Proc Natl Acad Sci U S A **105**(43): 16671-16676.

Kosmrlj, A., E. L. Read, Y. Qi, T. M. Allen, M. Altfeld, S. G. Deeks, F. Pereyra, M. Carrington, B. D. Walker and A. K. Chakraborty (2010). "Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection." Nature **465**(7296): 350-354.

Krangel, M. S. (2009). "Mechanics of T cell receptor gene rearrangement." Curr Opin Immunol **21**(2): 133-139.

Kuiken, C., C. Combet, J. Bukh, I. T. Shin, G. Deleage, M. Mizokami, R. Richardson, E. Sablon, K. Yusim, J. M. Pawlotsky and P. Simmonds (2006). "A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes." Hepatology **44**(5): 1355-1361.

Kuiken, C., P. Hraber, J. Thurmond and K. Yusim (2008). "The hepatitis C sequence database in Los Alamos." Nucleic Acids Res **36**(Database issue): D512-516.

Kuiken, C., K. Yusim, L. Boykin and R. Richardson (2005). "The Los Alamos hepatitis C sequence database." Bioinformatics **21**(3): 379-384.

Kuniholm, M. H., A. Kovacs, X. Gao, X. Xue, D. Marti, C. L. Thio, M. G. Peters, N. A. Terrault, R. M. Greenblatt, J. J. Goedert, M. H. Cohen, H. Minkoff, S. J. Gange, K. Anastos, M. Fazzari, T. G. Harris, M. A. Young, H. D. Strickler and M. Carrington (2010). "Specific human leukocyte antigen class I and II alleles associated with hepatitis C virus viremia." Hepatology **51**(5): 1514-1522.

Kuo, G., Q. L. Choo, H. J. Alter, G. L. Gitnick, A. G. Redeker, R. H. Purcell, T. Miyamura, J. L. Dienstag, M. J. Alter, C. E. Stevens and et al. (1989). "An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis." Science **244**(4902): 362-364.

Lafuente, E. M. and P. A. Reche (2009). "Prediction of MHC-peptide binding: a systematic and comprehensive overview." Curr Pharm Des **15**(28): 3209-3220.

Lange, C. M., K. Roomp, A. Dragan, J. Nattermann, M. Michalk, U. Spengler, V. Weich, T. Lengauer, S. Zeuzem, T. Berg and C. Sarrazin (2010). "HLA class I allele associations with HCV genetic variants in patients with chronic HCV genotypes 1a or 1b infection." J Hepatol.

Lata, S., M. Bhasin and G. P. Raghava (2009). "MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes." BMC Res Notes **2**: 61.

Lauer, G. M., E. Barnes, M. Lucas, J. Timm, K. Ouchi, A. Y. Kim, C. L. Day, G. K. Robbins, D. R. Casson, M. Reiser, G. Dusheiko, T. M. Allen, R. T. Chung, B. D. Walker and P. Klenerman (2004). "High resolution analysis of cellular immune responses in resolved and persistent hepatitis C virus infection." Gastroenterology **127**(3): 924-936.

Lavanchy, D. (2009). "The global burden of hepatitis C." Liver Int **29 Suppl 1**: 74-81.

Lechner, F., D. K. Wong, P. R. Dunbar, R. Chapman, R. T. Chung, P. Dohrenwend, G. Robbins, R. Phillips, P. Klenerman and B. D. Walker (2000). "Analysis of successful immune responses in persons infected with hepatitis C virus." J Exp Med **191**(9): 1499-1512.

Lefranc, M. P. (2001). "Nomenclature of the human T cell receptor genes." Curr Protoc Immunol **Appendix 1**: Appendix 1O.

Leisegang, M., S. Wilde, S. Spranger, S. Milosevic, B. Frankenberger, W. Uckert and D. J. Schendel (2010). "MHC-restricted fratricide of human lymphocytes expressing survivin-specific transgenic T cell receptors." J Clin Invest **120**(11): 3869-3877.

Leslie, A., D. Kavanagh, I. Honeyborne, K. Pfafferott, C. Edwards, T. Pillay, L. Hilton, C. Thobakgale, D. Ramduth, R. Draenert, S. Le Gall, G. Luzzi, A. Edwards, C. Brander, A. K. Sewell, S. Moore, J. Mullins, C. Moore, S. Mallal, N. Bhardwaj, et al. (2005). "Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA." J Exp Med **201**(6): 891-902.

Lever, A. M. and B. Berkhout (2008). "2008 Nobel prize in medicine for discoverers of HIV." Retrovirology **5**: 91.

Levy, J. A. (2009). "HIV pathogenesis: 25 years of progress and persistent challenges." Aids **23**(2): 147-160.

Li, L. P., J. C. Lampert, X. Chen, C. Leitao, J. Popovic, W. Muller and T. Blankenstein (2010). "Transgenic mice with a diverse human T cell antigen receptor repertoire." Nat Med **16**(9): 1029-1034.

Li, Y., Y. Huang, J. Lue, J. A. Quandt, R. Martin and R. A. Mariuzza (2005). "Structure of a human autoimmune TCR bound to a myelin basic protein self-peptide and a multiple sclerosis-associated MHC class II molecule." EMBO J **24**(17): 2968-2979.

Liang, X., L. U. Weigand, I. G. Schuster, E. Eppinger, J. C. van der Griendt, A. Schub, M. Leisegang, D. Sommermeyer, F. Anderl, Y. Han, J. Ellwart, A. Moosmann, D. H. Busch, W. Uckert, C. Peschel and A. M. Krackhardt (2010). "A single TCR alpha-

chain with dominant peptide recognition in the allorestricted HER2/neu-specific T cell repertoire." J Immunol **184**(3): 1617-1629.

Lin, H. H., S. Ray, S. Tongchusak, E. L. Reinherz and V. Brusic (2008). "Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research." BMC Immunol **9**: 8.

Livak, F., M. Tourigny, D. G. Schatz and H. T. Petrie (1999). "Characterization of TCR gene rearrangements during adult murine T cell development." J Immunol **162**(5): 2575-2580.

Lund, O., M. Nielsen, C. Kesmir, A. G. Petersen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, G. Roder, S. Justesen, S. Buus and S. Brunak (2004). "Definition of supertypes for HLA molecules using clustering of specificity matrices." Immunogenetics **55**(12): 797-810.

Lundegaard, C., K. Lamberth, M. Harndahl, S. Buus, O. Lund and M. Nielsen (2008). "NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11." Nucleic Acids Res **36**(Web Server issue): W509-512.

Ma, Z. and T. H. Finkel (2010). "T cell receptor triggering by force." Trends Immunol **31**(1): 1-6.

Macdonald, W. A., Z. Chen, S. Gras, J. K. Archbold, F. E. Tynan, C. S. Clements, M. Bharadwaj, L. Kjer-Nielsen, P. M. Saunders, M. C. Wilce, F. Crawford, B. Stadinsky, D. Jackson, A. G. Brooks, A. W. Purcell, J. W. Kappler, S. R. Burrows, J. Rossjohn and J. McCluskey (2009). "T cell allorecognition via molecular mimicry." Immunity **31**(6): 897-908.

Madrid, R., K. Janvier, D. Hitchin, J. Day, S. Coleman, C. Noviello, J. Bouchet, A. Benmerah, J. Guatelli and S. Benichou (2005). "Nef-induced alteration of the early/recycling endosomal compartment correlates with enhancement of HIV-1 infectivity." J Biol Chem **280**(6): 5032-5044.

Magiorkinis, G., E. Magiorkinis, D. Paraskevis, S. Y. Ho, B. Shapiro, O. G. Pybus, J. P. Allain and A. Hatzakis (2009). "The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis." PLoS Med **6**(12): e1000198.

Male, D., J. Brostoff, D. B. Roth and I. Roitt (2006). Immunology, Mosby Elsevir.

Mallal, S., D. Nolan, C. Witt, G. Masel, A. M. Martin, C. Moore, D. Sayer, A. Castley, C. Mamotte, D. Maxwell, I. James and F. T. Christiansen (2002). "Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir." Lancet **359**(9308): 727-732.

Marcus, F. (2008). Bioinformatics and systems biology: collaborative research and resources. Berlin-Heidelberg, Springer-Verlag.

Marrack, P., J. P. Scott-Browne, S. Dai, L. Gapin and J. W. Kappler (2008). "Evolutionarily conserved amino acids that control TCR-MHC interaction." Annu Rev Immunol **26**: 171-203.

Martin, A. M., D. Nolan, S. Gaudieri, C. A. Almeida, R. Nolan, I. James, F. Carvalho, E. Phillips, F. T. Christiansen, A. W. Purcell, J. McCluskey and S. Mallal (2004). "Predisposition to abacavir hypersensitivity conferred by HLA-B*5701 and a haplotypic Hsp70-Hom variant." Proc Natl Acad Sci U S A **101**(12): 4180-4185.

Masemola, A. M., T. N. Mashishi, G. Khoury, H. Bredell, M. Paximadis, T. Mathebula, D. Barkhan, A. Puren, E. Vardas, M. Colvin, L. Zijenah, D. Katzenstein, R. Musonda, S. Allen, N. Kumwenda, T. Taha, G. Gray, J. McIntyre, S. A. Karim, H. W. Sheppard, et al. (2004). "Novel and promiscuous CTL epitopes in conserved regions of Gag targeted by individuals with early subtype C HIV type 1 infection from southern Africa." J Immunol **173**(7): 4607-4617.

Mason, D. (1998). "A very high level of crossreactivity is an essential feature of the T-cell receptor." Immunol Today **19**(9): 395-404.

Mathers, C. D. and D. Loncar (2006). "Projections of global mortality and burden of disease from 2002 to 2030." PLoS Med **3**(11): e442.

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochim Biophys Acta **405**(2): 442-451.

Matthews, P. C., A. Prendergast, A. Leslie, H. Crawford, R. Payne, C. Rousseau, M. Rolland, I. Honeyborne, J. Carlson, C. Kadie, C. Brander, K. Bishop, N. Mlotshwa, J. I. Mullins, H. Coovadia, T. Ndung'u, B. D. Walker, D. Heckerman and P. J. Goulder (2008). "Central role of reverting mutations in HLA associations with human immunodeficiency virus set point." J Virol **82**(17): 8548-8559.

Mazza, C., N. Auphan-Anezin, C. Gregoire, A. Guimezanes, C. Kellenberger, A. Roussel, A. Kearney, P. A. van der Merwe, A. M. Schmitt-Verhulst and B. Malissen (2007). "How much can a T-cell antigen receptor adapt to structurally distinct antigenic peptides?" EMBO J **26**(7): 1972-1983.

McHutchison, J. G. (2004). "Understanding hepatitis C." Am J Manag Care **10**(2 Suppl): S21-29.

McKiernan, S. M., R. Hagan, M. Curry, G. S. McDonald, A. Kelly, N. Nolan, A. Walsh, J. Hegarty, E. Lawlor and D. Kelleher (2004). "Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source." Hepatology **40**(1): 108-114.

McMichael, A. J. and S. L. Rowland-Jones (2001). "Cellular immune responses to HIV." Nature **410**(6831): 980-987.

Mehta, C. and N. Patel (1986). "FEXACT: A Fortran subroutine for Fisher's exact test on unordered r*c contingency tables." ACM Transactions on Mathematical Software(12): 154-161.

Melek, M., M. Gellert and D. C. van Gent (1998). "Rejoining of DNA by the RAG1 and RAG2 proteins." Science **280**(5361): 301-303.

Menke, M., B. Berger and L. Cowen (2008). "Matt: local flexibility aids protein multiple structure alignment." PLoS Comput Biol **4**(1): e10.

Middleton, D., L. Menchaca, H. Rood and R. Komerofsky (2003). "New allele frequency database: http://www.allelefrequencies.net." Tissue Antigens **61**(5): 403-407.

Miretti, M. M., E. C. Walsh, X. Ke, M. Delgado, M. Griffiths, S. Hunt, J. Morrison, P. Whittaker, E. S. Lander, L. R. Cardon, D. R. Bentley, J. D. Rioux, S. Beck and P. Deloukas (2005). "A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms." Am J Hum Genet **76**(4): 634-646.

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

Morgan, R. A., M. E. Dudley, J. R. Wunderlich, M. S. Hughes, J. C. Yang, R. M. Sherry, R. E. Royal, S. L. Topalian, U. S. Kammula, N. P. Restifo, Z. Zheng, A. Nahvi, C. R. de Vries, L. J. Rogers-Freezer, S. A. Mavroukakis and S. A. Rosenberg (2006). "Cancer regression in patients after transfer of genetically engineered lymphocytes." Science **314**(5796): 126-129.

Morgan, R. A., M. E. Dudley, Y. Y. Yu, Z. Zheng, P. F. Robbins, M. R. Theoret, J. R. Wunderlich, M. S. Hughes, N. P. Restifo and S. A. Rosenberg (2003). "High efficiency TCR gene transfer into primary human lymphocytes affords avid recognition of melanoma tumor antigen glycoprotein 100 and does not alter the recognition of autologous melanoma antigens." J Immunol **171**(6): 3287-3295.

Moskophidis, D. and R. M. Zinkernagel (1995). "Immunobiology of cytotoxic T-cell escape mutants of lymphocytic choriomeningitis virus." J Virol **69**(4): 2187-2193.

Moss, P. A., R. J. Moots, W. M. Rosenberg, S. J. Rowland-Jones, H. C. Bodmer, A. J. McMichael and J. I. Bell (1991). "Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide." Proc Natl Acad Sci U S A **88**(20): 8987-8990.

Mungall, A. J., S. A. Palmer, S. K. Sims, C. A. Edwards, J. L. Ashurst, L. Wilming, M. C. Jones, R. Horton, S. E. Hunt, C. E. Scott, J. G. Gilbert, M. E. Clamp, G. Bethel, S. Milne, R. Ainscough, J. P. Almeida, K. D. Ambrose, T. D. Andrews, R. I. Ashwell, A. K. Babbage, et al. (2003). "The DNA sequence and analysis of human chromosome 6." Nature **425**(6960): 805-811.

Murphy, K., P. Travers and M. Walport (2008). Janeway's Immunobiology. New York, NY, Garland Science, Taylor & Francis Group.

Neumann-Haefelin, C., D. N. Frick, J. J. Wang, O. G. Pybus, S. Salloum, G. S. Narula, A. Eckart, A. Biezynski, T. Eiermann, P. Klenerman, S. Viazov, M. Roggendorf, R. Thimme, M. Reiser and J. Timm (2008a). "Analysis of the evolutionary forces in an immunodominant CD8 epitope in hepatitis C virus at a population level." J Virol **82**(7): 3438-3451.

Neumann-Haefelin, C., J. Timm, H. C. Spangenberg, N. Wischniowski, N. Nazarova, N. Kersting, M. Roggendorf, T. M. Allen, H. E. Blum and R. Thimme (2008b). "Virological and immunological determinants of intrahepatic virus-specific CD8+ T-cell failure in chronic hepatitis C virus infection." Hepatology **47**(6): 1824-1836.

Nielsen, M., C. Lundegaard, P. Worning, S. L. Lauemoller, K. Lamberth, S. Buus, S. Brunak and O. Lund (2003). "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations." Protein Sci **12**(5): 1007-1017.

O'Sullivan, D., T. Arrhenius, J. Sidney, M. F. Del Guercio, M. Albertson, M. Wall, C. Oseroff, S. Southwood, S. M. Colon, F. C. Gaeta and et al. (1991). "On the interaction of promiscuous antigenic peptides with different DR alleles. Identification of common structural motifs." J Immunol **147**(8): 2663-2669.

Panina-Bordignon, P., A. Tan, A. Termijtelen, S. Demotz, G. Corradin and A. Lanzavecchia (1989). "Universally immunogenic T cell epitopes: promiscuous binding to human MHC class II and promiscuous recognition by T cells." Eur J Immunol **19**(12): 2237-2242.

Patel, K., S. Norris, L. Lebeck, A. Feng, M. Clare, S. Pianko, B. Portmann, L. M. Blatt, J. Koziol, A. Conrad and J. G. McHutchison (2006). "HLA class I allelic diversity and progression of fibrosis in patients with chronic hepatitis C." Hepatology **43**(2): 241-249.

Patil, R., G. T. Clifton, J. P. Holmes, A. Amin, M. G. Carmichael, J. D. Gates, L. H. Benavides, M. T. Hueman, S. Ponniah and G. E. Peoples (2010). "Clinical and immunologic responses of HLA-A3+ breast cancer patients vaccinated with the HER2/neu-derived peptide vaccine, E75, in a phase I/II clinical trial." J Am Coll Surg **210**(2): 140-147.

Paul, W., Ed. (1998). Fundamental Immunology. New York, Raven Press.

Paul, W. E., Ed. (2008). Fundamental Immunology. Philadelphia, PA, Lippincott Williams & Wilkins.

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

Pereyra, F., X. Jia, P. McLaren, A. Telenti, P. de Bakker, B. Walker, S. Ripke, C. Brumme, S. Pulit, M. Carrington, C. Kadie, J. Carlson, D. Heckerman, R. Graham and R. Plenge (2010). "The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation." Science.

Perrin, L., L. Kaiser and S. Yerly (2003). "Travel and the spread of HIV-1 genetic variants." Lancet Infect Dis **3**(1): 22-27.

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

Peters, B., J. Sidney, P. Bourne, H. H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J. V. Ponomarenko, M. Sathiamurthy, S. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson and A. Sette (2005a). "The immune epitope database and analysis resource: from vision to blueprint." PLoS Biol **3**(3): e91.

Peters, B., J. Sidney, P. Bourne, H. H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, D. Nemazee, J. V. Ponomarenko, M. Sathiamurthy, S. P. Schoenberger, S. Stewart, P. Surko, S. Way, S. Wilson and A. Sette (2005b). "The design and implementation of the immune epitope database and analysis resource." Immunogenetics **57**(5): 326-336.

Plantier, J. C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, V. Lemee, F. Damond, D. L. Robertson and F. Simon (2009). "A new human immunodeficiency virus derived from gorillas." Nat Med **15**(8): 871-872.

Player, M. A., K. C. Barracchini, T. B. Simonis, L. Rivoltini, F. Arienti, C. Castelli, A. Mazzocchi, F. Belli, G. Parmiani and F. M. Marincola (1996). "Differences in frequency distribution of HLA-A2 subtypes between North American and Italian white melanoma patients: relevance for epitope specific vaccination protocols." J Immunother Emphasis Tumor Immunol **19**(5): 357-363.

Prilliman, K. R., K. W. Jackson, M. Lindsey, J. Wang, D. Crawford and W. H. Hildebrand (1999). "HLA-B15 peptide ligands are preferentially anchored at their C termini." J Immunol **162**(12): 7277-7284.

Prlic, A., T. A. Down and T. J. Hubbard (2005). "Adding some SPICE to DAS." Bioinformatics **21 Suppl 2**: ii40-41.

Prlic, A., T. A. Down, E. Kulesha, R. D. Finn, A. Kahari and T. J. Hubbard (2007). "Integrating sequence and structural biology with DAS." BMC Bioinformatics **8**: 333.

Rammensee, H., J. Bachmann, N. P. Emmerich, O. A. Bachor and S. Stevanovic (1999). "SYFPEITHI: database for MHC ligands and peptide motifs." Immunogenetics **50**(3-4): 213-219.

Rast, J. P., M. K. Anderson, S. J. Strong, C. Luer, R. T. Litman and G. W. Litman (1997). "alpha, beta, gamma, and delta T cell antigen receptor genes arose early in vertebrate phylogeny." Immunity **6**(1): 1-11.

Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J. A. Rafalski, E. A. Whitehorn, K. Baumeister and et al. (1985). "Complete nucleotide sequence of the AIDS virus, HTLV-III." Nature **313**(6000): 277-284.

Rauch, A., I. James, K. Pfafferott, D. Nolan, P. Klenerman, W. Cheng, L. Mollison, G. McCaughan, N. Shackel, G. P. Jeffrey, R. Baker, E. Freitas, I. Humphreys, H. Furrer, H. F. Gunthard, B. Hirschel, S. Mallal, M. John, M. Lucas, E. Barnes, et al. (2009). "Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen-restricted immune pressure." Hepatology **50**(4): 1017-1029.

Ray, S. C., L. Fanning, X. H. Wang, D. M. Netski, E. Kenny-Walsh and D. L. Thomas (2005). "Divergent and convergent evolution after a common-source outbreak of hepatitis C virus." J Exp Med **201**(11): 1753-1759.

Reche, P. A. and E. L. Reinherz (2007). "Definition of MHC supertypes through clustering of MHC peptide-binding repertoires." Methods Mol Biol **409**: 163-173.

Reichmann, D., O. Rahat, M. Cohen, H. Neuvirth and G. Schreiber (2007). "The molecular architecture of protein-protein binding sites." Curr Opin Struct Biol **17**(1): 67-76.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky and B. Korber (2000). "HIV-1 nomenclature proposal." Science **288**(5463): 55-56.

Robinson, J., M. J. Waller, S. C. Fail, H. McWilliam, R. Lopez, P. Parham and S. G. Marsh (2009). "The IMGT/HLA database." Nucleic Acids Res **37**(Database issue): D1013-1017.

Roomp, K., I. Antes and T. Lengauer (2010). "Predicting MHC class I epitopes in large datasets." BMC Bioinformatics **11**: 90.

Roomp, K., N. Beerenwinkel, T. Sing, E. Schülter, J. Büch, S. Sierra-Aragon, M. Däumer, D. Hoffmann, R. Kaiser, T. Lengauer and J. Selbig (2006). "Arevir: A Secure Platform for Designing Personalized Antiretroviral Therapies Against HIV." Lecture Notes in Computer Science **4075**: 185-194.

Roomp, K. and F. S. Domingues (2011). "Predicting interactions between T cell receptors and MHC-peptide complexes." Mol Immunol **48**(4): 553-562.

Rosenberg, S. A. and M. E. Dudley (2009). "Adoptive cell therapy for the treatment of patients with metastatic melanoma." Curr Opin Immunol **21**(2): 233-240.

Rosenberg, S. A., N. P. Restifo, J. C. Yang, R. A. Morgan and M. E. Dudley (2008). "Adoptive cell transfer: a clinical path to effective cancer immunotherapy." Nat Rev Cancer **8**(4): 299-308.

Rudolph, M. G., R. L. Stanfield and I. A. Wilson (2006). "How TCRs bind MHCs, peptides, and coreceptors." Annu Rev Immunol **24**: 419-466.

Ruppert, J., J. Sidney, E. Celis, R. T. Kubo, H. M. Grey and A. Sette (1993). "Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules." Cell **74**(5): 929-937.

Sabahi, A. (2009). "Hepatitis C Virus entry: the early steps in the viral replication cycle." Virol J **6**: 117.

Salloum, S., C. Oniangue-Ndza, C. Neumann-Haefelin, L. Hudson, S. Giugliano, M. aus dem Siepen, J. Nattermann, U. Spengler, G. M. Lauer, M. Wiese, P. Klenerman, H. Bright, N. Scherbaum, R. Thimme, M. Roggendorf, S. Viazov and J. Timm (2008). "Escape from HLA-B*08-restricted CD8 T cells by hepatitis C virus is associated with fitness costs." J Virol **82**(23): 11803-11812.

Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y.

Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, et al. (2010). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **38**(Database issue): D5-16.

Schonbach, C., J. L. Koh, D. R. Flower, L. Wong and V. Brusic (2002). "FIMM, a database of functional molecular immunology: update 2002." Nucleic Acids Res **30**(1): 226-229.

Schub, A., I. G. Schuster, W. Hammerschmidt and A. Moosmann (2009). "CMV-specific TCR-transgenic T cells for immunotherapy." J Immunol **183**(10): 6819-6830.

Schuster, I. G., D. H. Busch, E. Eppinger, E. Kremmer, S. Milosevic, C. Hennard, C. Kuttler, J. W. Ellwart, B. Frankenberger, E. Nossner, C. Salat, C. Bogner, A. Borkhardt, H. J. Kolb and A. M. Krackhardt (2007). "Allorestricted T cells with specificity for the FMNL1-derived peptide PP2 have potent antitumor activity against hematologic and other malignancies." Blood **110**(8): 2931-2939.

Scollay, R. G., E. C. Butcher and I. L. Weissman (1980). "Thymus cell migration. Quantitative aspects of cellular traffic from the thymus to the periphery in mice." Eur J Immunol **10**(3): 210-218.

Scott-Browne, J. P., J. White, J. W. Kappler, L. Gapin and P. Marrack (2009). "Germline-encoded amino acids in the alphabeta T-cell receptor control thymic selection." Nature **458**(7241): 1043-1046.

Seifert, U., H. Liermann, V. Racanelli, A. Halenius, M. Wiese, H. Wedemeyer, T. Ruppert, K. Rispeter, P. Henklein, A. Sijts, H. Hengel, P. M. Kloetzel and B. Rehermann (2004). "Hepatitis C virus mutation affects proteasomal epitope processing." J Clin Invest **114**(2): 250-259.

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

Shafer, R. W., R. Kantor and M. J. J. Gonzales (2000). "The Genetic Basis of HIV-1 Resistance to Reverse Transcriptase and Protease Inhibitors." AIDS Rev **2**: 211-228.

Shankarkumar, U., A. Pawar, K. Ghosh, S. Bajpai and A. Pazare (2010). "Human leucocyte antigen class II DRB1 and DQB1 associations in human immunodeficiency virus-infected patients of Mumbai, India." Int J Immunogenet **37**(3): 199-204.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Shatsky, M., R. Nussinov and H. J. Wolfson (2004). "A method for simultaneous alignment of multiple protein structures." Proteins **56**(1): 143-156.

Shieh, D. C., D. T. Lin, B. S. Yang, H. L. Kuan and K. J. Kao (1996). "High frequency of HLA-A*0207 subtype in Chinese population." Transfusion **36**(9): 818-821.

Shiffman, M. L. (2010). "Treatment of hepatitis C in 2011: what can we expect?" Curr Gastroenterol Rep **12**(1): 70-75.

Sidney, J., B. Peters, N. Frahm, C. Brander and A. Sette (2008). "HLA class I supertypes: a revised and updated classification." BMC Immunol **9**: 1.

Siggs, O. M., L. E. Makaroff and A. Liston (2006). "The why and how of thymocyte negative selection." Curr Opin Immunol **18**(2): 175-183.

Sim, G. K., C. Olsson and A. Augustin (1995). "Commitment and maintenance of the alpha beta and gamma delta T cell lineages." J Immunol **154**(11): 5821-5831.

Simmonds, P., A. Alberti, H. J. Alter, F. Bonino, D. W. Bradley, C. Brechot, J. T. Brouwer, S. W. Chan, K. Chayama, D. S. Chen and et al. (1994). "A proposed system for the nomenclature of hepatitis C viral genotypes." Hepatology **19**(5): 1321-1324.

Simmonds, P., J. Bukh, C. Combet, G. Deleage, N. Enomoto, S. Feinstone, P. Halfon, G. Inchauspe, C. Kuiken, G. Maertens, M. Mizokami, D. G. Murphy, H. Okamoto, J. M. Pawlotsky, F. Penin, E. Sablon, I. T. Shin, L. J. Stuyver, H. J. Thiel, S. Viazov, et al. (2005). "Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes." Hepatology **42**(4): 962-973.

Smith-Garvin, J. E., G. A. Koretzky and M. S. Jordan (2009). "T cell activation." Annu Rev Immunol **27**: 591-619.

Solberg, O. D., S. J. Mack, A. K. Lancaster, R. M. Single, Y. Tsai, A. Sanchez-Mazas and G. Thomson (2008). "Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies." Hum Immunol **69**(7): 443-464.

Spearman, P., S. Kalams, M. Elizaga, B. Metch, Y. L. Chiu, M. Allen, K. J. Weinhold, G. Ferrari, S. D. Parker, M. J. McElrath, S. E. Frey, J. D. Fuchs, M. C. Keefer, M. D. Lubeck, M. Egan, R. Braun, J. H. Eldridge, B. F. Haynes and L. Corey (2009). "Safety and immunogenicity of a CTL multiepitope peptide vaccine for HIV with or without GM-CSF in a phase I trial." Vaccine **27**(2): 243-249.

Starr, T. K., S. C. Jameson and K. A. Hogquist (2003). "Positive and negative selection of T cells." Annu Rev Immunol **21**: 139-176.

Stein, L. D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva and S. Lewis (2002). "The generic genome browser: a building block for a model organism system database." Genome Res **12**(10): 1599-1610.

Stephens, R., R. Horton, S. Humphray, L. Rowen, J. Trowsdale and S. Beck (1999). "Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC." J Mol Biol **291**(4): 789-799.

Stewart-Jones, G. B., A. J. McMichael, J. I. Bell, D. I. Stuart and E. Y. Jones (2003). "A structural basis for immunodominant human T cell receptor recognition." Nat Immunol **4**(7): 657-663.

Suzuki, T., K. Ishii, H. Aizaki and T. Wakita (2007). "Hepatitis C viral life cycle." Adv Drug Deliv Rev **59**(12): 1200-1212.

Swets, J. A. (1988). "Measuring the accuracy of diagnostic systems." Science **240**(4857): 1285-1293.

Team, R. D. C. (2003). R Reference Manual: Base Package, Network Theory.

Tester, I., S. Smyk-Pearson, P. Wang, A. Wertheimer, E. Yao, D. M. Lewinsohn, J. E. Tavis and H. R. Rosen (2005). "Immune evasion versus recovery after acute hepatitis C virus infection from a shared source." J Exp Med **201**(11): 1725-1731.

Thimme, R., J. Bukh, H. C. Spangenberg, S. Wieland, J. Pemberton, C. Steiger, S. Govindarajan, R. H. Purcell and F. V. Chisari (2002). "Viral and immunological determinants of hepatitis C virus clearance, persistence, and disease." Proc Natl Acad Sci U S A **99**(24): 15661-15668.

Thimme, R., C. Neumann-Haefelin, T. Boettler and H. E. Blum (2008). "Adaptive immune responses to hepatitis C virus: from viral immunobiology to a vaccine." Biol Chem **389**(5): 457-467.

Thomas, D. L., C. L. Thio, M. P. Martin, Y. Qi, D. Ge, C. O'Huigin, J. Kidd, K. Kidd, S. I. Khakoo, G. Alexander, J. J. Goedert, G. D. Kirk, S. M. Donfield, H. R. Rosen, L. H. Tobler, M. P. Busch, J. G. McHutchison, D. B. Goldstein and M. Carrington (2009). "Genetic variation in IL28B and spontaneous clearance of hepatitis C virus." Nature **461**(7265): 798-801.

Thompson, M. A., J. A. Aberg, P. Cahn, J. S. Montaner, G. Rizzardini, A. Telenti, J. M. Gatell, H. F. Gunthard, S. M. Hammer, M. S. Hirsch, D. M. Jacobsen, P. Reiss, D. D. Richman, P. A. Volberding, P. Yeni and R. T. Schooley (2010). "Antiretroviral treatment of adult HIV infection: 2010 recommendations of the International AIDS Society-USA panel." Jama **304**(3): 321-333.

Threlkeld, S. C., P. A. Wentworth, S. A. Kalams, B. M. Wilkes, D. J. Ruhl, E. Keogh, J. Sidney, S. Southwood, B. D. Walker and A. Sette (1997). "Degenerate and promiscuous recognition by CTL of peptides presented by the MHC class I A3-like superfamily: implications for vaccine development." J Immunol **159**(4): 1648-1657.

Timm, J., G. M. Lauer, D. G. Kavanagh, I. Sheridan, A. Y. Kim, M. Lucas, T. Pillay, K. Ouchi, L. L. Reyor, J. Schulze zur Wiesch, R. T. Gandhi, R. T. Chung, N. Bhardwaj, P. Klenerman, B. D. Walker and T. M. Allen (2004). "CD8 epitope escape and reversion in acute HCV infection." J Exp Med **200**(12): 1593-1604.

Timm, J., B. Li, M. G. Daniels, T. Bhattacharya, L. L. Reyor, R. Allgaier, T. Kuntzen, W. Fischer, B. E. Nolan, J. Duncan, J. Schulze zur Wiesch, A. Y. Kim, N. Frahm, C. Brander, R. T. Chung, G. M. Lauer, B. T. Korber and T. M. Allen (2007). "Human leukocyte antigen-associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution." Hepatology **46**(2): 339-349.

Timm, J. and M. Roggendorf (2007). "Sequence diversity of hepatitis C virus: implications for immune control and therapy." World J Gastroenterol **13**(36): 4808-4817.

Tong, J. C., T. W. Tan and S. Ranganathan (2007). "Methods and protocols for prediction of immunogenic epitopes." Brief Bioinform **8**(2): 96-108.

Toseland, C. P., D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama and D. R. Flower (2005). "AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data." Immunome Res **1**(1): 4.

UNAIDS (2009). "UNAIDS 2009 AIDS Epidemic Update." **UNAIDS**.

UNAIDS (2010). "UNAIDS Report on the Global AIDS Epidemic 2010." **UNAIDS**.

Urbani, S., B. Amadei, E. Cariani, P. Fisicaro, A. Orlandini, G. Missale and C. Ferrari (2005). "The impairment of CD8 responses limits the selection of escape mutations in acute hepatitis C virus infection." J Immunol **175**(11): 7519-7529.

Vahlne, A. (2009). "A historical reflection on the discovery of human retroviruses." Retrovirology **6**: 40.

Vella, S., M. Giuliano, P. Pezzotti, M. G. Agresti, C. Tomino, M. Floridia, D. Greco, M. Moroni, G. Visco, F. Milazzo and et al. (1992). "Survival of zidovudine-treated patients with AIDS compared with that of contemporary untreated patients. Italian Zidovudine Evaluation Group." Jama **267**(9): 1232-1236.

Vita, R., L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette and B. Peters (2010). "The immune epitope database 2.0." Nucleic Acids Res **38**(Database issue): D854-862.

von Boehmer, H., I. Aifantis, F. Gounari, O. Azogui, L. Haughn, I. Apostolou, E. Jaeckel, F. Grassi and L. Klein (2003). "Thymic selection revisited: how essential is it?" Immunol Rev **191**: 62-78.

Wakita, T., T. Pietschmann, T. Kato, T. Date, M. Miyamoto, Z. Zhao, K. Murthy, A. Habermann, H. G. Krausslich, M. Mizokami, R. Bartenschlager and T. J. Liang (2005). "Production of infectious hepatitis C virus in tissue culture from a cloned viral genome." Nat Med **11**(7): 791-796.

Wang, J. H., X. Zheng, X. Ke, M. T. Dorak, J. Shen, B. Boodram, M. O'Gorman, K. Beaman, S. J. Cotler, R. Hershow and L. Rong (2009). "Ethnic and geographical differences in HLA associations with the outcome of hepatitis C virus infection." Virol J **6**: 46.

Webster, J. P. (2009). Natural History of Host-Parasite Interactions (Advances in Parasitology). London, UK, Academic Press Inc.

Weiner, A., A. L. Erickson, J. Kansopon, K. Crawford, E. Muchmore, A. L. Hughes, M. Houghton and C. M. Walker (1995). "Persistent hepatitis C virus infection in a chimpanzee is associated with emergence of a cytotoxic T lymphocyte escape variant." Proc Natl Acad Sci U S A **92**(7): 2755-2759.

Werlen, G., B. Hausmann, D. Naeher and E. Palmer (2003). "Signaling life and death in the thymus: timing is everything." Science **299**(5614): 1859-1863.

Wertheim, J. O. and M. Worobey (2009). "Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2." PLoS Comput Biol **5**(5): e1000377.

Wertheimer, A. M., C. Miner, D. M. Lewinsohn, A. W. Sasaki, E. Kaufman and H. R. Rosen (2003). "Novel CD4+ and CD8+ T-cell determinants within the NS3 protein in subjects with spontaneously resolved HCV infection." Hepatology **37**(3): 577-589.

Winoto, A. and D. Baltimore (1989). "Separate lineages of T cells expressing the alpha beta and gamma delta receptors." Nature **338**(6214): 430-432.

Word, J. M., S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, B. K. Presley, J. S. Richardson and D. C. Richardson (1999a). "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms." J Mol Biol **285**(4): 1711-1733.

Word, J. M., S. C. Lovell, J. S. Richardson and D. C. Richardson (1999b). "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation." J Mol Biol **285**(4): 1735-1747.

Worman, H. J. (2002). The Hepatitis C Sourcebook, Mcgraw-Hill Professional.

Wozniak, A. L., S. Griffin, D. Rowlands, M. Harris, M. Yi, S. M. Lemon and S. A. Weinman (2010). "Intracellular Proton Conductance of the Hepatitis C Virus p7 Protein and Its Contribution to Infectious Virus Production." PLoS Pathog **6**(9).

Wucherpfennig, K. W., E. Gagnon, M. J. Call, E. S. Huseby and M. E. Call (2010). "Structural biology of the T-cell receptor: insights into receptor assembly, ligand recognition, and initiation of signaling." Cold Spring Harb Perspect Biol **2**(4): a005140.

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

Yin, Y. and R. A. Mariuzza (2009). "The multiple mechanisms of T cell receptor cross-reactivity." Immunity **31**(6): 849-851.

Yokomaku, Y., H. Miura, H. Tomiyama, A. Kawana-Tachikawa, M. Takiguchi, A. Kojima, Y. Nagai, A. Iwamoto, Z. Matsuda and K. Ariyoshi (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.

Yu, K., N. Petrovsky, C. Schonbach, J. Y. Koh and V. Brusic (2002). "Methods for prediction of peptide binding to MHC molecules: a comparative study." Mol Med **8**(3): 137-148.

Yusim, K., C. Kesmir, B. Gaschen, M. M. Addo, M. Altfeld, S. Brunak, A. Chigaev, V. Detours and B. T. Korber (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.

Yusim, K., R. Richardson, N. Tao, A. Dalwani, A. Agrawal, J. Szinger, R. Funkhouser, B. Korber and C. Kuiken (2005). "Los alamos hepatitis C immunology database." Appl Bioinformatics **4**(4): 217-225.

Zhang, G. L., A. M. Khan, K. N. Srinivasan, J. T. August and V. Brusic (2005). "MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides." Nucleic Acids Res **33**(Web Server issue): W172-179.

Zhang, H., C. Lundegaard and M. Nielsen (2009). "Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods." Bioinformatics **25**(1): 83-89.

Zhang, Y., Y. Liu, K. M. Moxley, L. Golden-Mason, M. G. Hughes, T. Liu, M. H. Heemskerk, H. R. Rosen and M. I. Nishimura (2010). "Transduction of human T cells with a novel T-cell receptor confers anti-HCV reactivity." PLoS Pathog **6**(7): e1001018.

Zimbwa, P., A. Milicic, J. Frater, T. J. Scriba, A. Willis, P. J. Goulder, T. Pillay, H. Gunthard, J. N. Weber, H. T. Zhang and R. E. Phillips (2007). "Precise identification of a human immunodeficiency virus type 1 antigen processing mutant." J Virol **81**(4): 2031-2038.