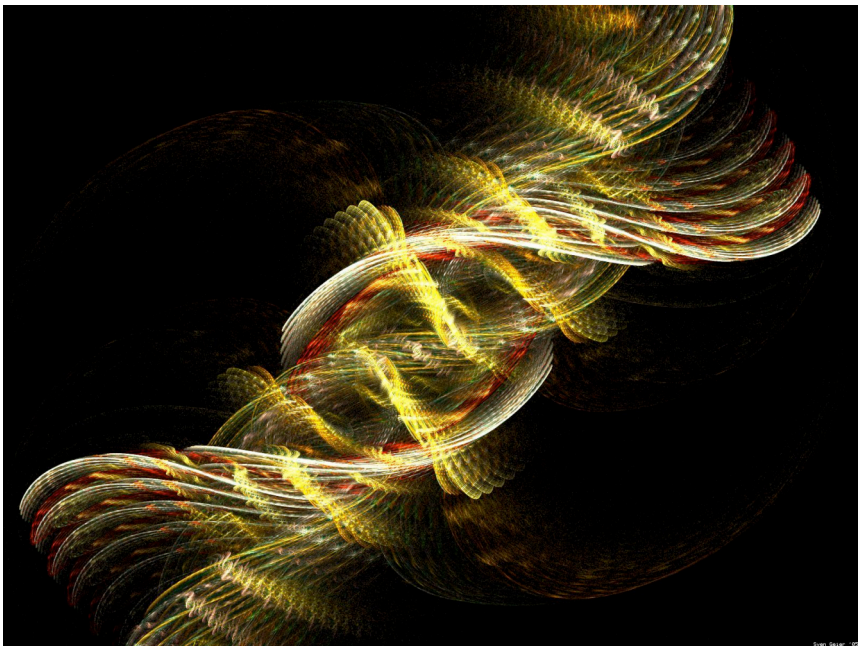
Bioinformatics in transplantation immunology

PhD thesis

by

Malene Erup Larsen



September 2010

Center for Biological Sequence Analysis
Department of Systems Biology
Technical University of Denmark

Cover image courtesy of Sven Geier:
http://www.sgeier.net/fractals/flam3/fractals/DNA.jpg

Preface

This thesis is submitted as the requirement for obtaining the PhD degree at the Department of Systems Biology at the Technical University of Denmark. The work was carried out at the Center for Biological Sequence Analysis (CBS), under the supervision of PhD Mette Voldby Larsen, PhD Thomas Skøt Jensen, and Professor Søren Brunak (main supervisor). The work presented in Chapters 2, 3 and 4 was done in collaboration with the Allogeneic Hematopoietic Cell Transplantation Laboratory, Department of Hematology, Rigshospitalet, Copenhagen, Denmark and Laboratory of Experimental Immunology, University of Copenhagen, Denmark. The work presented in Chapter 5 was done in collaboration with Laboratory of Experimental Immunology, University of Copenhagen, Denmark and Peter Medawar Building for Pathogen Research, University of Oxford, England.

English summary

Within the last 3 decades, allogeneic hematopoietic cell transplantation (allo-HCT) has become an effective treatment option for a number of malignant and non-malignant hematological diseases. According to the Center for International Blood and Marrow Transplant Research (www.cibmtr.org) around 25,000 allo-HCTs are now performed worldwide on an annual basis. These treatments are, however, plagued by an unpredictable risk of severe side effects such as graft-versus-host disease (GVHD). GVHD is a result of alloreactivity against healthy patient cells and occurs, when immune cells from a donor are transplanted into a patient and exposed to antigens, which they perceive as non-self and therefore react against. These antigens, called minor histocompatibility antigens (mHags), have received increasing attention in recent years, due to their role in GVHD as well as in the so-called graft-versus-tumor (GVT) effect. The GVT effect is also a result of donor reactivity against patient cells, however, in this case the effect is curative, as it targets the diseased hematopoietic tissue of the patient, thus preventing relapse of the malignancy. mHags are peptide fragments, encoded by polymorphic genes being disparate between patient and donor, which are presented on the surface of antigen presenting cells of the patient. To date, only around 50 mHags are known, but more are continuously being identified. The currently known mHags are believed to be only the tip of the iceberg, considering the millions of genetic differences in the human population. Identifying more mHags, and understanding their role in alloreactivity is crucial in order to better understand the interaction between GVHD and the GVT effect. In particular, there is an increased focus on identifying therapeutically relevant mHags with a hematopoietically restricted expression.

The focus of my PhD work has been the prediction of mHags using bioinformatics methods. The identification of mHags by means of traditional methods is a tedious task. Usually, an mHag is located to a small genomic region by experimental methods, and prediction methods for peptide/HLA binding are used to help identifying the exact peptide fragment constituting the mHag. We instead apply reverse immunology, beginning with the predictions which we use to compile a set of candidate mHags for experimental validation.

In this thesis, I present two such projects. In the first project, I use the prediction method *NetMHCpan*, which has been developed at CBS, to predict candidate mHags from the Y chromosome. These mHags arise due to differences between the genes on the Y chromosome, and their homologues on the X chromosome. They are believed to be involved in the higher alloreactivity observed, when the hematopoietic cells of a female donor are transplanted into a male patient. In the second project, I similarly predict candidate mHags caused by non-synonymous single nucleotide polymorphisms (nsSNPs) in proteins, where mHags have previously been identified. Experimental validations of the predicted mHags are currently being carried out at Laboratory of Experimental Immunology, University of Copenhagen.

In relation to the second project, I present a study demonstrating the correlation between the number of predicted mHag disparities, between a patient and donor, and transplantation outcome in a Danish patient cohort. Interestingly, no association is seen when only nsSNPs are considered, supporting the hypothesis that a peptide fragment, encompassing a given nsSNP, should be able to bind to one of the patient's HLA molecules to be clinically relevant.

Lastly, I present a new online prediction tool *HLArestrictor*, based on *NetMHCpan*, for the patient-specific prediction of epitopes within peptides or proteins. We developed *HLArestrictor* in order to offer the researchers a quicker overview of *NetMHCpan* predictions when adressing the common scientific question of identifying the HLA restriction element and minimal epitope within a peptide eliciting a T cell response in a given patient. *HLArestrictor* is thus also suitable for mHag prediction.

Dansk resumé

Knoglemarvstransplantation fra en matchende donor, også kaldet allogen hæmatopoietisk celletransplantation (allo-HCT), har indenfor de seneste 3 årtier udviklet sig til en effektiv behandling af alvorlige blodsygdomme. Ifølge Center for International Blood and Marrow Transplant Research (www.cibmtr.org) foretages der nu ca. 25.000 behandlinger årligt på verdensplan. Desværre er disse behandlinger plaget af en uforudsigelig risiko for alvorlige bivirkninger såsom graft-versus-host disease (GVHD). GVHD opstår ved, at donors immunsystem, som er transplanteret ind i patienten, reagerer mod antigener præsenteret på overfladen af patientens raske celler, som derved fejlagtigt angribes. Disse antigener, kaldet minor histocompatibility antigens (mHags), er peptidfragmenter kodet af gener, i hvilke patienten har en mutation, som donoren ikke har. mHags er de seneste år blevet genstand for stigende opmærksomhed p.g.a. deres betydning for GVHD såvel som for den såkaldte graft-versus-tumor (GVT)-effekt. GVT-effekten opstår også ved, at donors immunsystem reagerer mod patientens celler, men i dette tilfælde på en konstruktiv måde, idet det her er patientens syge blodceller der angribes. Man kender i dag kun ca. 50 mHags, men der opdages hele tiden flere. De mHags, der kendes i dag, forventes at være toppen af isbjerget, idet der findes millioner af gen-forskelle i det humane genom. Det er vigtigt at få identificeret flere mHags for bedre at forstå deres betydning i.f.t. til GVHD og GVT-effekten. Især arbejdes der på at identificere flere terapeutisk relevante mHags udtrykt i blodceller.

Mit PhD-arbejde har fokuseret på at forudsige mHags v.h.a. bioinformatiske metoder. De traditionelle, tidskrævende, eksperimentelle metoder bruges typisk til at lokalisere en mHag til et mindre område af genomet, hvorefter forudsigelsesmetoder kan hjælpe med at identificere det præcise peptid, der udgør mHag'en. Vores metode er omvendt, idet vi starter med forudsigelserne og bruger disse til at finde en række mulige mHags, som derefter testes eksperimentelt.

I denne afhandling præsenterer jeg to sådanne projekter. I det første projekt bruger jeg forudsigelsesmetoden *NetMHCpan*, som er udviklet på CBS, til at forudsige mulige mHags fra Y-kromosomet. Disse mHags opstår p.g.a. forskelle mellem gener på Y-kromosomet og deres homologer på X-kromosomet. De antages at være involveret i den højere forekomst af GVHD, der er forbundet med transplantationer med en mandlig patient og en kvindelig donor. I det andet projekt forudsiger jeg på tilsvarende vis mulige mHags, der skyldes enkelt-mutationsforskelle i proteiner, hvorfra mHags tidligere er identificeret. De forudsagte mulige mHags bliver i øjeblikket testet eksperimentelt på Laboratoriet for Eksperimentel Immunologi på Panum Instituttet.

I forbindelse med det andet projekt præsenterer jeg et studie, der viser sammenhængen mellem antallet af forudsagte mHag-forskelle mellem patient og donor og behandlingsresultater efter transplantationer i en dansk patientgruppe. Her ser vi, at patienter med få forudsagte mHag-forskelle har en bedre overlevelsesprocent end patienter med mange forskelle.

Endeligt præsenterer jeg et nyt, online forudsigelsesværktøj *HLArestrictor* baseret på *NetMHCpan* til brug for patientspecifik forudsigelse af epitoper fra peptider eller proteiner. Idéen med *HLArestrictor* er at give forskere et hurtigere overblik over forudsigelserne i den typiske videnskabelige situation, hvor man ønsker at finde det præsenterende vævstype-molekyle samt det præcise peptidfragment, der giver anledning til et immunrespons i en given patient. *HLArestrictor* er således også velegnet til forudsigelse af mHags.

Acknowledgements

This thesis would not have been possible without the help and support from a number of people. The Center for Biological Sequence Analysis (CBS) is characteristic for its special spirit founded by Professor Søren Brunak who deserves great thanks for providing a friendly and scientifically stimulating atmosphere. As my main supervisor, Søren has provided the basis for making my PhD study possible. He has been a great source of inspiration and guidance throughout my time at CBS.

My daily supervisor, Mette Voldby Larsen deserves the greatest thanks of all. Her scientific overview and talent for project planning together with her kind and helpful personality makes her the best supervisor I could ever imagine. Also, I am very grateful for her thorough proofreading of this thesis.

I also thank Ole Lund for welcoming me in the immunology group. His leadership is admirable and forms the basis of a scientifically as well as socially successful group. He is always a good source of advice and infectious optimism.

A special thanks goes to Morten Nielsen for our fruitful collaboration on the *HLArestrictor*. Whether in Denmark or Argentina, he was always available and full of interesting and productive ideas.

I also greatly appreciate our collaboration with everyone at Rigshospitalet and Panum. A special thank you goes to Brian Kornblit, a bright and friendly person with whom I enjoyed to work during the writing of our paper. Also, I thank Søren Buus and Lars Vindeløv for sharing their great experience as senior members of the collaboration.

I thank Thomas Skøt for supervising me in the beginning of my PhD study, and for introducing me to the interesting field of bioinformatics together with Rasmus Wernersson. Thanks, also to Nicholas Gauthier and Lars Juhl, for our collaboration on the Cyclebase paper.

I would like to thank everyone I worked with at CBS, especially the members of the immunology group and the ISB group, as well as the many nice office mates I have had during my time at CBS. Thanks to Nico for the nice LaTeX template used for this thesis.

The CBS administration deserves great thanks for always being available and making everything work smoothly at CBS. The same is true for the system administration led by Kristoffer Rapacki, who is always available for an interesting chat.

I would like to acknowledge all my co-authors and internal as well as external collaborators not mentioned already. I would also like to thank my fellow PhD students for providing such a great scientific as well as social atmosphere. Especially Rikke Rentsch deserves special thanks for our collaboration and friendship.

Finally, I would like to thank my friends and family for all their support and kind interest in my studies. Morten, I am deeply grateful for everything you have done for me during my PhD, for being my LaTeX expert, for your thorough proofreading, and for always cheering me up.

Papers included in this thesis

In this thesis, the following two papers are presented. The first one is published and the second one is recently submitted. Additionally, the work presented in Chapters 2 and 4 will be submitted as papers once the experimental work has been finalized by our collaborators.

- **Degree of predicted minor histocompatibility antigen mismatch correlates with poorer clinical outcomes of nonmyeloablative allogeneic hematopoietic cell transplantation**
  Malene Erup Larsen, Brian Kornblit, Mette Voldby Larsen, Tania Nicole Masmas, Morten Nielsen, Martin Thiim, Peter Garred, Anette Stryhn, Ole Lund, Soren Buus, and Lars Vindelov.
  Biol Blood Marrow Transplant Oct. 2010, 16(10):1370-81

- **HLArestrictor – a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides or proteins**
  Malene Erup Larsen, Henrik Kløverpris, Anette Stryhn, Catherine K. Koofhethile, Stuart Sims, Thumbi Ndung'u, Philip Goulder, Søren Buus, Morten Nielsen.
  Submitted to Immunogenetics Aug. 2010

Additional papers

During my PhD study, I have been involved in the following publications. However, these do not fit into the main topics of the thesis, and are therefore not included.

- **Cyclebase.org – a comprehensive multi-organism online database of cell-cycle experiments**
  N.P. Gauthier, **M.E. Larsen**, R. Wernersson, U. de Lichtenberg, L.J. Jensen, S. Brunak, T.S. Jensen.
  Nucleic Acids Res. Jan 2008, 36 (Database issue), D854-9

- **OCT4 and downstream factors are expressed in human somatic urogenital epithelia and in culture of epididymal spheres**
  D.M. Kristensen, J.E. Nielsen, M. Kalisz, M.D. Dalgaard, K. Audouze, **M.E. Larsen**, G.K. Jacobsen, T. Horn, S. Brunak, N.E. Skakkebaek, H. Leffers.
  Mol Hum Reprod. Jan. 2010 [Epub ahead of print]

- **UniSH2, a quantitative pan-domain predictive method for phosphotyrosyl peptide recognized by the SH2 domain family**
  H. Zhang, O. Lund, C. Lundegaard, **M.E. Larsen**, M. Nielsen
  Manuscript in preparation

Abbreviations

aGVHD	acute graft-versus-host disease
allo-HCT	allogeneic hematopoietic cell transplantation
ANN	artificial neural network
cGVHD	chronic graft-versus-host disease
CI	confidence interval
CTL	cytotoxic T lymphocyte
ELIspot	enzyme-linked immunospot assay
ER	endoplasmic reticulum
GVHD	graft-versus-host disease
GVT	graft-versus-tumor
GWAS	genome-wide association study
HCT	hematopoietic cell transplantation
HLA	human leukocyte antigen
HR	hazard ratio
HWE	Hardy-Weinberg equilibrium
ICS	intracellular cytokine staining
IFNγ	interferon-gamma
LD	linkage disequilibrium
MC	myeloablative conditioning
mHag	minor histocompatibility antigen
MHC	major histocompatibility complex
MUD	matched unrelated donor
NMA	non-myeloablative
NMC	non-myeloablative conditioning
nsSNP	non-synonymous single nucleotide polymorphism
OS	overall survival
PBMC	peripheral blood mononuclear cell
PFS	progression free survival
RI	relapse incidence
RRM	relapse related mortality
SNP	single nucleotide polymorphism
TAP	transporter associated with antigen processing
TCR	T cell receptor
TRM	treatment related mortality

Contents

Chapter 1

Introduction

Allogeneic hematopoietic cell transplantation (allo-HCT) can be a powerful curative treatment of a number of malignant and non-malignant hematological diseases (Copelan, 2006). However, serious complications such as graft-versus-host disease (GVHD) or graft rejection are common, and the occurrences of these are still difficult to predict. To understand the mechanisms of the immune system involved in the positive curative effects as well as the negative side effects of an allo-HCT treatment, an understanding of the normal functions of the immune system is necessary. The human immune system has evolved as a defense against pathogens and cancers, whereas the presence of allogenic hematopoietic cells in the body is an artificial situation. The allo-HCT treatment aims at cheating this highly optimized system in order to rid the body of the malignancy, while minimizing the side effects.

1.1 The adaptive immune system

The immune system consists of two major parts - the innate and the adaptive immune system. The innate immune system is the evolutionarily older, unspecific immune system which provides immediate response to invading pathogens by recognizing features common to these. The innate immune system has no memory of previous infections and is therefore not capable of optimizing its response in subsequent encounters with the same pathogen (Murphy et al.). The innate immune system alone is of little relevance in a transplantation setting but is important due to its interactions with the adaptive immune system.

The adaptive immune system consists of two major cell types called B and T lymphocytes, responsible for, respectively, the humoral and cellular immunity. This part of the immune system provides a slower but specific response to pathogens and responds more effectively if the same pathogen is encountered again. The main role of the B lymphocytes is the production of antibodies specific to a unique antigen. These antibodies are excreted into the blood where they bind and inactivate the antigens. The T lymphocytes, also called T cells, instead focus on identifying and killing infected or cancerous host cells. This introduction will focus only on the function of the T cells, since this part of the immune system is most important to HCT treatments. Figure 1.1 gives an overview of the different cell types of the immune system and the blood.

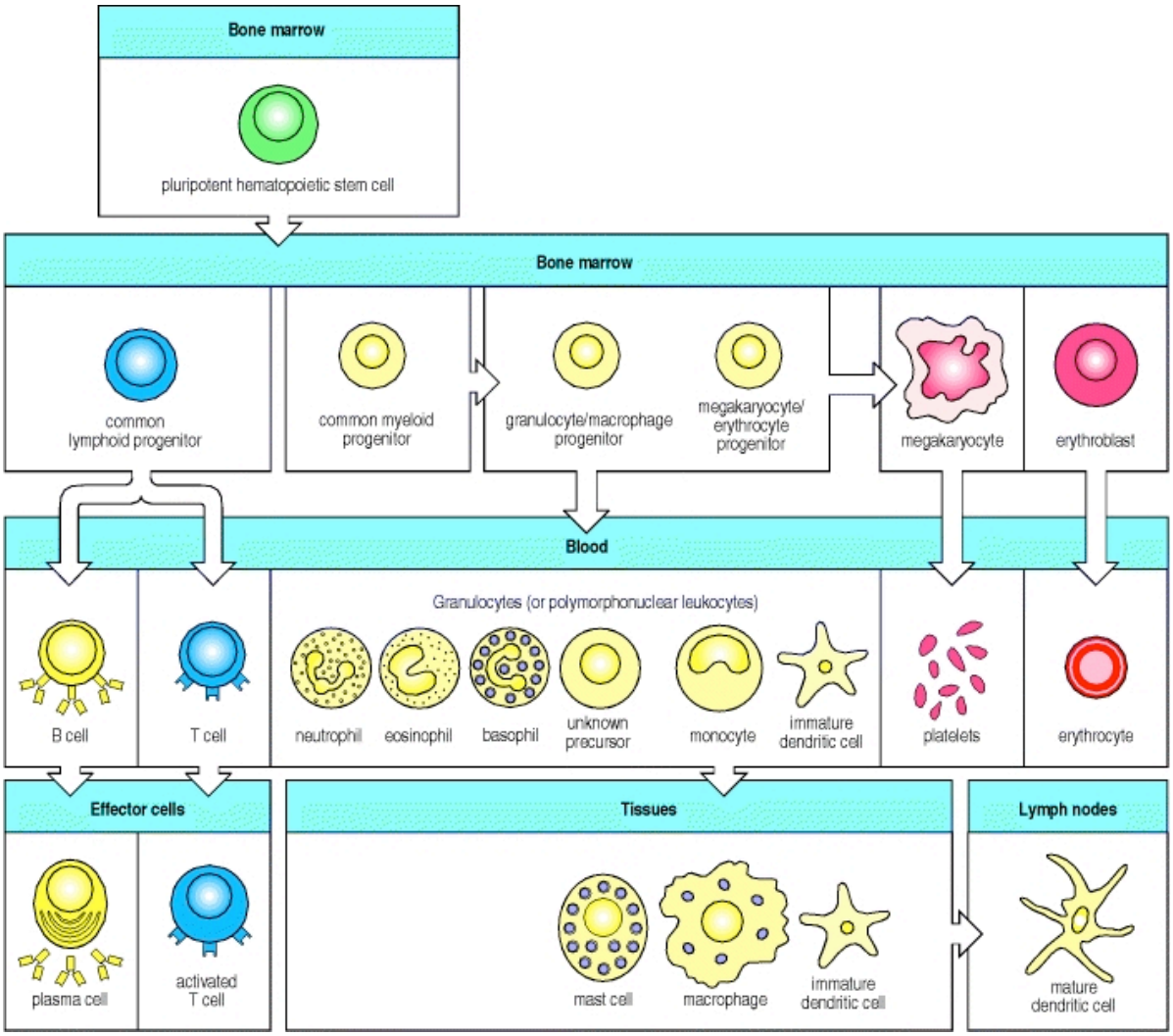


Figure 1.1: **Overview of the different cell types of the blood.** Hematopoietic stem cells in the bone marrow differentiate into leukocytes (white blood cells), erythrocytes (red blood cells) and thrombocytes (blood platelets). Leukocytes are divided into lymphocytes (B and T cells), responsible for the adaptive immunity, and myeloid cells, involved in both the innate and adaptive immune response. Erythrocytes carry oxygen in the blood. Thrombocytes are produced by megakaryocytes, and are responsible for blood clotting. From (Murphy et al.)

1.1.1 Antigen presentation

The mechanism, by which the T cells are able to identify if a cell is infected or not, depends on the presentation of protein fragments on the cell surface. The fragments, called antigens, are peptides of 8-11 amino acids, which result from the continual degradation of proteins present in the cell cytoplasm. Hereby, the cell displays a snapshot of the internal situation including any potential foreign proteins.

The antigen processing and presentation pathway is presented in Figure 1.2. All proteins present in the cell cytoplasm, whether they belong to the cell's own proteome (self) or originate from a virus or bacteria infecting the cell (non-self), are routinely tagged for degradation by the attachment of a small protein called ubiquitin. The tagged proteins are then degraded by the proteasome into peptides of 4-20 amino acids, which can then be transported by the transporter associated with antigen processing (TAP) into the endoplasmic reticulum (ER).

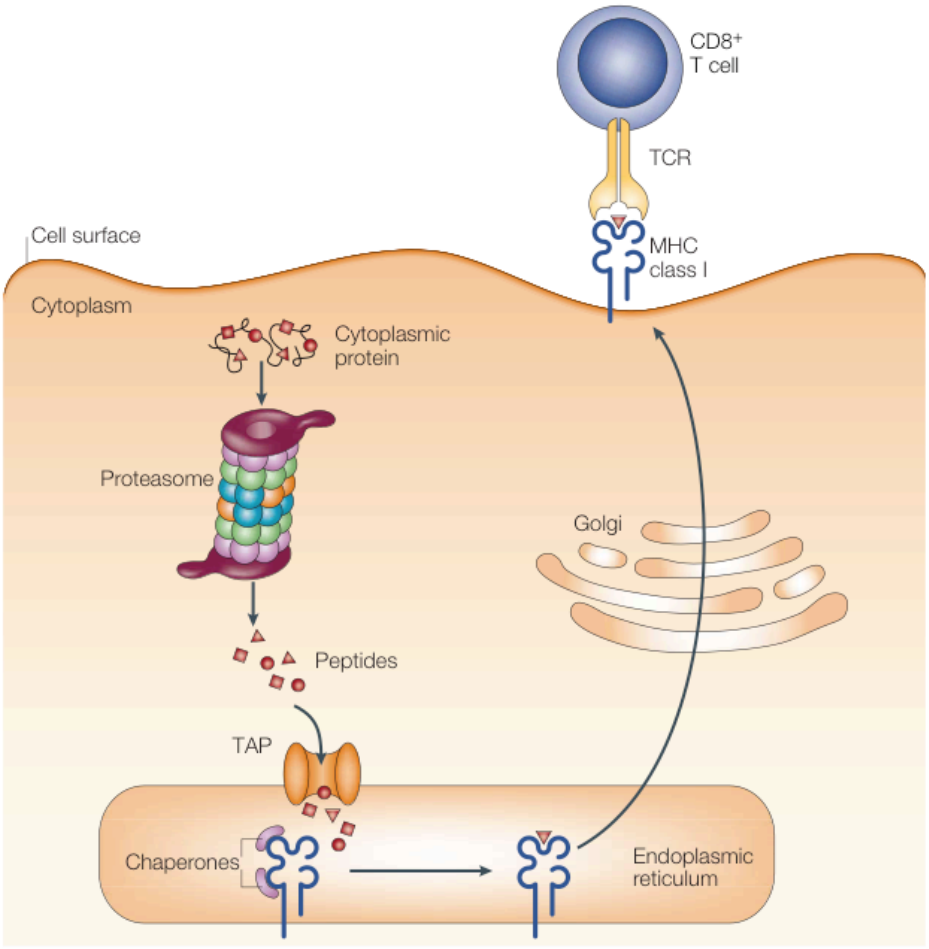


Figure 1.2: **The MHC class I antigen presentation pathway.** All proteins present in the cell cytoplasm are routinely degraded into peptides by the proteasome. The peptides are transported into the ER by the TAP molecule, where they are further processed by amino peptidases. MHC class I molecules in the ER then selectively bind peptides forming a peptide/MHC complex which is transported via the Golgi apparatus to the cell surface, where it is presented to passing T cells. From (Yewdell et al., 2003).

In the ER, further N-terminal shortening of the peptides by amino peptidases may take place, before peptides of 8-11 amino acids can form complexes with the major histocompatibility complex (MHC) class I molecules (Yewdell et al., 2003). Most (~70%) of these peptides, also called MHC ligands, are 9 amino acids long (see Table 1.1). The binding of peptides to MHC molecules is the most selective step in the pathway, as only a few percent of the peptides or less will have a sequence that fits a particular MHC molecule (Yewdell and Bennink, 1999). Once a stable peptide/MHC complex has formed, it is transported via the Golgi apparatus to the cell surface, ready for inspection by circulating T cells.

1.1.2 The HLA system

The human leukocyte antigen (HLA) system is the human variant of the MHC. It consists of a number of genes located on chromosome 6 in two main regions called MHC class I and II. Class I is again divided into the classical HLA-A, -B, and -C genes, which are highly polymorphic with more than 3,500 different alleles in the human population (Robinson et al., 2001, 2003,

Length of MHC ligands	8	9	10	11	12
Prevalence	7.5%	68.3%	16.4%	5.6%	1.6%

Table 1.1: **Lengths of MHC ligands**. The table shows the percentage distribution of the different peptide lengths of MHC ligands listed in the SYFPEITHI database (Rammensee et al., 1999)

Gene	A	B	C	E	F	G
Alleles	1,193	1,800	829	9	21	46
Proteins	891	1,419	623	3	4	15

Table 1.2: **HLA class I polymorphism**. The classical loci (A, B, and C) are significantly more polymorphic than the non-classical loci (E, F and G). Source: IMGT/HLA Database (Robinson et al., 2009)

2009). In addition, the MHC class I region contains the non-classical HLA-E, -F, and -G genes which are relatively conserved (O'Callaghan and Bell, 1998). Table 1.2 gives the number of different alleles for the six loci. The class I genes are expressed in all human cells with a nucleus and present peptide fragments from inside the cell as described in Section 1.1.1. Class II genes called HLA-DR, -DQ, and -DP are only expressed in specialized antigen presenting cells such as dendritic cells and macrophages. They present peptide fragments derived from extracellular proteins including pathogens (Murphy et al.).

1.1.3 Binding motifs

Each HLA molecule is characterized by its own binding motif, which can be visualized with a sequence logo, which is a graphical representation of multiple sequences as described in (Schneider and Stephens, 1990). For a given HLA molecule the sequence logo is a useful illustration of amino acid preferences for each position in 9mer peptides capable of binding to the HLA molecule. An example of a sequence motif logo is shown in Figure 1.3. For each of the 9 amino acid positions, the frequency of each of the 20 possibly amino acids, within peptides known to bind the HLA molecule, is represented by the height of the corresponding letter. The height of all the letters in a stack corresponds to the information content at the given position, that is, how conserved the occurrence of specific amino acids is. The colors of the letters represent the physical and chemical properties of the amino acids. Acidic amino acids (D and E) are red, basic amino acids (H, K and R) are blue, hydrophobic amino acids (A, C, F, I, L, M, P, V, and W) are black, and neutral amino acids (G, N, Q, S, T, and Y) are green. Sequence motif logos for all human HLA alleles are available at the website www.cbs.dtu.dk/biotools/MHCMotifViewer which was developed at CBS by Rapin et al. (2008).

1.1.4 T cells

Hematopoietic stem cells in the bone marrow differentiate into all the different cells of the immune system including T cells (see Figure 1.1). The precursor T cells produced in the bone marrow go through a complex education process in the thymus, where they learn to distinguish self from non-self. T cells are characterized by the T cell receptor (TCR) expressed on their surface and their co-receptor, which can be either CD8 or CD4. A T cell expressing the CD8 receptor is then denoted as CD8+.

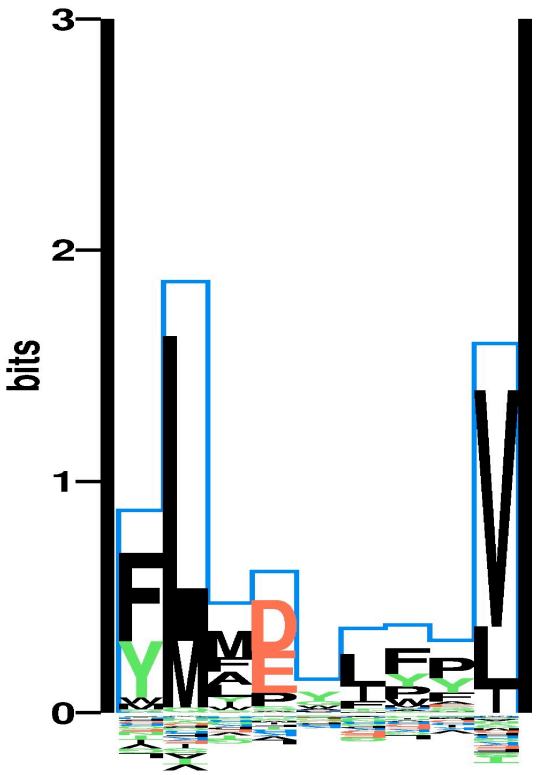


Figure 1.3: **Sequence motif logo for HLA-A*0201.** The sequence logo shows that the HLA-A*0201 molecule preferably binds peptides having the hydrophobic (black) amino acids leucine (L) or methionine (M) at position 2 and valine (V) or leucine (L) at position 9. The other positions are less conserved, but also have amino acids preferences as shown.

TCR recombination

The TCR usually consists of an α and a β chain and is practically unique for each T cell due to somatic recombination of genes taking place in the thymus. The α chain is encoded by a variable (V), a joining (J) and a constant (C) gene segment; the first two are found in multiple copies ($\sim$70 V_α and 61 J_α segments) on chromosome 6, while only one copy of the C_α gene segment exists. The β chain is encoded by similar V, J and C gene segments and an additional diversity (D) gene segment. The V and J genes likewise exist in multiple copies (52 V_β and 13 J_β segments), while the D_β and C_β genes each exist in two copies (Murphy et al.). The multiple copies of the gene segments together with the random addition of nucleotides between V(D)J gene segments result in a huge number of recombination possibilities with a potential TCR diversity of 10^{18}.

T cell training in the thymus

As the T cell precursors migrate from the bone marrow to the thymus, they do not yet express any TCR or co-receptor on their surface. During their early development in the thymus, they begin expressing their unique TCR and both the CD8 and the CD4 co-receptors. They then undergo a *positive* selection, in which they interact with self-peptide/self-MHC complexes presented by specialized cells in the thymus. These cells synthesize and degrade all proteins encoded by the individual's genome, including tissue specific proteins. T cells, which do not

recognize any self-peptide/self-MHC complex, undergo apoptosis in order to ensure that all cells leaving the thymus can recognize peptides bound to self-MHC molecules. T cells recognizing peptides bound to MHC class I molecules are destined to become cytotoxic T lymphocytes (CTLs) expressing only the CD8 co-receptor. T cells recognizing peptides bound to MHC class II molecules are destined to become T helper cells or regulatory T cells expressing the CD4 co-receptor. Before leaving the thymus, the T cells also undergo a *negative* selection, by which they are induced to undergo apoptosis if they interact too strongly with any self-peptide/self-MHC complex. The mechanisms of *positive* and *negative* selection are illustrated in Figure 1.4. The few percent of the T cells that survive both the positive and negative selection are then ideally specific to self-MHC molecules, but should not be activated by self-peptides. However, some T cells escape the negative selection and are instead inactivated after leaving the thymus by other mechanisms, such as regulatory T cells or anergy, in order to prevent them from causing autoimmune diseases.

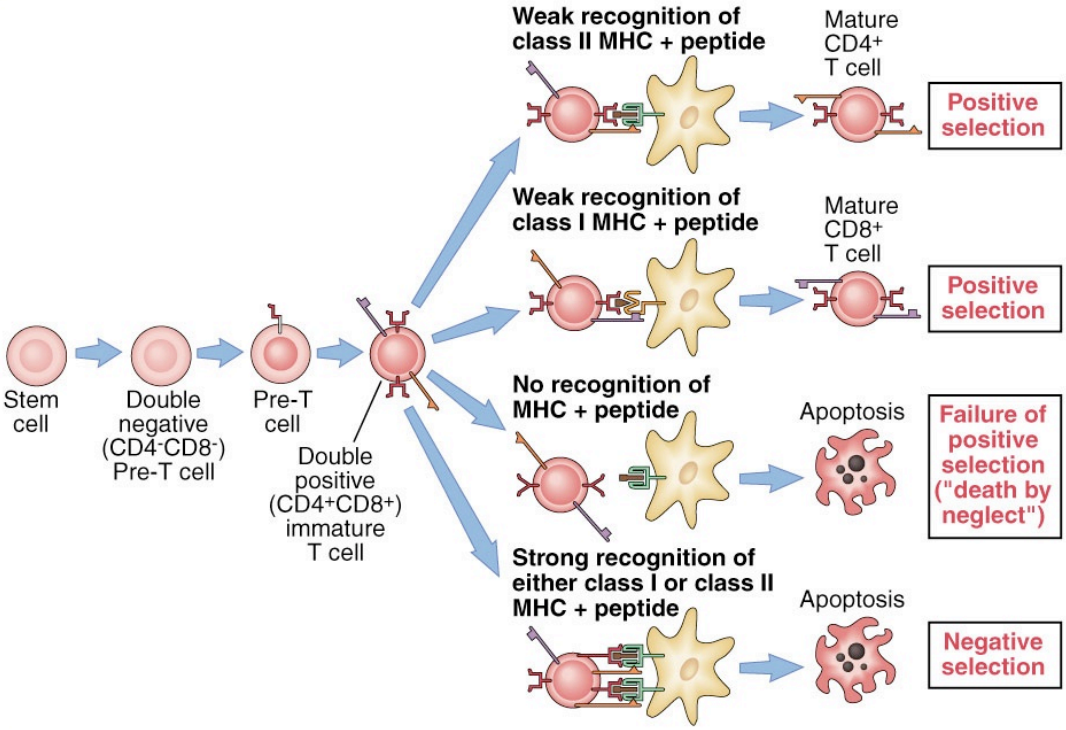


Figure 1.4: **Positive and negative selection in the thymus.** Double positive T cells expressing both the CD8 and CD4 co-receptor are presented to class I and II MHC molecules in complex with self-peptides. **Positive selection:** T cells recognizing peptides bound to MHC class I molecules differentiate into CD8+ T cells, while T cells recognizing peptides bound to MHC class II molecules differentiate into CD4+ T cells. T cells, which do not recognize any self-peptide/MHC complexes undergo apoptosis. **Negative selection:** T cells interacting too strongly with self-peptide/MHC complexes are negatively selected and induced to undergo apoptosis. Elsevier illustration used with permission from Elsevier. All rights reserved.

T cell activation

T cells, which are newly released from the thymus, are called naive T cells. They migrate into the lymph nodes, where they await to be activated. If a tissue becomes infected with a pathogen, dendritic cells are activated and pick up antigens from the site of infection, before migrating to the lymph nodes. They then present the antigen to the naive T cells together with a co-stimulatory signal. If the antigen/MHC complex is recognized by the naive T cell, the T cell differentiates into an effector T cell; a CTL in the case of CD8+ cells or a T helper cell in the case of CD4+ cells. If the co-stimulatory signal is missing, the naive T cell instead becomes anergic (inactivated), since the antigen is then unlikely to originate from a pathogen.

Activated CTLs then undergo clonal expansion and are released into the general circulation ready to identify and induce apoptosis in any cell expressing its antigen/MHC complex. T helper cells do not kill other cells but assist in the activation of both B cells and CTLs. When the infection is cleared, most of the effector cells die by apoptosis, while some differentiate into memory T cells (CD8+ or CD4+). The memory T cells specific to the antigen are present for the rest of the individual's life at a level 100-1000 fold above the frequency of the original pool of naive T cells specific to the antigen (Murphy et al.). If the same antigen is encountered again, the memory T cells quickly mount an immune response, more efficient than that of the primary infection.

1.2 Immunological bioinformatics

An important application of bioinformatics in immunology is the prediction of T cell epitopes from a given protein and the corresponding HLA restriction. Traditional experimental epitope identification is tedious, and using prediction methods to guide the search for the optimal epitope within a longer peptide or protein, can reduce the experimental workload significantly. Furthermore, epitope predictors have paved the way for the field of reverse immunology which was pioneered by Rappuoli (2000) a decade ago, where it was used to identify vaccine candidate epitopes against serogroup B meningococcus (Pizza et al., 2000). Normally, the starting point of traditional epitope discovery is an isolated CTL clone recognizing an unknown disease epitope, and the experimental task is then to identify this epitope as well as the presenting HLA molecule, possibly guided by predictions when the antigen has been localized to a small genomic region. In reverse immunology, the starting point could instead be the entire genome of a pathogen, from which epitope candidates are predicted and subsequently tested experimentally for T cell recognition.

Currently, the publicly available epitope predictors are very useful for predicting MHC class I epitopes, while MHC class II predictors are still relatively inaccurate. The predictor *NetMHCpan* (Hoof et al., 2009; Nielsen et al., 2007) has been developed in the Immunological Bioinformatics group at CBS and is considered one of the most accurate class I predictors (Lin et al., 2008a; Zhang et al., 2009). *NetMHCpan* can predict ~74% of actual epitopes, while the most accurate class II predictor *NetMHCIIpan* (Nielsen et al., 2008) can only predict ~50% of actual epitopes (Lin et al., 2008b). Throughout this thesis, only class I predictions are used, and all predictions are done with *NetMHCpan*.

1.2.1 Artificial neural networks and *NetMHCpan*

NetMHCpan can predict peptide/MHC binding for all HLA-alleles with a known sequence by means of an artificial neural network (ANN) that has been trained on experimental peptide binding data for a subset of the HLA alleles and the primary sequence of the HLA molecules. *NetMHCpan* is the successor of *NetMHC* (Nielsen et al., 2003), which only predicts binders restricted to the alleles used for training the method. An ANN is a machine learning technique suitable for solving non-linear problems by means of interconnected layers of neurons or computational units, typically an input layer, one or more hidden layers, and an output layer (see Figure 1.5). The binding of a 9mer peptide in the binding groove of the MHC molecule as illustrated in Figure 1.6, is an example of a non-linear problem, since the binding strength is not simply a sum of the binding forces on each of the 9 amino acids. Exchanging a small amino acid with a larger one can, for instance, mean that the peptide will no longer fit in the binding groove, while exchanging two small amino acids with one large can have very little effect on the binding.

An ANN like *NetMHC* is trained for each HLA allele by presenting it to amino acid sequences (the input layer) of peptides with known binding affinity (the output layer). This is done in several rounds continuously allowing the connections or weights between the neurons to adjust, such that their final calculation of the binding affinity approaches the true value. Usually, the training set is divided into n subsets, and in each training round, the network is trained on n-1 subsets. After each training round the performance of the network is tested on the last subset, and to avoid overfitting, the training is stopped when the performance reaches its maximum.

Finally, the network is presented to peptides on which it has not been trained, in order to evaluate the predictive performance on unknown peptides. In practice, this is done by a so-called leave-one-out cross validation where the training, described above, is done several times with all but one peptide. The network has then in practice never seen the peptide, it is validated on. An average prediction performance of the predictor can then be calculated.

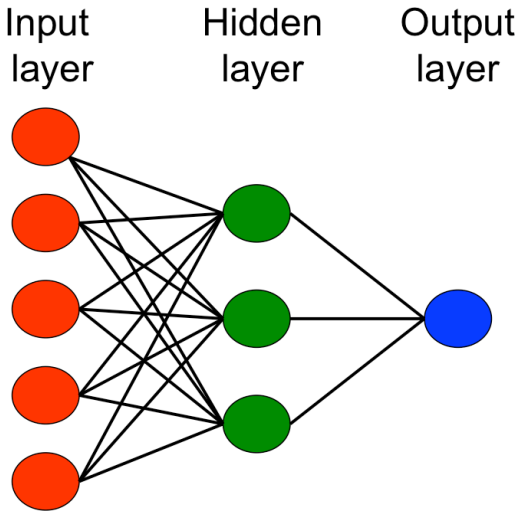


Figure 1.5: **Simplified schematics of an ANN.** The neurons between the layers are assigned weights during the training of the network. Once trained, the ANNs of the *NetMHC* predictor can predict the binding affinity (corresponding to the output layer) when given an unknown amino acid sequence (corresponding to the input layer).

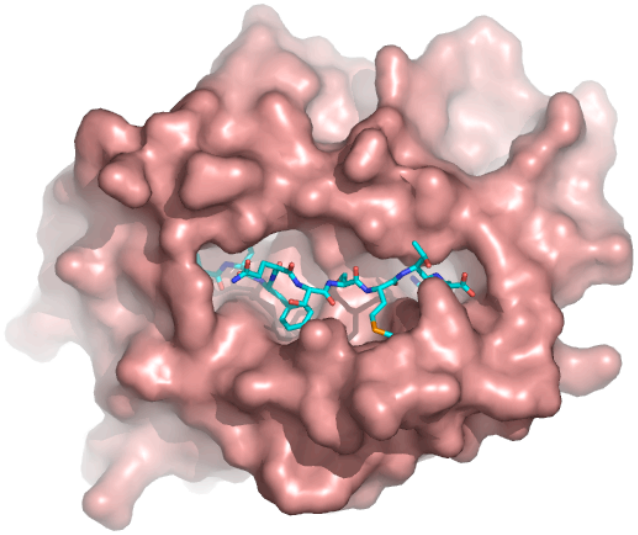


Figure 1.6: **MHC molecule binding a 9mer peptide.** PDB structure 1DUZ (Khan et al., 2000). The figure illustrates how the 9mer peptide (LLFGYPVYV) fits into the binding groove of the MHC molecule (HLA-A*0201).

After the training is complete, the ANN described above can predict the binding affinity of a new peptide to the same HLA allele. *NetMHC* thus consists of one trained ANN for each of the alleles where sufficient training data was available. However, the diversity of the HLA system means that many alleles are then not covered. This problem was solved with *NetMHCpan*, which is also trained on the protein sequence of the MHC molecule itself, thus transferring information from MHC molecules with available binding data to MHC molecules sharing sequence similarity with these.

Binding thresholds

When given a peptide and an HLA molecule, *NetMHCpan* predicts the binding strength between the two, measured by the IC50 binding affinity in nM and a %random value. The binding affinity is the quantitative measure, corresponding to the input data on which *NetMHCpan* has been trained. It denotes the concentration of peptides needed for occupying half of the HLA molecules and thus a small IC50 value corresponds to strong binding. It is generally accepted that immunogenic peptides are characterized by an affinity threshold of 500 nM (Assarsson et al., 2007; Sette et al., 1994). However, for a number of alleles, there is little available binding affinity data, which means that *NetMHCpan* can only predict for these alleles with limited accuracy and therefore does not, as standard, report any binding affinity value.

The %random value is calculated for all alleles and corresponds to the predicted binding strength of the query peptide compared to the predicted binding strengths of 1 million random peptides to the chosen HLA molecule. Thus a %random value of 1 means that for the chosen HLA molecule, only 1% of random peptides are predicted to have a stronger binding than the query peptide. For some alleles, this is likely to be a more precise measure of a peptide's binding capabilities, as it has been suggested that not all HLA molecules present peptides at the same binding threshold (Rao et al., 2009; Stranzl et al., 2010). Defining a fixed percentage threshold is, however, equally artificial, as it is also unlikely that different alleles will bind the same percentage of random peptides.

In reality, each HLA molecule probably has its own binding affinity threshold corresponding to a unique percentage of random peptides. At present, however, such individual thresholds cannot be determined for each of the more than >3500 known HLA alleles (Robinson et al., 2009), and thus it is necessary to use either a nM threshold or a % random threshold as a compromise.

In this thesis, the standard 500 nM threshold was used in Chapters 2 and 4 for the selection of peptides to purchase for validation experiments. In Chapters 3 and 5, %random thresholds between 0.5 and 2 were used, while the *HLArestrictor* method described in Chapter 5 was designed to work with both kinds of thresholds in its standard setting.

1.2.2 Other predictors

Some of the earlier prediction tools, still being frequently used by experimentalists, are the *BIMAS* (Parker et al., 1994) and *SYFPEITHI* (Rammensee et al., 1999) methods. They are both weight-matrix-based, meaning that they use a position specific scoring matrix (PSSM) (Altschul et al., 1997) to calculate how well each of the 8, 9, 10, or 11 amino acids of a peptide fits with the binding specificity observed for a given HLA molecule. For a 9mer for instance, a PSSM is a 20×9 matrix which for each of the 9 amino acid positions assigns a score for each of the 20 different amino acids based on how frequently the amino acid is observed in the given position among peptides known to bind the HLA molecule. A prediction score is then calculated by summing up the scores for a given peptide. These kinds of approaches cannot capture non-linear effects as described in Section 1.2.1 and are outperformed by non-linear predictors (Lin et al., 2008a; Lundegaard et al., 2010).

Binding predictions for MHC class II molecules are still much less accurate than for MHC class I (Nielsen et al., 2010), which is the reason why MHC class II binding predictions are not considered in this thesis. Producing reliable predictions is mainly complicated by the fact that MHC class II molecules are open at both ends, allowing the ends of a binding peptide to extend outside of the binding groove. Therefore, identifying the binding core of a peptide is difficult. One of the most used class II predictors is *TEPITOPE* (Hammer et al., 1994). Again, this is a non-linear predictor outperformed by the ANN based *NetMHCIIpan* predictor (Lin et al., 2008b; Nielsen et al., 2010).

Some predictors include other parts of the antigen presentation pathway. Some of the most accurate of these are *MHCpathway* (Tenzer et al., 2005) and *NetCTL* (Larsen et al., 2005), which integrate predictions of MHC class I binding, C-terminal proteasomal cleavage, and TAP transport efficiency (Lundegaard et al., 2007). A pan-specific version of the latter, *NetCTLpan* (Stranzl et al., 2010), has recently been developed. While these different steps in the pathway are all important to define the final epitopes, it has been shown that in most cases, the ability of the integrated methods to predict epitopes do not outperform the prediction methods based solely on MHC class I binding prediction (Stranzl et al., 2010).

1.3 Experimental epitope validation

Various methods exist for *in vitro* validation of T cell epitopes. The 3 methods described in the following are some of the most commonly used. The enzyme-linked immunospot assay (ELIspot) and the intracellular cytokine staining (ICS) methods are useful for examining many peptides simultaneously, often in the form of 15mers, which are processed by antigen presenting cells prior to the procedure. Validation by tetramers is more specific, as information about the exact peptide/MHC complex and T cell is retained.

1.3.1 ELIspot

The ELIspot technique was developed by Czerkinsky et al. (1983) and is used to measure various immune responses including antigen specific T cell responses. As illustrated in Figure 1.7, T cells are incubated with antigens on an ELIspot plate where they, if activated, will produce cytokines such as interferon-gamma (IFNγ). The surface of the ELIspot plate is coated with cytokine specific antibodies which bind the cytokines. After washing, a second biotinylated antibody is added along with some additional chemicals to reveal the cytokine production from each single T cell as one spot on the plate. The spot density of each well is then examined to assess whether the T cells responded to antigen(s) in the well. Often, each well on the ELIspot plate will be loaded with a pool of around 10 peptides and the pool contents will be designed using a matrix such that each peptide will be found in exactly two wells on the plate. This procedure allows for a unique identification of the antigenic peptides, while saving time and materials. If spots are observed in more than two, say four, wells within a matrix design, an extra assay must be made to identify the two peptides causing the response.

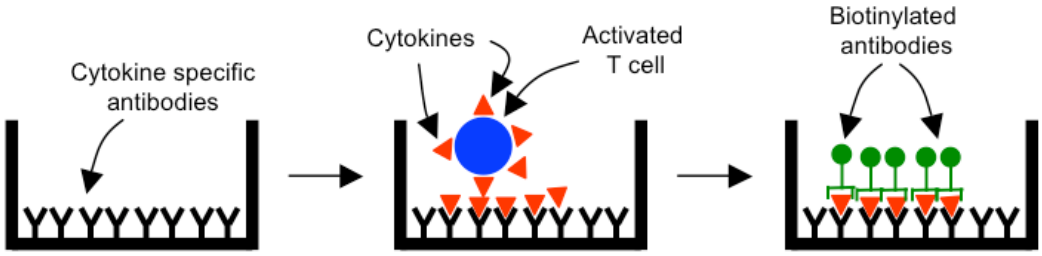


Figure 1.7: **The ELIspot assay.** The surface of a plastic well is coated with cytokine specific antibodies. T cells incubated with antigens are added. Activated T cells will secrete cytokines which bind the cytokine specific antibodies. T cells are washed away and biotinylated antibody is added. More chemicals are added (not shown) to reveal the bound cytokine as spots on the plate.

1.3.2 Intracellular cytokine staining

ICS is another method which allows the monitoring of cytokine production in T cells by multicolor staining of intracellular cytokines followed by flow cytometry (Murphy et al.). Whereas in the ELIspot assay, the T cells eliciting a cytokine response are washed away during the process, ICS allows for the identification of the exact T cell responsible for the response. Compared to ELIspot, ICS can monitor several different markers simultaneously on a single cell level. Also, it can distinguish the T cells by staining with antigens specific to CD4 and CD8. In ICS, the cells are chemically treated to inhibit the secretion of cytokines, which are therefore accumulated inside the cells. The cells are then fixed and permeabilized so that fluorochrome labeled antibodies can bind to the cytokines. These antibodies can then be detected by flow cytometry.

1.3.3 Tetramers

Tetramer stainings are even more precise than ICS, since they are used to verify the exact HLA molecule and peptide sequence recognized by a T cell. Labeling T cells with their specific peptide/MHC complex is difficult, since the affinity between the TCR and the peptide/MHC complex is often too low. In order to boost the sensitivity, a method was developed by Altman et al. (1996) by which four peptide/MHC complexes labeled with biotin are bound to the tetrameric molecule streptavidin as shown in Figure 1.8.

The resulting tetramer can simultaneously bind to several TCRs on a T cell, thus forming a more stable bond. The streptavidin molecules are coupled to fluorescent dye molecules allowing for the T cell to be monitored and quantized by flow cytometry (Murphy et al.). The T cells can also be stained with antibodies specific to CD8, CD4 or CD3 molecules, providing information about the nature of the responding T cells (See Figure 1.9).

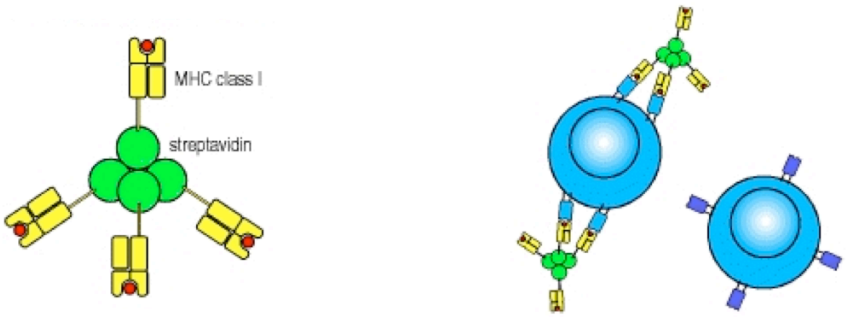


Figure 1.8: **Left: The tetramer molecule.** The tetrameric molecule streptavidin binds 4 peptide/MHC molecules to which biotin groups have been added. **Right: Tetramers binding a T cell.** T cells expressing a TCR specific to the peptide/MHC molecule can more easily form a stable bond with the tetramer compared to a single peptide/MHC molecule. T cells with a different specificity do not bind the tetramers. From (Murphy et al.).

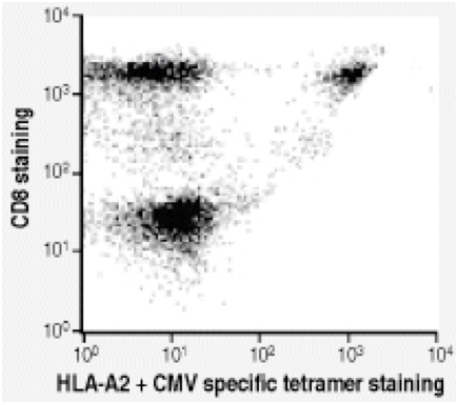


Figure 1.9: **A typical tetramer stain plot.** Apart from the tetramer staining, T cells are typically stained with antibodies specific to CD4 or, as here, CD8. In this case, only the T cells corresponding to the dots in the top right corner recognize the peptide/MHC complex (the tetramer staining is read on the X-axis). These are all CD8+ T cells (CD8 staining is read on the Y-axis) and as expected, none of the CD8- cells (in the lower part of the plot) recognize the peptide/MHC class I molecule. From (Murphy et al.).

1.4 Hematopoietic cell transplantation

HCT is a widely used treatment applied in a number of malignant diseases such as acute myelogenous leukemia, chronic myelogenous leukemia, multiple myeloma, Hodgkin lymphoma, or non-Hodgkin lymphoma. HCT can also be used for curing non-malignant hematologic diseases such as aplastic anemia. All patients studied in this thesis were treated for malignant diseases. Prior to a HCT, hematopoietic cells from a suitable donor are either procured directly from the bone marrow, from umbilical cord blood, or, more commonly, harvested as peripheral blood stem cells from the donor's blood. The latter is done using a process called apheresis, which filters out the leukocytes. Prior to this procedure, the donor is treated with granulocyte colony stimulating factor, which promotes the release of stem cells from the bone marrow to the peripheral blood (Soiffer). The hematopoietic cells, also called the graft, are then injected into the blood stream of the patient. Here, they find their way to the bone marrow in a process called homing (Soiffer). The patient is prepared for the transplantation by one of the two following methods.

1.4.1 Myeloablative conditioning

Myeloablative conditioning (MC) is the traditional high dose conditioning regimen. It usually consists of a combination of chemotherapy and total body irradiation in doses high enough to eradicate the cancerous and myeloid cells (See Figure 1.1) as well as suppress the patient's own immune system prior to the transplantation. Thus ablating the patient's myeloid cells, the conditioning would be lethal without the replacement of hematopoietic cells from the donor. The advantages of this treatment is that relapse of the malignant disease is mostly prevented, and that the graft from the donor is well received because of the immunosuppression. However, the side effects of the conditioning are severe, which makes the treatment unavailable to elderly (above 50-60 years) or medically infirm patients due to a high risk of treatment related mortality (TRM) (Rowe et al.).

1.4.2 Non-myeloablative conditioning

Non-myeloablative conditioning (NMC) is a newer, reduced intensity conditioning regimen. It eradicates neither the myeoloid nor the cancerous cells, while still providing an immunosuppressive effect. The method instead relies on the graft-versus-tumor (GVT), also called graft-versus-leukemia, effect for eradicating the malignant cells as well as the immune system of the patient. The GVT effect results from the recognition of the patient's tumor cells as non-self by the donor's CTLs and is coupled to the undesirable complication GVHD described below. NMC is associated with an increased relapse rate and is therefore mainly used in elderly or medically infirm patients, who cannot bear the side effects of the high dose treatment. On the positive side, the incidence of TRM is lower after NMC compared to MC. The different patient demographics of MC patients vs. NMC patients make it difficult to compare the progression free survival (PFS) and overall survival (OS). However, several studies (Aoudjhane et al., 2005; Hari et al., 2008; Kim et al., 2006; Storb, 2007; Valcarcel et al., 2005) find that OS is comparable between the two treatments and that the higher relapse incidence in NMC patients is balanced by the lower TRM. Other studies conclude that OS is higher with the high dose treatment (Mengarelli et al., 2002).

1.4.3 GVHD and the GVT effect

In GVHD, healthy patient cells are perceived as non-self by the donor immune system and are therefore attacked. Tissues typically affected are liver, skin and the gastrointestinal tract. GVHD is traditionally divided into acute GVHD (aGVHD) occurring within the first 100 days after transplantation, and chronic GVHD (cGVHD) occurring after the first 100 days. Standard criteria are typically used for grading acute and chronic GVHD. aGVHD is divided into grades I-IV where, by definition, grade I does not require treatment and grade IV is fatal (Przepiorka et al., 1995; Sullivan, 2003). cGVHD is classified as limited or extensive according to the Seattle criteria (Lee et al., 2003; Shulman et al., 1980). New classification criteria have recently been developed (Filipovich et al., 2005) introducing a clinical scoring system for each organ involved in cGVHD and taking into consideration that aGVHD can occur after 100 days, especially after NMC treatments (Mielcarek et al., 2003).

GVHD is linked to the GVT effect in which the malignant patient cells are perceived as non-self by the donor CTLs. The coupling between the two effects has been observed in experiments with T cell depletion, where T cells are removed from the graft prior to transplantation. This procedure eliminates the risk of GVHD, but increases the risk of relapse due to the missing GVT effect (Soiffer, 2003).

The relation between GVHD and the GVT effect can also be illustrated by the observation that the graft source is of particular relevance after NMC as the GVT effect is carried out by the CTLs, which are more abundant in peripheral blood stem cells compared to bone marrow. Higher PFS and OS has been observed in patients receiving peripheral blood stem cells after NMC whereas bone marrow is associated with more relapse (Maris et al., 2003). Also, after MC HCTs, peripheral blood stem cells have been associated with higher incidences of acute and chronic GVHD in a meta-analysis by Cutler et al. (2001). In another meta-analysis by al Jurf et al. (2005), a decreased relapse rate, higher PFS and OS, and increased risk of cGVHD was observed. Thus there is a balance between the risk of relapse due to a lack of the GVT effect and the risk of GVHD, which is especially important in NMC HCT, where the eradication of the leukemia relies on the donor CTLs.

1.4.4 Donor matching

The HLA loci most relevant for HCTs are the HLA-A, -B, -C, -DR and -DQ. As each person has two copies of each gene, this represents 10 possible variations in a given patient. If all 10 alleles are matched by the chosen donor, the patient-donor pair is said to be 10/10 matched, which is considered to be the optimal situation. Transplantations are, however, routinely performed with 9/10 matched or even down to 5/10 matched donors, the latter being the case in haploidentical transplantations from a related donor, which can be used if a perfect match is not available, or the need for transplantation is urgent (Koh and Chao, 2008). Within the last 20 years, HLA donor matching has improved significantly, as the donor registers have expandend and HLA typing is now done using DNA based techniques (Karanes et al., 2008). As all patients studied in this thesis are 10/10 matched, the focus here will be the fully matched situation. In siblings, there is a 25% chance of a 10/10 match (Welniak et al., 2007), since the HLA alleles are often inherited as haplotypes due to the fact that the whole MHC region only spans a less than 5 Mb long region on chromosome 6 with a high degree of linkage disequilibrium (LD) (Walsh et al., 2003). The concept of LD, is described in Section 1.6.2. If no fully matched sibling is available, a matched unrelated donor (MUD) can be used. Unrelated donors are found through national and international donor registries such as *Bone Marrow Donors Worldwide*, which currently

lists more than 14 million donors. Even with that many donors to choose from, a perfect match cannot always be made, as the chance of succes also depends on how rare the patient's HLA types are (Tiercy et al., 2007).

1.5 Minor histocompatibility antigens

The fact that alloreactivity can occur, even in 10/10 matched allogeneic HCT, was discovered in mice as early as 1956 by Barth et al. The antigens hypothesized to be involved in this mechanism were named minor histocompatibility antigens (mHags). Today it is known that mHags result from genetic disparities between patient and donor (Hambach et al., 2007). In a fully matched allogeneic HCT, the HLA molecules of the patient are identical to those on which the incoming donor CTLs were trained in the thymus of the donor. Still, due to genetic differences in the rest of the donor and patient proteomes, the donor CTLs might recognize, in complex with the patient's HLA molecules, peptides not encountered during their thymal training, and thus perceived as non-self. As illustrated in Figure 1.10, this leads to the induction of apoptosis in patient cells and thereby the GVT effect and GVHD. The disparate, immunogenic self-peptides, the mHags, mostly result from nonsynonymous single nucleotide polymorphisms (nsSNPs) in autosomal genes. However, mHags may also be caused by gene deletions, genetic variation in noncoding regions affecting gene transcription, or the presence of Y chromosome-encoded proteins in sex-mismatched HCT (Kawase et al., 2007; Mullally and Ritz, 2007; Murata et al., 2003; Spaapen and Mutis, 2008). mHags with a broad tissue expression, especially in those tissues in which GVHD is most commonly observed, may induce GVHD (Akatsuka et al., 2003b; Goulmy et al., 1996; Perez-Garcia et al., 2005), whereas

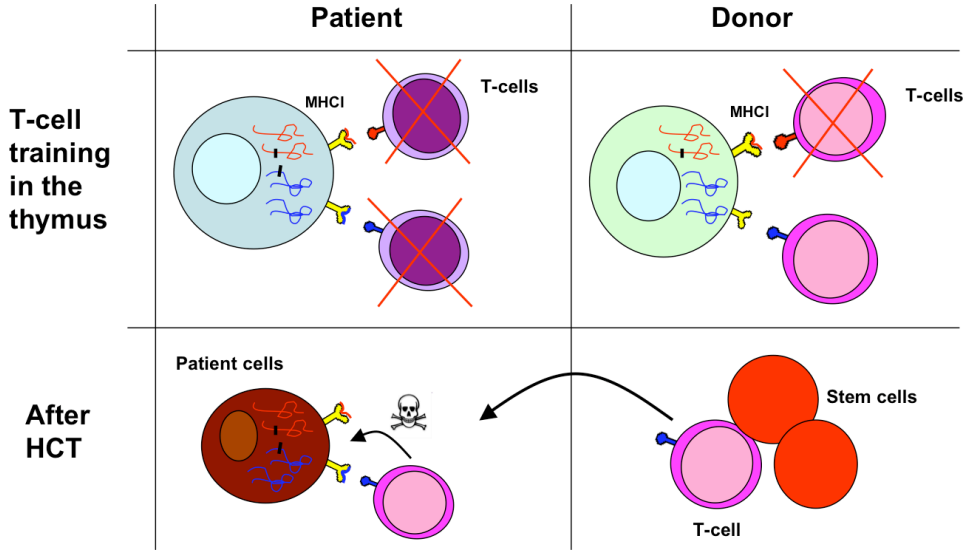


Figure 1.10: **CTL response to mHags - the GVHD and GVT effect.** T cells are continuously trained in the thymus and deleted if they recognize self antigens. If a nsSNP, present in the patient but not in the donor, results in the recognition of a self antigen, the corresponding T cell is deleted in the patient only. After the HCT, the donor CTL recognizes the patient self antigen, the mHag, as non-self and initializes apoptosis in the patient cell. Figure courtesy of Mette V. Larsen.

mHags, which are expressed only in hematopoietic tissue, may induce the GVT effect (de Rijke et al., 2005; Stumpf et al., 2009; van Bergen et al., 2007). The presence of mHag-specific CTLs posttransplantation is also involved in graft rejection (Goulmy et al., 1977; Voogt et al., 1990). Here the situation is turned, such that the patient's CTLs recognize mHags presented by the infused donor cells and induce apoptosis in these.

Self-peptides with the ability to bind to the HLA molecules are not equally distributed throughout the human genome. In a recent study by Juncker et al. (2009), it was shown that human proteins are more likely to contain MHC ligands if they are localized in the intracellular parts of the cell, including the cytoplasm and nucleus. Also, the study showed, that proteins with a higher mRNA expression level more often contain MHC ligands as shown in Figure 1.11. These observations make it probable that also mHags are more likely to be found within these intracellular, highly expressed proteins.

1.5.1 Identification of mHags

The database, dbMinor (Spierings et al., 2006), lists approximately 30 mHags as shown in Table 1.3. Since 2006 several more mHags have been identified, such that around half a hundred mHags are known to date. A commonly used method of mHag identification is peptide elution from HLA class I molecules with subsequent fractioning by high performance liquid chromatography. The immunogenic peptides identified in this way are then sequenced by mass spectrometry (Bleakley and Riddell, 2004; Brickner et al., 2001, 2006; den Haan et al., 1995, 1998; Meadows et al., 1997; Pierce et al., 1999, 2001; Spierings et al., 2003a; van Bergen et al., 2007; Wang et al., 1995).

Another common identification method is expressional cloning, where RNA, from a cell known to express the mHag, is used to prepare cDNA, which is then cloned into an expression vector. The vector is then transfected into a cell together with a plasmid encoding the restricting HLA allele. Those of the cells that stimulate CTLs specific to the mHag are then positive for the cDNA encoding the mHag. The exact sequence of the immunogenic peptide is then identified by repeating the procedure with truncated versions of the cDNA or by using predictions. (Bleakley and Riddell, 2004; Dolstra et al., 1997, 1999; Kawase et al., 2007; Murata et al., 2003; Rosinski et al., 2008; Spierings et al., 2003b; Stumpf et al., 2009; Terakura et al., 2007; Vogt et al., 2000b, 2002; Warren et al., 2000, 2006).

In a few cases, bioinformatics methods have been used for the identification of mHags. One example is Mommaas et al. (2002), where the *BIMAS* predictor was used to search for additional mHags containing the same nsSNP as the already known HLA-A2 restricted mHag HA-1 (VLHDDLLEA). In this way, the neighboring HLA-B60 restricted mHag HA-1/B60 (KECVLHDDL) was identified.

A method similar to expressional cloning uses genetic linkage analysis to identify mHags. In genetic linkage analysis, pedigrees are analyzed, to identify genomic regions associated with a specific phenotype, here, the presence of a specific, unidentified mHag capable of eliciting a CTL response. In this approach, only the restricting HLA allele is transfected into cell lines. Genetic loci specific to the cell lines eliciting a CTL response, are then identified by linkage analysis (Akatsuka et al., 2003a; de Rijke et al., 2005).

Recently, genome-wide association analysis (GWAS) has been used to identify mHags. Shortly, a GWAS uses SNP microarrays to identify associations between SNP variants and phenotypes in cohorts of non-related individuals. Using a GWAS, disease related mutations, or in this case mHags, can be located to small genomic regions with a high degree of LD, represented by so-called tag SNPs on the microarray. Kawase et al. (2008) used a panel of

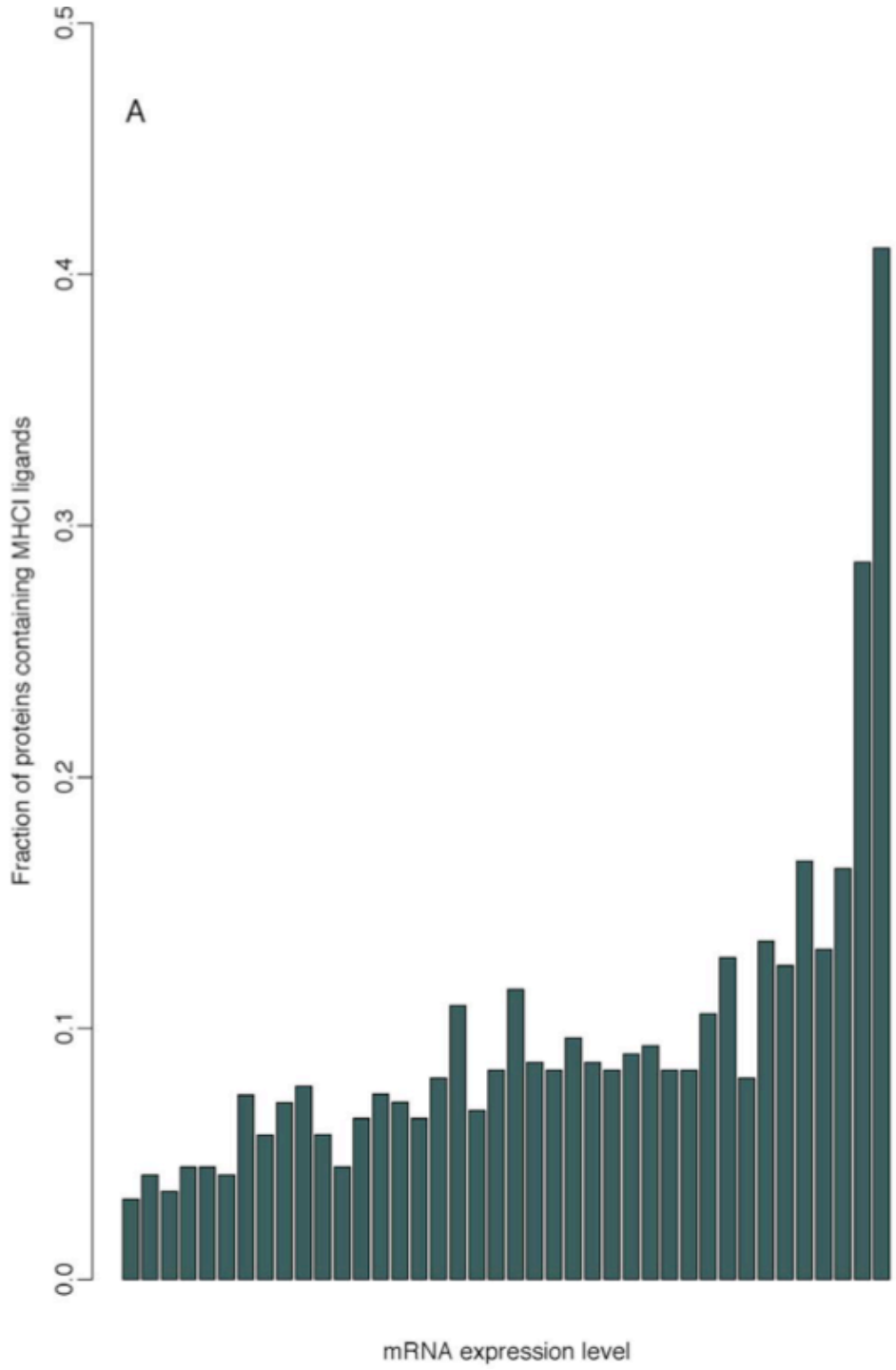


Figure 1.11: **Expression level of MHC ligands.** An equal number of proteins is grouped into each bin on the X-axis and sorted by the mRNA expression levels of the proteins within hematopoietic tissue cells. The height of each bar indicates the fraction of proteins, with a given expression level, which contain MHC class I ligands. It is seen that the fraction increases with the expression level of the proteins. Of the 2.5% proteins with the highest expression level, corresponding to the rightmost bar, as many as 41% contain MHC ligands. From (Juncker et al., 2009).

mHag†	HLA restriction	Peptide sequence‡	mHag gene††
mHags encoded by genes on autosomal chromosomes			
HA-2^{V}	A*0201	YIGEVLVSV	***MYO1G***
HA-1^{H}	A*0201	VLHDDLLEA	***HMHA1***
HA-1^{H}	B60	KECVLHDDL	***HMHA1***
HA-1^{H}	A*0206	VLHDDLLEA	***HMHA1***
HB-1^{H}	B44	EEKRGSLHVW	***HMHB1***
HB-1^{Y}	B44	EEKRGSLYVW	***HNHB1***
HA-8^{R}	A*0201	RTLDKVLEV	*KIAA0020*
HA-3^{T}	A1	VTEPGTAQY	*AKAP13*
UGT2B17	A29	AELLNIPFLY	*UGT2B17*
ACC1^{Y}	A24	DYLQYVLQI	***BCL2A1***
ACC2^{D}	B44	KEFEDDIINW	
LRH-1	B7	TPNQRQNVC	***P2RX5***
CTL-7A7^{R}	A3	RVWDLPGVLK	***PANE1***
ACC-5^{R}	A*3101	ATLPLLCAR	*CTSH*
ACC-4^{R}	A*3303	WATLPLLCAR	
RDR173^{H}	B7	RPHAIRRPLAL	***ECGF1***
DNR-7^{R}	A3	SLPRGTSTPK	***SP110***
LB-ADIR-1^{F}	A*0201	SVAPALALFPA	***TOR3A***
ACC-6	B44	MEIFIEVFSHF	***HMSD***
mHags encoded by X-homolog genes on Y chromosomes			
SMCY	B7	SPSVDKARAEL	*JARID1D*
SMCY	A*0201	FIDSYICQV	*JARID1D*
DFFRY	A*0101	IVDCLTEMY	*USP9Y*
UTY	B8	LPHNHTDL	*UTY*
UTY	B60	RESEEESVSL	*UTY*
DBY	DQ5	HIENFSDIDMGE	*DDX3Y*
DBY	DRB1*1501	GSTASKGRYIPPHLRNREA	*DOX3Y*
RPS4Y	DRB3*0301	VIKVNDTVQI	*RPS4Y1*
RPS4Y	B*5201	TIRYPDPVI	*RPS4Y1*
ACC-3	A*3303	EVLLRPGLHFR	*TMSB4Y*

Table 1.3: **Currently known mHags**. The disparate amino acids are underlined in the peptide sequences. From (Akatsuka et al., 2007).

approximately 100 cell lines expressing the HLA-A*2402 molecule. Based on whether or not the cells were recognized by the isolated CTL clone, they performed a genome wide association analysis to identify the genomic region encoding the corresponding mHag. In a similar approach, Kamei et al. (2009) used HapMap cell lines already genotyped. In both cases the *BIMAS* predictor was used to identify the exact sequence of the mHags.

All the mHag identification methods mentioned above start with a CTL clone specific to an unknown mHag. By various means, the corresponding mHag is then identified. This traditional approach to mHag discovery is time-consuming and only identifies few mHags restricted to the more common HLA alleles. Considering the diversity of the HLA system with >3,500 known alleles (Robinson et al., 2009), as well as the >120,000 known allelic nsSNP variants (UniProt, 2009), it seems likely that many mHags have yet to be identified. Bioinformatics methods can address this task in a high-throughput way by integrating SNP data, tissue-specific gene expression, and predictions of peptide/MHC binding as described in more detail in Chapter 4. Applying reverse immunology to mHag discovery is not as simple as in disease epitope discovery, due to the much larger genome of humans and the need for genotyping patients and donors individually to identify the genetic disparities. In this thesis, reverse immunology is applied on a small scale, considering only a fraction of the genome, however, as the cost of genotyping continues to drop, applying full-scale reverse immunology to mHag discovery is within reach.

1.5.2 Adoptive immunotherapy

In the event of relapse after a transplantation, donor lymphocyte infusions can be given in order to establish the GVT effect as decribed by Kolb et al. (1995) and Collins et al. (1997). It has been shown that the GVT effect after donor lymphocyte infusion is caused by donor CTLs specific to mHags expressed uniquely by the patient (Marijt et al., 2003). A promising perspective of these findings is to expand *in vitro* and infuse only those CTLs that are specific to mHags from genes with a restricted hematopoietic tissue expression. This would selectively induce the GVT effect while avoiding GVHD. Until now, only transient remission of relapsed leukemia has been observed in patients treated with such mHag specific CTLs (Riddell et al., 2006; Warren et al., 2010), while durable remission has been shown in mice (Fontaine et al., 2001). The treatment of other cancers with mHag specific CTLs is also being investigated, and promising results have been obtained in the treatment of metastatic melanoma (Dudley et al., 2002).

However, a general problem associated with such treatments is that the healthy tissue is likely to be attacked by the infused CTLs as well. Unless mHags expressed uniquely in tumors are discovered, the method will thus mainly be applicable in the treatment of cancers in tissues not necessary for survival, such as breast or prostate tissue. In leukemia treatment, donor CTLs specific to hematopoietically restricted patient mHags will ideally be infused after an allo-HCT. Thus the destruction of the healthy patient hematopoietic tissue by the CTLs is acceptable, as the graft from the donor should already have taken over (Spaapen and Mutis, 2008). As only a few of the mHags known today are therapeutically relevant, more mHags need to be discovered, which are sufficiently expressed in hematopoietic tissue only and presented by HLA alleles that are frequent in the population. Further, mHag candidates should ideally have a population frequency between 26% and 78% optimizing the probability that the mHag is present in the patient and absent in the donor (Spaapen and Mutis, 2008).

1.6 Statistical methods

Various statistical methods are applied in this thesis, especially in Chapter 3, where the influence of different parameters on patient outcome is investigated. Here, an overview of the statistical methods is given, to provide a background for the choice of methods. Most of the statistical analyses were carried out in the statistical computing language R (www.r-project.org).

1.6.1 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium (HWE) assumption is a basic equation in population genetics concerning the distribution of genotypes (Hartwell et al.). If two alleles, A and a of a gene are present in a population with allele frequencies q and r, respectively, then the three different genotypes AA, Aa and aa will have frequencies q^2, $2qr$, and r^2 which should sum to 1

$$q^2 + 2qr + r^2 = 1. \tag{1.1}$$

This equation assumes a large, closed population of randomly mating individuals where the genotypes in question have no influence on the survival. In this thesis, it is assumed that patients and donors should adhere to the HWE, and the equation is used as a control of the quality of genotyping data.

1.6.2 Linkage disequilibrium

Two genetic loci are said to be in linkage disequilibrium (LD) if they are not inherited independently of each other (Barnes). If, for instance, two SNPs are located close to each other on the same chromosome, they are likely to be inherited together. A common measure of LD is the correlation coefficient r^2, where the value $r^2 = 1$ means that the two SNPs are in complete LD and thus always inherited together. In such cases, it is only necessary to genotype an individual for one of the two SNPs, since the other SNP variant is then given. A haplotype defines a number of SNPs in LD, between which recombination is unlikely to occur in related individuals, an example being the human MHC region.

1.6.3 Fisher's exact test

Fisher's exact test is useful when dealing with relatively small sample sizes (Altman). It is used to compare frequencies between groups, for example frequencies of certain genotypes observed in a group of patients compared to a group of donors, as exemplified in Table 1.4. The null hypothesis is then that there is no frequency difference between the two groups. The biological interpretation in this case, is that the genotype of an individual does not affect the chance of acquiring a hematological disease.

1.6.4 Kaplan Meier

In survival analysis, a Kaplan Meier curve is a useful tool for estimation of the probability of surviving past a certain time, for example the probability that a patient is still alive a certain number of years after an allo-HCT. The Kaplan Meier estimator uses life-time data to estimate the survival probability as function of time (Altman). Such data is usually censored, meaning that not all patients are followed up to a given time. If a patient is censored before the given time, the outcome after censoring is unknown. However, the fact that the patient survived

Genotype	Patients	Donors
GG	102	106
AG	21	15
AA	2	4

Table 1.4: **Example of Fisher's exact test**. The table denotes the number of patients vs. donors having the 3 different genotypes for SNP rs9061. Fisher's exact test gives P = 0.45 for this table, confirming the null hypothesis that there is no statistical difference in genotype distribution between patients and donors.

until the date of censoring, represents important information and thus the patient should not be excluded from the analysis.

The Kaplan Meier method takes censoring into account when estimating the survival probability. The principle of the method is to consider each time point t_j, where an event (here a death) takes place. If the proportion of patients, surviving $t_j - 1$ days is p_{j-1}, then the probability p_j of surviving t_j days is p_{j-1} times the proportion of patients not dying on day t_j. This can be formulated as

$$p_j = p_{j-1} \times \frac{r_j - f_j}{r_j} \tag{1.2}$$

where r_j is the number of patients still at risk (meaning that they have not been censored) just before day t_j, and f_j is the number of deaths on day t_j. The Kaplan Meier estimator of survival as a function of time t is then given by (Gooley et al., 1999)

$$\text{KM}(t) = \prod_j \left(\frac{r_j - f_j}{r_j} \right). \tag{1.3}$$

Two patient groups with different survival curves, such as shown in Figure 3.3A (page 52), can be compared with the nonparametric logrank test, where the null hypothesis assumes no difference in survival between the two groups.

1.6.5 Cumulative incidence

In survival analysis, there is only two possible outcomes: The patient will, unless censored, either die before or live up until a given time. If instead considering the probability that a patient dies from TRM after an allo-HCT, then a competing risk is present, since the patient could also die from other causes. The Kaplan Meier method cannot be used for such problems, instead a cumulative incidence (CI) analysis is applied.

Considering two competing events TRM and RRM with Kaplan Meier estimators $\text{KM}_1(t)$ and $\text{KM}_2(t)$, respectively, the overall Kaplan Meier survivor function $\text{KM}_{12}(t)$ is the product of the two individual Kaplan Meier estimators (Gooley et al., 1999),

$$\text{KM}_{12}(t) = \text{KM}_1(t)\text{KM}_2(t). \tag{1.4}$$

It has been shown that the CI estimator can be written as (Gooley et al., 1999; Kalbfleisch and Prentice)

$$\text{CI}(t) = \sum_j \frac{f_j}{n_{j-1}} \text{KM}_{12}(t_j) \tag{1.5}$$

where f_j is the number of deaths at time t_j from the event related to KM_1, i.e. TRM, and n_{j-1} is the number of patients still at risk beyond time t_{j-1}. The influence of RRM is included in the CI estimator through the overall Kaplan Meier survivor function $KM_{12}(t)$.

Two patient groups with different CI curves such as shown in Figure 3.3B (page 52) can be compared with Gray's K test (Gooley et al., 1999; Gray, 1988), where the null hypothesis assumes no difference in TRM between the two groups.

1.6.6 Cox regression

Cox regression analysis, also known as proportional hazards regression analysis, is an advanced statistical method, especially useful for multivariate survival analysis where the impact of several covariates on the outcome of interest, e.g. survival, is considered (Altman). In this thesis, Cox regression was used for the analysis presented in Table 3.7 (page 51). The analysis work was assisted by a professional statistician, as it was complicated by the time-dependent nature of two of the covariates (aGVHD and cGVHD). It is beyond the scope of this thesis to explain the mathematics behind Cox regression, instead, I will briefly explain, how the analysis results in Table 3.7 should be interpreted.

The results of three time-dependent multivariate Cox regression analyses are presented in Table 3.7: OS, PFS, and TRM. As an example, the interpretation of the OS analysis is given below. The influences of the three covariates 'Number of predicted mHags', 'Acute GVHD grade III-IV', and 'Extensive chronic GVHD' on the outcome variable OS are considered. Each patient is input into the analysis with information on the time of death or censoring, number of predicted mHags (more or less than the median of 3), and time of aGVHD or cGVHD or censoring. The hazard ratio (HR) denotes the relative risk associated with each of the covariates, for instance, patients having more than 3 predicted mHags are 2.2 times more likely to die within the first 5 years after allo-HCT, compared to patients with maximum 3 predicted mHags. The HR of 2.2 lies within confidence interval (1.2-4.0) and the association between the number of predicted mHags and OS is significant since $P < 0.05$ (P=0.014). Similarly, patients developing aGVHD have a 3.2 times greater risk of death within 5 years, compared to patients with no aGVHD, while cGVHD is not significantly associated with OS in our patient cohort.

1.6.7 Matthews correlation coefficient

The Matthews correlation coefficient (MCC) used in Chapter 5 measures the quality of binary predictions, such as whether or not the correct HLA restriction element is predicted for a given epitope. In this type of predictions, the following 4 outcomes are possible, here using prediction of HLA restriction as an example:

- True positive (TP): The predicted HLA restriction element for an epitope is the same as the validated one.
- False positive (FP): The predicted HLA restriction element is not validated.
- True negative (TN): The HLA restriction element is neither predicted nor validated.
- False negative (FN): The validated HLA restriction element is not predicted.

The specificity is defined as

$$\text{specificity} = \frac{TN}{TN + FP} \tag{1.6}$$

and a specificity of 1 (100%) means that no false positives are predicted. The sensitivity is defined as

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{1.7}$$

and a sensitivity of 1 (100%) means that all the true positives are correctly predicted.

The MCC is defined as (Baldi et al., 2000)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{1.8}$$

Assuming that no false predictions are made, $FP = 0$ and $FN = 0$, the MCC becomes

$$MCC = \frac{TP \times TN}{\sqrt{(TP)(TP)(TN)(TN)}} = 1 \tag{1.9}$$

The denominator is used for normalization such that the MCC can assume values between -1 and 1, where 0 corresponds to random performance, 1 corresponds to perfect predictions, and -1 corresponds to inverse predictions. Together with the sensitivity and specificity, the MCC is thus a useful measure of prediction performance.

1.7 Reading guidelines

In the following chapters, I present my contribution to the field of bioinformatics in relation to transplantation immunology. Chapter 2 describes the bioinformatics predictions of mHag candidates from the Y chromosome and preliminary experimental validation results. Chapter 3 presents the first paper of this thesis, which describes the correlation between the number of predicted mHags, encompassing nsSNPs within known mHag source proteins, and transplantation outcome in a Danish patient cohort. The study described in the paper emerged from the project described in Chapter 4. The aim of that study is to apply reverse immunology to identify novel mHags around nsSNPs in proteins, where mHags have previously been identified, and selected proteins expressed in hematopoietic tissues. In Chapter 5, I present the second paper of this thesis, describing a new online prediction tool *HLArestrictor*, based on *netMHCpan*, for the patient-specific prediction of epitopes within peptides or proteins. In the paper, the method is benchmarked using a large dataset of HIV interferonγ ELIspot responses and a smaller dataset of tetramer validated HIV epitopes and HLA restriction elements. Concluding remarks on the work presented in this thesis are given in Chapter 6.

Chapter 2

Prediction of mHags from the Y chromosome

This chapter concerns an ongoing project in collaboration with Allogeneic Hematopoietic Cell Transplantation Laboratory, Department of Hematology, Rigshospitalet and Laboratory of Experimental Immunology, University of Copenhagen. The purpose of the project is to discover novel mHags encoded by the Y chromosome using computational predictions and subsequent experimental validations. This chapter describes the prediction of candidate mHags using *NetMHCpan* and the criteria used for the final selection of the peptides most suitable for experimental validation. Preliminary experimental validations carried out by PhD student Annika H. Rasmussen at the Laboratory of Experimental Immunology are also described.

2.1 Introduction

The outcome of allo-HCTs has been shown to be influenced by donor/patient gender combinations. Increased alloreactivity resulting in a higher incidence of GVHD, a lower OS, and reduced relapse risk has been reported in female to male (F->M) transplantations (Gahrton et al., 2005; Stern et al., 2008). A selective GVT effect has also been observed by Randolph et al. (2004) in a large study of more than 3,000 transplanted patients. The effects of (F->M) transplantations are illustrated in Figure 2.1, and although the differences are relatively small, they are nevertheless significant in large cohorts. As a result, the use of female unrelated donors in male patients has generally been reduced during the last 10 years, while this is not the case for related donors, which are still preferred to unrelated, regardless of gender (Stern et al., 2008).

The increased alloreactivity in sex-mismatched HCTs is believed to be caused by mHags encoded by the Y chromosome. In accordance with this, as early as 1977, CTLs specific to an unidentified Y-antigen were observed by Goulmy et al. (1977). As much as a third of the approximately 30 mHags, listed in dbMinor (Spierings et al., 2006), are encoded by genes from the Y chromosome (Akatsuka et al., 2007). An example is the HLA-B*2705 restricted mHag SRDSRGKPGY from the *DDX3Y* gene recently identified by Rosinski et al. (2008). CTLs recognizing this mHag were shown to react against leukemic cells expressing the mHag. The *DDX3Y* gene contains two additional known mHags (Akatsuka et al., 2007). However, since it is a broadly expressed gene, the GVT effect could be accompanied by a GVHD effect. A couple of the H-Y encoded mHags are of special interest due to their tissue specificity, which makes them candidate GVT mHags (Spaapen and Mutis, 2008). These are the LPHNHTDL

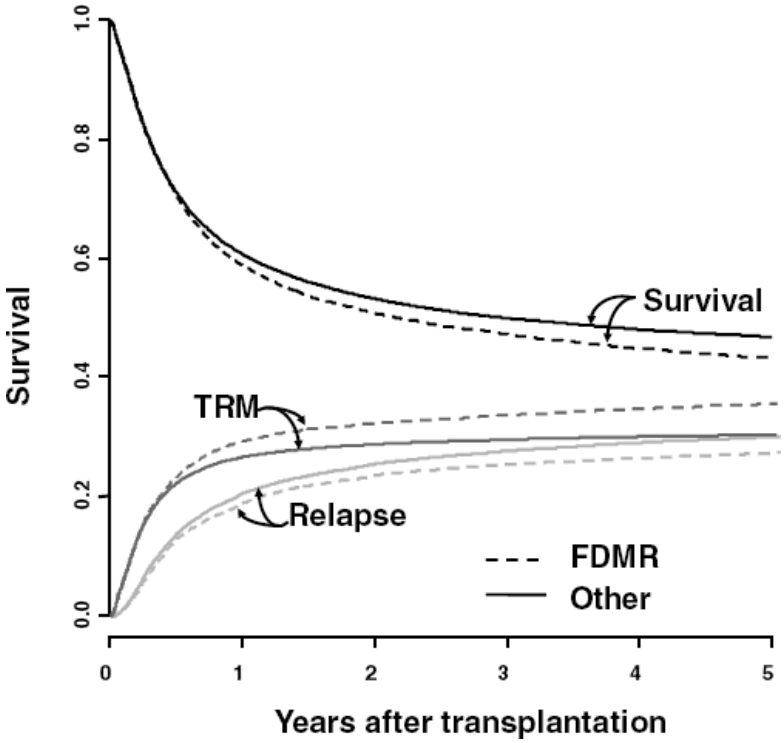


Figure 2.1: **Effects of (F->M) transplantations on outcome.** It is seen that female donor/male recipient (FDMR) patients have a slightly, but significantly, lower overall survival, a higher treatment related mortality and lower relapse incidence. These differences are all indications of the increased alloreactivity in (F->M transplantations). From (Stern et al., 2008).

peptide from the *UTY* gene (Warren et al., 2000) and the TIRYPDPVI peptide from the *RPS4Y* gene (Ivanov et al., 2005). Although the two genes are broadly expressed, CTLs specific to the mHags surprisingly recognize only activated hematopoietic cells and malignant cells expressing the mHags. A reason for this could be an elevated expression level of the proteins in rapidly proliferating cells, which was observed for the RPS4Y gene.

mHags are also believed to be involved in the higher risk of graft rejection found in some studies of male to female (M->F) transplantations (Gahrton, 2007). Graft rejection is a relatively rare event, which occurs when the patient's CTLs recognize and induce apoptosis in the hematopoietic cells from the donor. H-Y specific mHags involved in graft rejection have been identified from the *SMCY*, *DFFRY* and *UTY* genes (Vogt et al., 2000a,b).

2.1.1 The Y chromosome

The genes on the male-specific region of the Y chromosome are described in detail in Skaletsky et al. (2003). The region consists of 3 different regions as shown in Figure 2.2 with a total of 27 genes or gene families:

- The X transposed region which has arised from a major transposition from the X chromosome 3-4 million years ago. Two genes with very close X homologues (99% identity) are found in this region.
- The X degenerate region with 16 genes and a number of pseudogenes which all have homologues on the X chromosome with a sequence similarity of 60-96%. This region is

believed to originate from the ancient autosome that diverged into the modern X and Y chromosomes.

- The ampliconic region, which is not homologous to the X chromosome. In this region each gene is found in multiple, almost identical copies comprising 9 gene families. These genes are all expressed in testis only, which is an immunoprivileged site (Pelletier and Byers, 1992). In general, the blood-testis barrier prevents leukocytes from entering the testis where they would perceive the developing sperm cells as non-self due to genetic recombination. Likewise, after an allo-HCT, the donor CTLs also do not enter the testis. Thus disparate peptides, originating from genes uniquely expressed there, are unlikely to play a role in alloreactivity.

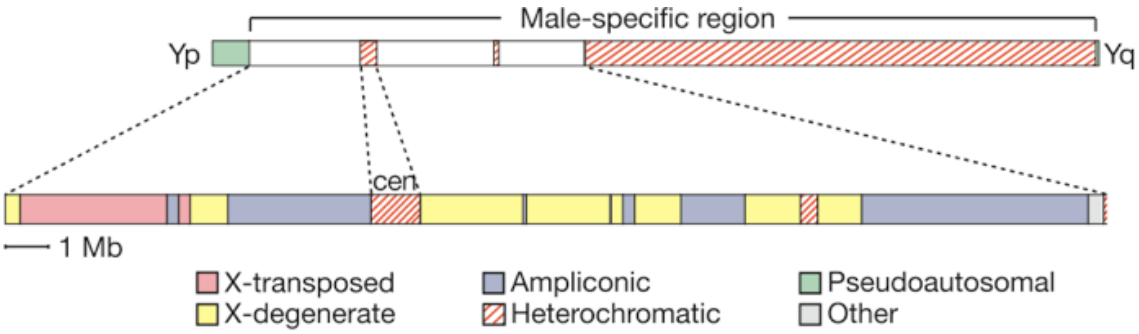


Figure 2.2: Schematic overview of the Y chromosome. From Skaletsky et al. (2003).

2.1.2 The aim of this study

The ongoing study presented here aims at identifying novel mHags encoded by genes on the Y chromosome by means of bioinformatics predictions. Traditionally, the starting point of mHag identification is an isolated CTL clone, known to be involved in alloreactivity. The mHag and restricting HLA allele recognized by the CTL clone is then identified by one of the methods described in Section 1.5.1. Often prediction methods are used as one of the final steps in the procedure, when the mHag has been localized to a small genomic region.

In our approach, the idea is instead to start with the predictions. As shown below, the number of raw mHag predictions from relevant genes on the Y chromosome is overwhelming. Predictions are done for multiple HLA alleles, here, the alleles represented by our patient cohort, and for all peptides of 8-11 amino acids within several selected genes, resulting in thousands of raw peptide/HLA predictions. Therefore, the bioinformatical challenge lies in narrowing down the number of mHag candidates to validate experimentally by intracellular cytokine staining (ICS) and by tetramers. Recently, some preliminary ICS results have been obtained and will be presented at the end of this chapter.

2.2 Materials, methods, and prediction results

2.2.1 Patient set

The patient set used for this study consists of 32 male patients and their HLA-identical female donors of which 26 were siblings, 1 was a mother, and 5 were MUDs. The patients were all treated at the Allo-HCT Laboratory, Department of Hematology, Rigshospitalet between April 2000 and November 2007 with an allo-HCT with peripheral blood graft from their donor after NMC.

For related donors, donor selection was based on serological typing for HLA-A, -B and -C and on molecular typing for HLA class II. For MUDs, donor selection was based on molecular typing for HLA-A, B, C, DRB1, and DQB1. The patient set represents 31 different HLA alleles (12 HLA-A and 19 HLA-B) as shown in Table 2.1. All patients were treated for malignant hematological diseases such as acute myelogenous leukemia/myelodysplastic syndrome (n=13), non-Hodgkin's lymphoma (n=12), chronic lymphocytic leukemia (n=3), multiple myeloma (n=2), and Hodgkin lymphoma (n=2). Donor treatment, conditioning regimen, and supportive care have been described in Kornblit et al. (2008).

2.2.2 Proteins selected for prediction

Of the 27 genes or gene families on the Y chromosome, the 9 ampliconic gene families only expressed in testis were excluded. Thus, 18 H-Y genes (*TGIF2LY, PCDH11Y, SRY, RPS4Y1, ZFY, AMELY, TBL1Y, PRKY, USP9Y, DBY, UTY, TMSB4Y, NLGN4Y, CYorf15A, CYorf15B, SMCY, EIF1AY*, and *RPS4Y2*) remained to be screened for potential mHags. The Ensembl

HLA-A alelle	No. of patients with the allele	HLA-B allele	No. of patients with the allele
A*0201	16	B*4402	8
A*0301	7	B*0801	8
A*0101	6	B*4001	7
A*1101	6	B*0702	6
A*2402	6	B*5101	5
A*6801	6	B*3501	4
A*3201	2	B*1501	4
A*3101	2	B*1302	3
A*2902	2	B*1801	2
A*3001	2	B*2705	2
A*2501	1	B*4403	2
A*2301	1	B*5301	2
		B*3502	1
		B*3701	1
		B*3901	1
		B*1518	1
		B*4002	1
		B*5501	1
		B*5801	1

Table 2.1: Distribution of HLA alleles in the patient set.

database (www.ensembl.org) was used to obtain the 43 different protein products encoded by these 18 genes.

2.2.3 Prediction of mHags

Potential mHags in the 43 H-Y encoded non-ampliconic proteins were predicted with *NetMHCpan* (Nielsen et al., 2007) using all possible combinations of the 31 HLA alleles and all peptides of lengths 8-11 that could be generated from the proteins. A peptide was considered to be a potential epitope if the affinity score was below 500 nM. Note, that a lower affinity score corresponds to stronger binding. The number of raw epitope predictions was 38,573 peptide/HLA combinations comprising 7,390 distinct peptides. Naturally, it is not possible to experimentally validate this huge number of peptides, thus a number of filtering steps were applied as described in the following.

2.2.4 Homologue filtering

A homologue filtering step was included to filter out the predicted mHags that are also encoded by X or autosomal genes in the human genome, mostly the X homologues of the 18 selected H-Y genes. The sequences of all human proteins, except those from the Y chromosome, were obtained from the Ensembl database and were scanned for each of the 7,390 predicted epitopes. If a full match was found, the peptide was excluded from the selection, decreasing the number of distinct peptides by more than 50%, resulting in a set of 14,792 peptide/HLA pairs with 3,182 distinct peptides.

2.2.5 Submer filtering

Often, a peptide predicted to be an mHag will contain one or more shorter peptides, or submers, also predicted as mHags for the same HLA allele. To further reduce the number of peptides to validate experimentally, these submers were filtered out, see Figure 2.3. This ensures that the selected set of peptides is more diverse. After this filtering step, 3,969 peptide/HLA pairs remained with 2,207 distinct peptides.

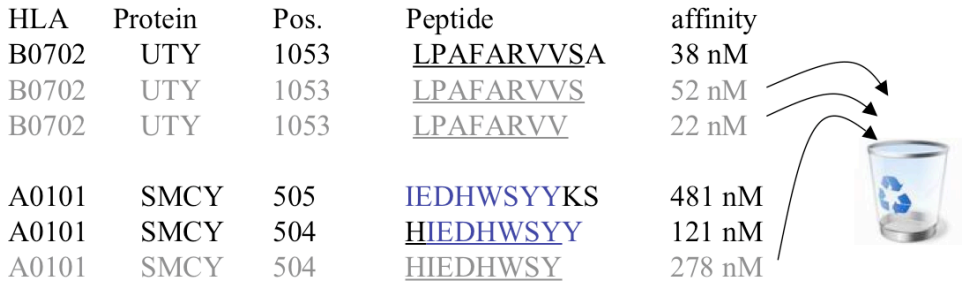


Figure 2.3: **Submer filtering of predicted binders. Top:** If one or more peptides is contained within a larger peptide, only the longest one is kept. **Bottom:** Peptides are only discarded if they are a true submer of a longer peptide. In this example, two peptides are both kept even though they share an 8 amino acid long sequence marked in blue. Binding affinities are not taken into consideration, as long as they are below 500 nM.

2.2.6 Final selection of peptides

Ideally the 2,207 potential mHags should be tested experimentally, but due to financial and logistics reasons, a further reduction of the number of peptides was necessary. For each HLA allele, only the predicted top 30 strongest binders were therefore included in the final set. Furthermore, this was done only for the 15 most common HLA alleles in the patient set (defined by presence in at least 3 patients). To further reduce the peptide set, while keeping maximum diversity between the peptides, a final submer filtering was made. The final submer filtering did not take into account which HLA alleles, a predicted binder was restricted by, thus reducing the number of peptides further. The result of this final selection was a set of 324 peptides to be validated experimentally in patient samples. The peptides are listed in Appendix A.

2.2.7 Backtracing submers

The different HLA alleles often have a significant overlap in the peptides which they can bind. Therefore it was necessary to determine the predicted binding affinity of each of the 324 peptides in the test set to each of the 31 HLA alleles in the patient set. Furthermore, a backtracing was made in order to investigate if any submers of each of the 324 peptides were predicted for a given HLA allele. If so, the longer peptide was included in the subset of peptides to be tested for the given HLA allele, even if the longer peptide itself was not a predicted mHag. The reason for this was that the submers can be generated from the longer versions by peptide cleavage in the ICS experiments.

2.2.8 T cell cytokine responses by ICS

IFNγ, tumor necrosis factor-alpha and interleukin 2 production by T cells, was monitored by ICS as described in (Christensen et al., 2002). T cells were stained with antigen specific to CD8 and CD4, in order to be able to distinguish between class I and II responses. Initially, 12x12 peptide matrixes were designed for each patient, comprising those of the 324 candidate peptides which were predicted to bind, or contained submer(s) predicted to bind, to any of the patients HLA molecules. Peptide mixes were incubated with peripheral blood mononuclear cells (PBMCs) obtained from the patients 6 months - 8 years posttransplantation. Submers were automatically generated *in vitro* by proteases in the cell medium. PBMCs consists of all blood cells with a round nucleus, where B cells serve as antigen presenting cells which presents the peptides and submers to T cells. Whenever T cell responses were observed in two or more peptide mixes, the corresponding peptides were tested individually.

2.2.9 Validation of peptide/HLA binding

Peptides eliciting a T cell cytokine response in the ICS measurements and with a predicted binding to any of the given patient's HLA molecules were screened using a luminescent oxygen channeling immunoassay (LOCI) in order to measure the binding affinity of the predicted peptide/HLA complexes. The LOCI method is described in (Harndahl et al., 2009). Peptides recognized by T cells, but where only submers were predicted binders, were not measured in binding assays.

2.2.10 Tetramer validations

Tetramer validations are planned for peptides with a T cell response confirmed by ICS, or for submers of these with predicted binding to any of the given patient's HLA molecules. The experiments will be carried out as described in (Leisner et al., 2008).

2.3 Preliminary validation results

The validations of the predicted peptides are carried out by our collaborators at Laboratory of Experimental Immunology. Until now, 8 patients have been tested by ICS assays, and 35 individual CD8+ T cell responses to 30 different peptides have been observed (see Table 2.2). Furthermore, 18 individual CD4+ T cell responses to 14 different peptides were also observed (see Appendix B). Only the longest version of a predicted peptide has been tested initially, and, if an ICS response is observed, the submers will be tested as well. Additionally, binding assays were made to confirm the binding of the peptides to the predicted HLA alleles, if these were available in the lab (also shown in Table 2.2). The next step is to characterize the T cell responses using tetramers, for the peptides which elicited a response in the ICS experiments. The tetramers should consist of the peptides or submers and the HLA molecules with the most promising prediction results.

Protein	Peptide	Predicted restriction	Measured binding
*Patient 289: A*0301, A*2402, B*0702, B*3508, C*0401, C*0702*			
UTY	**YFYYNAFHWAI**	A*2402(70nM)	A*2402(1632nM)
	YYNAFHWAI	A*2402(15nM), C0702(494nM))	
	YFYYNAFHW	A*2402(11nM)	
	FYYNAFHWA	A*2402(431nM)	
	FYYNAFHW	A*2402(7nM)	
*Patient 257: A*0101, B*0801, B*4001, C*0304, C*0701*			
USP9Y	**IVDCLTEMYY**	A*0101(36nM)	A*0101(21nM)
	IVDCLTEMY	A*0101(28nM)	
*Patient 297: A*2402, A*3101, B*3501, B*5101, C*0102, C*0401*			
USP9Y	**FPHTELANL**	B*3501(470nM), B*5101(1561nM)	B*3501(10nM), B*5101(753nM),
USP9Y	**ELFARSSDPR**	A*3101(344nM)	A*3101(11569nM)
	LFARSSDPR	A*3101(171nM)	
	FARSSDPR	A*3101(474nM)	
USP9Y	**SYMMDDLELI**	A*2402(71nM)	A*2402(10nM)
	YMMDDLELI	A*2402(404nM)	
	YMMDDLEL	B*3501(156nM)	
PRKY	**LVTMGTGTFGR**	A*3101(216nM)	A*3101(None)
	VTMGTGTFGR	A*3101(31nM)	
	TMGTGTFGR	A*3101(34nM)	
	LVTMGTGTF	B*3501(144nM)	
	MGTGTFGR	A*3101(252nM)	
AMELY	**RPPYSSYGY**	B*3501(84nM)	B*3501(95nM)
	PPYSSYGY	B*3501(217nM)	
*Patient 287: A*0101, A*2501, B*1801, B*3701, C*0602, C*1203*			
USP9Y	**VALFSSCPVAY**	None	
	FSSCPVAY	A*0101(133nM)	
USP9Y	**FQILHDRFF**	B*1801(2252nM), B*3701(3258nM)	B*1801(4964nM), B*3701(384nM)
SRY	**TEAEKWPFF**	B*1801(35nM), B*3701(476nM)	B*1801(17nM), B*3701(340nM)

Protein	Peptide	Predicted restriction	Measured binding
*Patient 627: A*0201, A*6802, B*0702, B*5301, C*0401, C*0702*			
USP9Y	**YSLEYFQFVKK**	None	
	YSLEYFQFV	A*6802(32nM)	
	YSLEYFQF	B*5301(110nM), C*0702(84nM)	
	SLEYFQFV	A*0201(68nM)	
SMCY	**SLLERGQQLGV**	A*0201(23nM)	A*0201(28nM)
	SLLERGQQL	A*0201(53nM)	
ZFY	**TFVPIAWAAAY**	None	
	TFVPIAWAAA	A*6802(467nM)	
	FVPIAWAAA	A*6802(65nM)	
	VPIAWAAAY	B*5301(13nM)	
	FVPIAWAA	A*0201(464nM), A*6802(84nM)	
SRY	**LPADPASVL**	B*0702(9nM), B*5301(175nM)	B*0702(198nM), B*5301(1700nM)
*Patient AET: A*2402, A*6802, B*4002, B*4402, C*0304, C*0501*			
UTY	**MIKYCLLKILK**	None	
	KYCLLKIL	A*2402(1278nM)	
USP9Y	**RMILPMSRAFR**	None	
	MILPMSRAF	C*0304(531nM)	
SRY	**RYSHWTKL**	A*2402(166nM)	A*2402(152nM)
UTY	**YFYYNAFHWAI**	A*2402(70nM)	A*2402(1632nM)
	FYYNAFHWAI	A*2402(13nM)	
	YFYYNAFHW	A*2402(11nM)	
	YYNAFHWAI	A*2402(15nM)	
	FYYNAFHW	A*2402(7nM)	
USP9Y	**GSSDFQVHFLK**	None	
	SDFQVHFL	B*4002(246nM)	
USP9Y	**MVRVLTVIKEY**	None	
	MVRVLTVI	C*0304(717nM)	

Protein	Peptide	Predicted restriction	Measured binding
*Patient 283: A*0201, B*5101, B*5801, C*0702, C*0718*			
UTY	**LPSCPTNFCIF**	B*5101(1104nM)	B*5101(None)
	LPSCPTNFCI	B*5101(189nM)	
CYorf15A	**ILNRETLLDFV**	A*0201(13nM)	A*0201(139nM)
	ILNRETLL	A*0201(242nM)	
UTY	**YFYYNAFHWAI**	A*0201(328nM), C*0702(237nM)	A*0201(4595nM)
	YFYYNAFHW	B*5801(239nM)	
PCDH11Y	**KSLTTTMQFK**	None	
	KSLTTTMQF	B*5801(11nM)	
SMCY	**SLMASSPTSI**	A*0201(17nM)	A*0201(27nM)
	MASSPTSI	B*5801(294nM)	
	LMASSPTSI	A*0201(36nM)	
CYorf15B	**KVADVDLAVPV**	A*0201(20nM)	A*0201(30nM)
	KVADVDLAV	A*0201(46nM)	
USP9Y	**VALFSSCPVAY**	B*5801(570nM)	B*5801(None)
	ALFSSCPVA	A*0201(75nM)	
	ALFSSCPV	A*0201(5nM)	
DDX3Y	**RQSSGSSSSGF**	None	
	QSSGSSSSGF	B*5801(614nM)	
RPS4Y1	**YPDPVIKV**	B*5101(1247nM)	B*5101(None)
DDX3Y	**FLLPILSQIYT**	A*0201(12nM)	A*0201(6nM)
	FLLPILSQI	A*0201(4nM)	
	LLPILSQI	A*0201(293nM)	
*Patient 611: A*1101, A*2902, B*4002, B*4403, C*0202, C*1601*			
NLGN4Y	**SSKMFNYFK**	A*1101(5nM)	A*1101(59nM)
	SKMFNYFK	A*1101(150nM)	
AMELY	**YQSMIRPPY**	A*2902(46nM)	A*2902(43nM)
	QSMIRPPY	A*2902(292nM)	
USP9Y	**LEYFQFVKKLL**	B*4002(1805nM)	B*4002(65nM)
	LEYFQFVKKL	B*4002(44nM)	

Table 2.2: T cell cytokine responses observed by ICS in 8 tested (F->M) patients. Peptides marked in bold elicited at T cell response. Predicted non-binders were tested if they contained any submers predicted to bind to any of the patient's HLA alleles. The submers (peptides in normal font) are not tested yet, but could give rise to some of the responses, due to peptide cleavage. The binding affinity of some of the peptides to available HLA alleles were measured in binding assays. The results of these are added in the last column, where 'None' means that no binding could be measured (affinity > 20,000 nM).

2.4 Discussion and outlook

Although reverse immunology has been applied in disease epitope discovery for vaccine development for a decade (Mora et al., 2003), the concept has only recently been applied for mHag identification. This was done by Ofran et al. (2010), who predicted 41 HY mHag candidates restricted to HLA-A*0201 of which 13 9mers or 10mers elicited a response in ELIspot assays and were confirmed to bind to the HLA-A*0201 molecule. Although the mHags were not confirmed by tetramers, the study represents a proof of principle for our approach.

Starting with more than 38,000 predicted raw peptide/HLA combinations comprising 7,390 distinct peptides, the largest bioinformatical challenge in this project has been to narrow down the number of mHag candidates to validate experimentally. This was done through a number of filtering steps. More than half of the peptides were found identically on the X chromosome and are thus unlikely to be H-Y mHags. A relatively large number of peptides were removed, since they were submers of longer peptides which were also predicted binders. However, due to financial and logistics reasons, a restrictive filtering step was implemented with the criteria that only the top 30 strongest binders for each of the more common HLA alleles should be included in the set of peptides to test experimentally.

The resulting 324 peptides selected for the ICS experiments were thus intended to cover as many of the 32 patients and 31 HLA-A and -B alleles as possible. HLA-C alleles were not included in the predictions when selecting the peptides since *NetMHCpan* is less accurate for HLA-C (Hoof et al., 2009). However, predictions for HLA-C were made subsequently and are included in Table 2.2.

A number of the peptides, eliciting a T cell response from the 8 tested patients, were confirmed in binding assays to bind to one of the patient's HLA molecules. These peptide/HLA complexes should therefore be synthesized as tetramers and tested for T cell recognition. In other cases, the peptides did not bind any of the tested HLA molecules. In those cases, it is likely that the response was caused by one of the peptide's submers which should then instead be tested with tetramers, guided by prediction results for the submers. The criteria used when selecting the test set of 324 peptides favors the longer versions of the peptides, while most known class I epitopes are 9 amino acids long. As mentioned, the reason for selecting the longer peptides, instead of the submers, was the fact that the shorter versions are generated *in vitro* by proteases in the ICS process and hence should be covered when testing by ICS. Based on the ICS results, a new test set should then be selected for tetramer validations.

In this study, the homolog filtering step was quite simple, as only predicted binders with exact matches on the X chromosome (or, in few cases, on some of the autosomes), were removed from the peptide selection. This step could have been more sophisticated, if the number of amino acid changes, or differences in binding strength between the X and Y version of the peptide had been considered. A predicted binder, whose X chromosome variant is not predicted to bind, is likely a better mHag candidate, since the X chromosome variant of the peptide might not be presented at all during the T cell training in the thymus of the female donor. Likewise, the number and positions of amino acid changes could affect the recognition of the peptide by the donor T cells, even if the binding of the peptide was unaffected.

Different methods for submer filtering were investigated during the peptide selection. One approach was to keep the strongest binder in a family of predicted binders of unequal length. This would probably have been the best solution if only one peptide selection was to be made. Since we here planned to comprise another selection of peptides and submers after observing T cell responses by ICS, we instead decided to keep the longest peptide within a peptide family. We also investigated the effects of filtering peptides by affinity if there was an 8 amino acids

sequence overlap, but one peptide was not a submer of the other. This approach was also not selected, as it sometimes resulted in an 'domino effect' where a peptide could 'win' over its neighbor and then subsequently get filtered out by another peptide.

The study presented here is, as mentioned, still ongoing, and has the aim of discovering previously unknown mHags encoded by genes on the Y chromosome, starting with bioinformatics predictions. If successful, this systematic and direct approach could significantly increase the number of discovered mHags. 35 CD8+ T cell responses in 8 patients have already been observed by ICS, thus it seems likely that mHags could indeed be identified in subsequent tetramer validations.

Chapter 3

Degree of predicted mHag mismatch correlates with poorer clinical outcome of allo-HCT

This chapter presents the first study to demonstrate the correlation between the number of predicted mHags and outcome after allogeneic NMC HCT. We use *NetMHCpan* to predict nsSNP-related mHags specific to each patient-donor pair in proteins known to contain mHags and demonstrate that the number of such mHags provides a strong correlate of the transplantation outcome.

Our data suggest that some of the proteins known to contain mHags are likely to contain several additional mHags that have yet to be identified, and that the presence of multiple mHags confers a higher risk of mortality after NMC HCT. Furthermore, our data suggest a possible role for *in silico* based mHag prediction, in both donor selection and in selecting candidate mHags for further evaluation in *in vitro* and *in vivo* experiments.

The paper presented in this chapter emerged as part of a larger project described in Chapter 4, the aim of which is to predict and experimentally identify novel mHags around selected nsSNPs. The focus of the work presented here is to correlate the prediction of mHags with clinical transplantation outcome.

My efforts in the work presented in the following paper involved patient-specific predictions and statistical analyses of transplantation outcomes. In addition, I was main responsible for writing the manuscript which was published in *Biol Blood Marrow Transplant Oct. 2010, 16(10):1370-81.*

Degree of predicted minor histocompatibility antigen mismatch correlates with poorer clinical outcomes of nonmyeloablative allogeneic hematopoietic cell transplantation.

Malene Erup Larsen[1], Brian Kornblit[2], Mette Voldby Larsen[1], Tania Nicole Masmas[2], Morten Nielsen[1], Martin Thiim[1], Peter Garred[3], Anette Stryhn[4], Ole Lund[1], Søren Buus[4], Lars Vindelov[2].

[1]Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark
[2]Allogeneic Hematopoietic Cell Transplantation Laboratory, Department of Hematology, Rigshospitalet, Denmark
[3]Laboratory of Molecular Medicine, Department of Clinical Immunology, Rigshospitalet, Denmark
[4]Laboratory of Experimental Immunology, University of Copenhagen, Denmark

Abstract

In fully HLA-matched allogeneic hematopoietic cell transplantations (HCT), the main mechanism of the beneficial graft-versus-tumor effect and of the detrimental graft-versus-host disease is believed to be caused by donor cytotoxic T cells directed against disparate recipient minor histocompatibility antigens (mHags). The most common origin of disparate mHags is non-synonymous single nucleotide polymorphism (nsSNP) differences between donors and patients. At this time, only some 30 mHags have been identified and registered, but considering the numerous different HLA-types in the human population as well as all the possible nsSNP differences between any two individuals, it is likely that many mHags have yet to be discovered. The objective of the current study was to predict novel HLA-A and HLA-B restricted mHags in a cohort of patients treated with non-myeloablative conditioning allogeneic HCT (matched related donor, n=70; matched unrelated donor, n=56) for hematologic malignancies. Initially, the cohort was genotyped for 53 nsSNPs in 11 known mHag source proteins. Twenty-three nsSNPs within six mHag source proteins showed variation in the graft-versus-host direction. No correlation between the number of disparate nsSNPs and clinical outcome could be observed. Next, mHags in the graft-versus-host direction were predicted for each patient-donor pair. Using the *NetMHCpan* predictor, we identified peptides encompassing a nsSNP variant uniquely expressed by the patient and with predicted binding to any of the HLA-A or B molecules expressed by the patient and donor. Patients with more than the median of three predicted mHags had a significantly lower five-year overall survival (42% vs 70%, P=0.0060, adjusted hazard ratio (HR) 2.6, P=0.0047) and significantly higher treatment related mortality (39% vs 10%, P=0.0094, adjusted HR 4.6, P=0.0038). No association between number of predicted mHags and any other clinical outcome parameters was observed. Collectively, our data suggest that the clinical outcome of HCT is not affected by disparate nsSNPs *per se*, but rather by the HLA-restricted presentation and recognition of peptides encompassing these. Our data also suggest that 6 of the 11 proteins included in the current study could contain more mHags yet to be identified, and that the presence of multiple mHags confers a higher risk of mortality after non-myeloablative conditioning HCT. Furthermore, our data suggest a possible role for *in silico* based mHag predictions, in donor selection as well as in selecting candidate mHags for further evaluation in *in vitro* and *in vivo* experiments.

Keywords: Allo-HCT, mHags, nonsynonymous SNPs, GVHD, survival

3.1 Introduction

In recent years, the role of minor histocompatibility antigens (mHags) in HLA-matched allogeneic hematopoietic cell transplantation (HCT) has become increasingly evident. mHags are immunogenic HLA-presented peptides derived from protein products of polymorphic genes that are disparate between patient and donor (Hambach et al., 2007). Although most of these polymorphic proteins result from nonsynonymous single nucleotide polymorphisms (nsSNPs) in autosomal genes, mHags also may be caused by gene deletions, genetic variation in non-coding regions affecting gene transcription or the presence of Y chromosome-encoded proteins in sex-mismatched HCT (Kawase et al., 2007; Mullally and Ritz, 2007; Murata et al., 2003; Spaapen and Mutis, 2008). Depending on their tissue distribution, mHags with broad tissue expression may induce graft-versus-host-disease (GVHD), whereas mHags, which are expressed only in hematopoietic tissue, may induce a graft-versus-tumor (GVT) effect (Hambach et al., 2007).

Several studies have linked the presence of mHag-specific T cells posttransplantation with graft rejection (Goulmy et al., 1977; Voogt et al., 1990), GVHD (Akatsuka et al., 2003b; Goulmy et al., 1996; Perez-Garcia et al., 2005), and the GVT effect (de Rijke et al., 2005; Stumpf et al., 2009; van Bergen et al., 2007). Because GHVD is a major cause of transplantation-related morbidity and treatment-related mortality (TRM) (Ferrara et al., 2009), identification and characterization of mHags specifically expressed in hematopoietic but not other normal tissues could contribute to the development of selective GVT oriented immunotherapy by separating the beneficial GVT effect from GVHD.

Approximately 30 mHags have been identified (Spierings et al., 2006) by various methods, including peptide elution from the major histocompatibility complex (MHC) (Spierings et al., 2003a; van Bergen et al., 2007), expressional cloning (Dolstra et al., 1999; Kawase et al., 2007; Warren et al., 2006), genetic linkage analysis (Akatsuka et al., 2003a; de Rijke et al., 2005), and genome-wide association analysis (Kamei et al., 2009; Kawase et al., 2008; Ogawa et al., 2008). Common to all these methods is that they identify only a few mHags restricted to no more than few HLA types. Considering the diversity of the HLA system with >3500 known alleles (Robinson et al., 2001, 2003, 2009), as well as the >120,000 known allelic nsSNP variants (UniProt, 2009), it seems likely that many mHags have yet to be identified. If GVT-oriented immunotherapy is to be broadly applicable to a large number of patients, then the number of known mHags needs to be expanded in a systematic manner and on a larger scale using computerized methods (Kessler and Melief, 2007). This has been addressed by several different bioinformatics techniques using algorithms to integrate databases containing information about protein processing, MHC-peptide binding, SNP data, and tissue-specific gene expression (de Rijke et al., 2005; Deluca et al., 2009; Halling-Brown et al., 2006; Schuler et al., 2005).

NetMHCpan (Nielsen et al., 2007) is an MHC-peptide binding prediction tool capable of predicting the binding of peptides to any MHC molecule with a known protein sequence. The method is based on an Artificial Neural Network trained on experimental MHC-peptide binding data. In 2 recent comparisons, *NetMHCpan* has proven superior to other available predictors in predicting HLA class I binding (Lin et al., 2008a; Zhang et al., 2009). The purpose of the current project was to investigate the association between the number of predicted mHags and the outcome after allogeneic HCT with nonmyeloablative (NMA) conditioning. Based on patient and donor HLA-A and -B types and genotype for a number of nsSNPs, mHags were predicted in proteins already known to contain mHags. Known mHag source proteins were chosen because the previous discovery of mHags in these proteins indicates that they are

expressed in relevant tissues and have an expression and degradation frequency that allows peptides from the proteins to be presented by HLA molecules.

3.2 Materials and methods

3.2.1 Patients

This analysis includes data from 126 consecutive patients who underwent allo-HCT with a peripheral blood graft from an HLA-identical related or 10/10 allele-matched unrelated donor after NMA conditioning between April 2000 and July 2007 at the allo-HCT unit, Department of Hematology, Rigshospitalet, Copenhagen. For related donors, donor selection was based on serologic typing for HLA-A, -B, and -C, and on molecular typing for HLA class II. For unrelated donors, donor selection was based on molecular typing for HLA-A, -B, -C, -DRB1, and -DQB1. When available, HLA-identical siblings were preferred to matched unrelated donors, and cytomegalovirus serostatus and sex mismatch were taken into account when possible. Molecular class I typing of related patients and donors was performed retrospectively as part of this study.

All patients were treated for a malignant hematologic disease, including acute myelogenous leukemia/myelodysplastic syndrome (n = 58), non-Hodgkin lymphoma (n = 25) (follicular lymphoma, n = 15; diffuse large B cell lymphoma, n = 4; mantle cell lymphoma, n = 3; peripheral T cell lymphoma, n = 3), chronic lymphocytic leukemia (n = 18), multiple myeloma (n = 12), and Hodgkin's disease (n = 13). The diseases were classified as low, standard, or high risk according to Kahl et al. (2007). Detailed patient and donor demographic data are summarized in Table 3.1. Donor treatment, conditioning regimen, and supportive care were as described previously (Kornblit et al., 2008). All patients were conditioned with fludarabine 30 mg/m^2 for 3 days and 2 Gy of total body irradiation (TBI), except for 2 patients who were conditioned with 2 Gy TBI only. Acute and chronic GVHD (aGVHD, cGVHD) was diagnosed according to standard criteria (Sullivan, 2003). Informed consent was obtained from all patients, and the local Ethics Committee approved the study design.

3.2.2 Prediction of mHags

Eleven non-Y chromosomal proteins (see Table 3.2) known to contain mHags were selected from the dbMinor database (Spierings et al., 2006). The amino acid sequences of the 11 proteins were obtained from RefSeq (Pruitt et al., 2007), and the nsSNPs in these were identified using dbSNP (Smigielski et al., 2000). *NetMHCpan* was used to predict the binding to the HLA-A or -B molecules presented by the patients for all peptides with a length of 8-11 amino acids encompassing the nsSNPs. For each HLA allele, binding peptides were defined as those peptides with a predicted binding strength within the top 1% among random natural peptides. A total of 53 nsSNPs were selected for genotyping (see Table 3.3), all with a minor allele frequency of $\geq$1% in the HapMap CEU population (Consortium, 2003) and located within peptides predicted to bind to at least one of the HLA-A or -B molecules represented in the patient cohort. For a peptide to be considered a potential mHag in the graft-versus-host (GVH) direction for a given patient, the peptide should be predicted to bind at least one of the patient's HLA-A or -B molecules according to the foregoing definition, and the patient should carry the allele coding the binding peptide variant, whereas the donor should be homozygous for the alternative allele. This definition also allows for the donor's variant of the peptide to be

a predicted binder, because the donor's T cells might recognize the difference between the 2 variant peptides.

Variable	Patients (n = 126), n (%)
Patient age, years	
Median=53	
Range 19–69	
Donor age, years	
Median=44	
Range 19–68	
Patient age ≤ 40 years	18 (14%)
Patient age > 40 years	108 (86%)
Donor age ≤ 40 years	50 (40%)
Donor age > 40 years	76 (60%)
Type of donor	
Matched related	70 (56%)
Matched unrelated	56 (44%)
Sex of patient/donor	
Male/female	28 (22%)
Other combinations	98 (78%)
Underlying disease*	
Low risk	25 (20%)
Standard risk	63 (50%)
High risk	38 (30%)
CMV status of patient/donor	
CMV-negative/CMV-negative	25 (20%)
Other combinations	101 (80%)

*Underlying disease was classified as low, standard,or high risk according to Kahl et al. (2007).
CMV indicates cytomegalovirus.

Table 3.1: Patient and donor characteristics

Protein Symbol	Protein Name	Protein Length, aa	Known miHA	Sequence
HMHA1	Histocompatibility (minor) HA-1	1138	HA1, HA-1/B60	VLHDDLLEA, KECVLHDDL
MYO1G	Myosin 1G	1018	HA-2	YIGEVLVSV
AKAP13	A kinase (PRKA) anchor protein 13	2817	HA-3	VTEPGTAQY
KIAA0020	KIAA0020	649	HA-8	RTLDKVLEV
HMHB1	Histocompatibility (minor) HB1	41	HB-1H, HB-1Y	EEKRGSLHVW, EEKRGSLYW
BCL2A1	BCL2-related protein A1	175	ACC-1, ACC-2	DYLQYVLQI, KEFEDDIINW
LRH1	Purinergic receptor P2X5 isoform A	422	LRH-1	TPNQRQNVC
ECGF1	Endothelial cell growth factor	482	LB-ECGF-1H	RPHAIRRPLAL
CTSH	Cathepsin H	335	CTSH/A31, A33	ATLPLLCAR, WATLPLLCAR
TOR3A	Torsin family 3, member A	397	LB-ADIR-1F	SVAPALALFPA
SP110	SP110 nuclear body protein, isoform A	689	SP110(HwA-9)	SLPRGTSTPK

miHAs indicates minor histocompatibility antigen.

Table 3.2: Proteins selected from dbMinor and their reported mHags

Protein	nsSNPs	Predicted miHAs around nsSNP	Patients with nsSNP in GVH Direction	Patients with Predicted miHA(s)	HLA Types Covered	Example Predicted miHA	Known miHAs around nsSNP
SP110	rs9061	6	15	11*	A0301, A1101, A6801, B3801, B4001	KLTSKMNA(K/E)	
	rs28930679	2	28	11*	B4001	(A/V)EEDSEEMPSL	
	rs1135791	12	47	31	A0301, A1101, A3101, A3102, A3103, B0801, B2702, B4001	(M/T)TLGELLK	
	rs3948463	11	12	12	11 HLA-As, B3505, B1302, B5101, B5201	MLWSCTFCR(I/M)	
	rs3948464	17	25	2	A3001, A3101, B2702, B2705	RTKCARKSR(L/S)K	
HMHB1	rs161557	14	25	15	A3001, A0201, A0203, A2402, 8 HLA-Bs	(Y/H)VWKSELVEV	HB-1H, HB-1Y
AKAP13	rs745191	4	24	16	A0101, 11 HLA-Bs	PSDLALL(V/G)	
	rs2061821	8	33	26	A0101, A2902, A3002, A8001, A2501, A2601,11 HLA-Bs	V(M/T)EPGTAQY	HA-3
	rs2061822	8	35	21	A0301, A1101, 10 HLA-Bs	LMNPDATV(W/R)K	
	rs2061824	2	34	6	A3001, B4001	(R/C)EESADAPV	
	rs4075254	7	35	19	A0101, A1101, A0301, 7 HLA-Bs	NTDSSLQS(V/M)	
	rs4075256	6	35	24	B0702, B5501, B5601, B1401, B1402, B3901, B3701, B4001	RPLEDRA(V/A)GL	
	rs4843074	2	33	0	A0203, B3502, B3503	DALNCSQ(P/A)SPL	
	rs4843075	8	36	2	A6802, B4001, B4901, B1302, B3701, B4501, B5001	CEVSG(D/N)VTV	
	rs7162168	8	36	16	A0301, A3101, A3102, A3103, A6601, A6801, 7 HLA-Bs	V(M/T)RAPPSGR	
	rs7177107	4	15	4	A6801, A0301, B4501	KLCDNIVS(K/E)	
	rs34434221	4	5	0	A1101, B0702, B5501, B5601, B3801, B5101	(Q/K)PVDKISV	
	rs35624420	5	2	0	7 HLA-As, 7 HLA-Bs	RAVGLSTS(F/S)	
BCL2A1	rs1138357	16	29	25	10 HLA-As, 9 HLA-Bs	YLQ(Y/C)VLQI	ACC-1
	rs1138358	17	29	24	8 HLA-As, 7 HLA-Bs	VLQ(K/N)VAFSV	
	rs3826007	14	27	16	A3201, A2501, 14 HLA-Bs	KEFEDDII(G/D)II	ACC-2
MYO1G	rs3735485	10	27	12	8 HLA-As, B1518, B0801, B0809	D(M/T)HHRHHL	
	rs7792760	9	26	8	A0301, A3001, A3101, A3102, A3103, A6801, B3901	RLKTL(Q/R)DK	
KIAA0020	rs2173904	7	33	19*	A0301, A1101, A3001, A2301, A2402, 9 HLA-Bs	KSADH(R/P)TLDK	HA-8
	rs2270891	11	6	19*	A0201, A0301, A1101, A3001, A3201, 10 HLA-Bs	LE(V/L)QPEKL	
	rs10968457	3	6	2	A0301, A3001	KQFTGK(S/N)TK	

miHA indicates minor histocompatibility antigen, nsSNP, nonsynonymous single nucleotide polymorphism.
For each nsSNP, the following are listed: number of predicted miHAs, number of patients with the nsSNP difference in the GVH direction, the number of patients with at least one predicted miHA around the nsSNP, the HLA types to which the miHAs around the nsSNP are predicted to bind, an example miHA, and the name of any known miHAs around the nsSNP.
*Number of patients with predicted miHAs containing either of 2 close SNPs.

Table 3.3: Overview of the selected nsSNP with variation in the GVH direction, predicted mHags, and prevalence in patients

3.2.3 Genotyping

Pretransplantation DNA from patients and DNA from donors were genotyped for the 53 nsSNPs using a 12-plex format GenomeLab SNPstream genotyping system (Beckman Coulter, Brea, CA) according to the manufacturer's protocol. The genotype of each of the polymorphisms was validated in 5-10 samples by direct Sanger sequencing (ABI Prism 3100 Genetic Analyzer; Applied Biosystems, Foster City, CA) using PCR primers designed for the SNPstream Genotyping system (autoprimer.com; Beckman Coulter) and purification by ethanol precipitation as described previously (Kornblit et al., 2007). In some cases, failed or missing genotypes could be inferred from linkage disequilibrium (LD) with the successfully genotyped nsSNPs. The criterion for inferring genotypes in this way was complete LD (R^2=1) using the CEU population in the HapMap database (Consortium, 2003). To validate the genotyping assay in the event of departure from the Hardy-Weinberg equilibrium (HWE), a control population of 96 healthy Danish Caucasian blood donors was genotyped by direct sequencing for the relevant nsSNPs.

3.2.4 Statistical analysis

LD, expressed as the squared correlation coefficient, R^2, quantified between all pairs of biallelic loci was estimated using SNPAlyze version 4.0 (Dynacom, Yokohama, Japan). The HWE was assessed separately in the patient and donor populations, and analyzed using gene frequencies obtained by simple gene counting and the χ^2 test. Where applicable, Fisher's exact test was used to compare frequencies.

Cox regression was used to estimate the association between the number of nsSNP differences or predicted mHags and overall survival (OS), progression-free survival (PFS), relapse incidence (RI), relapse-related mortality (RRM), TRM, grade II-IV aGVHD, grade III-IV aGVHD, or extensive cGVHD. Probability of OS and PFS was estimated by the Kaplan-Meier method, and comparisons were made with the logrank test, whereas the cumulative incidences of RI, RRM, TRM, and GVHD were compared using Gray's K test (Gooley et al., 1999; Gray, 1988). In the estimates of RI, RRM, TRM, and GVHD, death before relapse, death with or without relapse, death without GVHD, and retransplantation were handled as competing events when appropriate (Gooley et al., 1999).

OS was measured from the time of transplantation until death from any cause. Patients still alive at the time of analysis were censored at the date of last follow-up. PFS was calculated from the date of transplantation to the date of first relapse or death. Patients who were alive and in remission were censored at date of last follow-up. TRM was defined as death in complete remission (CR) or death where it was not possible to assess disease status before death. RRM was defined as death during relapsed or progressive disease. In the multivariate Cox regression models, all of the covariates listed in Table 3.1, along with the presence of GVHD (time-dependent covariate), were entered one by one into a pairwise model together with the number of nsSNP differences or predicted mHags. The covariates were kept in the final model if they remained significant ($P < .05$) or altered the association with the number of nsSNP differences or predicted mHags by $>10\%$. All P values were 2-tailed, and $P < .05$ was considered significant.

				Patients					Donors				
nsSNP	Major allele A	Minor allele a	Fisher's exact test	Failed Genotypes, %	HWE	AA, %	Aa, %	aa, %	Failed Genotypes, %	HWE	AA, %	Aa, %	aa, %
rs9061	G	A	0.45	0	0.80	81.7	16.7	1.6	0.8	**0.01**	84.8	12.0	3.2
rs28930679	C	T	0.27	2.4	0.86	60.2	35.0	4.8	1.6	0.15	55.6	33.9	10.5
rs1135791	T	C	**0.05**	4.0	0.08	23.1	58.7	18.2	4.0	0.33	30.6	44.6	24.8
rs3948463	G	A	1.00	3.2	0.64	83.6	15.6	0.8	3.2	0.59	82.8	16.4	0.8
rs3948464	C	T	0.38	0.8	0.46	71.2	24.8	4.0	0	0.19	78.6	18.2	3.2
rs161557	C	T	0.75	2.4	0.83	57.7	36.6	5.7	3.2	0.42	59.8	32.8	7.4
rs745191	G	T	0.22	1.6	0.12	53.2	43.6	3.2	3.2	0.32	45.1	47.5	7.4
rs2061821	C	T	**0.04**	4.8	**0.01**	28.3	60.8	10.9	4.8	**0.003**	40.8	55.0	4.2
rs2061822	C	T	0.25	0.8	**0.01**	32.8	59.2	8	1.6	**0.02**	41.9	53.2	4.8
rs2061824	T	C	0.08	1.6	**0.002**	27.4	62.9	9.7	4.0	**0.01**	39.7	55.4	4.9
rs4075254	A	G	0.06	3.2	**0.01**	27.9	61.5	10.6	4.8	**0.01**	40.0	55.0	5.0
rs4075256	C	T	0.16	2.4	**0.005**	27.6	61.8	10.6	2.4	**0.02**	38.2	55.3	6.5
rs4843074	G	C	0.11	7.1	**0.001**	25.6	64.1	10.3	5.6	**0.02**	37.8	55.5	6.7
rs7177107	G	A	0.15	5.6	0.23	67.2	26.9	5.9	5.6	0.34	64.7	33.6	1.7
rs34434221	A	C	0.27	10.3	0.81	95.6	4.4	0	4.0	0.89	98.3	1.7	0
rs35624420	C	T	1.00	0	0.93	98.4	1.6	0	0	0.93	97.6	2.4	0
rs1138357	G	A	0.94	3.2	0.52	56.6	35.2	8.2	2.4	0.56	58.5	34.1	7.4
rs1138358	T	G	1.00	0.8	0.46	57.6	34.4	8.0	0.8	0.61	58.4	34.4	7.2
rs3826007	G	A	0.64	7.1	0.38	57.3	34.2	8.5	4.8	0.73	61.7	32.5	5.8
rs3735485	C	T	0.74	3.2	0.99	73.8	23.8	2.4	3.2	0.27	77.0	19.7	3.3
rs7792760	G	A	1.00	0	0.20	74.6	21.4	4.0	0.8	0.44	75.2	21.6	3.2
rs2173904	G	C	0.60	2.4	0.65	33.3	46.4	20.3	7.1	0.72	28.0	47.5	24.5
rs2270891	G	T	1.00	1.6	0.49	92.7	6.5	0.8	2.4	0.49	92.7	6.5	0.8

Observed frequencies of genotypes in patients and donors separately. The minor alleles were defined as the alleles with the lowest frequency, whereas the major alleles were defined as the alleles with the highest frequency. Differences in genotype distribution between patients and donors for each nsSNP were assessed by Fisher's exact test. P values <.05 are in bold type.

Table 3.4: Distribution of genotypes

3.3 Results

3.3.1 Transplantation outcome

In our cohort of 126 patients, the median follow-up was 837 days (range, 30-3178 days). The 5-year OS and PFS were 58% and 49%, respectively. The probability of grade II-IV aGVHD within the first year was 69%, and the 3-year probability of extensive cGVHD was 44%.

3.3.2 Genotyping of patients

The patient and donor cohorts were successfully genotyped for 31 of the 53 selected nsSNPs, and a variation in the GVH direction was observed in 23 of these. There was no significant difference in the distribution of genotype frequencies between patients and donors, except for rs2061821 (P = .036) and rs1135791 (P = .046) (see Table 3.4). Sixteen of 23 nsSNPs adhered to the HWE (P > .05). Of the 7 polymorphisms that departed from the HWE, 6 were in strong LD located in *AKAP13* (see Table 3.5), and 1 was located in *SP110* (rs9061). Genotypes of SNPs that failed the HWE assumption were validated by direct sequencing of approximately 10% of the patient and donor cohorts. Furthermore, to ensure unbiased genotyping, the assay for these 7 nsSNPs was further validated in a cohort of 96 healthy controls (data not shown). When genotypes in the 96 control individuals were analyzed together with donor samples, all 7 nsSNPs adhered to HWE (data not shown).

Three of the nsSNPs (rs4843075, rs7162168, and rs10968457) that failed genotyping were inferred based on complete LD ($R^2 = 1$) according to the HapMap CEU population to some of the 23 varying nsSNPs, thus resulting in a total of 26 varying nsSNPs. In detail, rs4843075 and rs7162168 in the *AKAP13* protein were in LD with a block of 5 nsSNPs (rs4843074, rs2061821, rs2061824, rs4075256, and rs4075254), which were successfully genotyped. In the *KIA0020* protein, rs10968457 was in LD with rs2270891.

3.3.3 Effect of number of nsSNPs in the GVH direction on outcome

The median number of nsSNP differences in the GVH direction between patient and donor was 4 (range, 0-17). Patients with $\leq$4 nsSNP differences in the GVH direction had a nonsignificant higher 5-year OS and PFS than patients with >4 nsSNP differences (see Table 3.6 and Figure 3.1A). Likewise, patients with $\leq$4 nsSNP differences had a nonsignificant lower 5-year TRM than patients with >4 nsSNP differences (see Figure 3.1B). No difference in outcome was observed for any of the other clinical parameters (P > .30).

Patient/Donor	Rs2061821	Rs2061822	Rs2061824	Rs4075254	Rs4075256
Rs2061822	0.85 / 0.97				
Rs2061824	1/1	0.86 /1			
Rs4075254	1/1	0.86 /1	1/1		
Rs4075256	1/1	0.86 /1	1/1	1/1	
Rs4843074	1/1	0.85 /1	1/1	1/1	1/1

Table 3.5: Pairwise linkage disequilibrium, expressed as R between *AKAP13* polymorphisms out of HWE, in the patient and donor populations

Parameter	nsSNP Differences ≤ 4 vs > 4*	HR (95% CI)	*P*	Predicted miHAs ≤3 vs >3†	HR (95%CI)	*P*
OS	66.3 % vs 48.9 %	1.7 (0.9–3.1)	.09	70.1 % vs 42.2 %	2.3 (1.2–4.2)	**.0060**
PFS	54.5 % vs 43.2 %	1.5 (0.9–2.6)	.13	58.3 % vs 36.7 %	2.0 (1.2–3.5)	**.0082**
TRM	12.3 % vs 34.0 %	2.2 (0.9–5.5)	.09	9.9 % vs 39.2 %	3.4 (1.3–8.9)	**.0094**
RRM	24.4 % vs 25.9 %	1.3 (0.6–3.0)	.63	22.2 % vs 30.7 %	1.7 (0.7–3.8)	.39
RI	37.3 % vs 30.6 %	1.2 (0.6–2.3)	.76	34.8 % vs 34.6 %	1.5 (0.8–2.9)	.45
Acute GVHD grade II-IV	72.7 % vs 75.7 %	1.0 (0.7–1.5)	.75	71.8 % vs 76.9 %	1.1 (0.7–1.7)	.95
Acute GVHD grade III-IV	21.2 % vs 20.1 %	1.0 (0.4–2.1)	.82	19.6 % vs 22.0 %	1.2 (0.5–2.5)	.82
Extensive chronic GVHD	62.3 % vs 63.7 %	0.9 (0.5–1.6)	.41	62.2 % vs 63.8 %	1.0 (0.6–1.7)	.36

HR, hazard ratio; CI, confidence interval; *P* values <.05 are in bold type.
*Patient–donor pairs are divided into those with ≤4 or >4 nsSNP differences in the GVH direction.
†Patient–donor pairs are divided into those with ≤3 or >3 predicted miHAs.

Table 3.6: Univariate analyses of the effect of number of predicted mHags on different clinical outcome parameters after 5 years

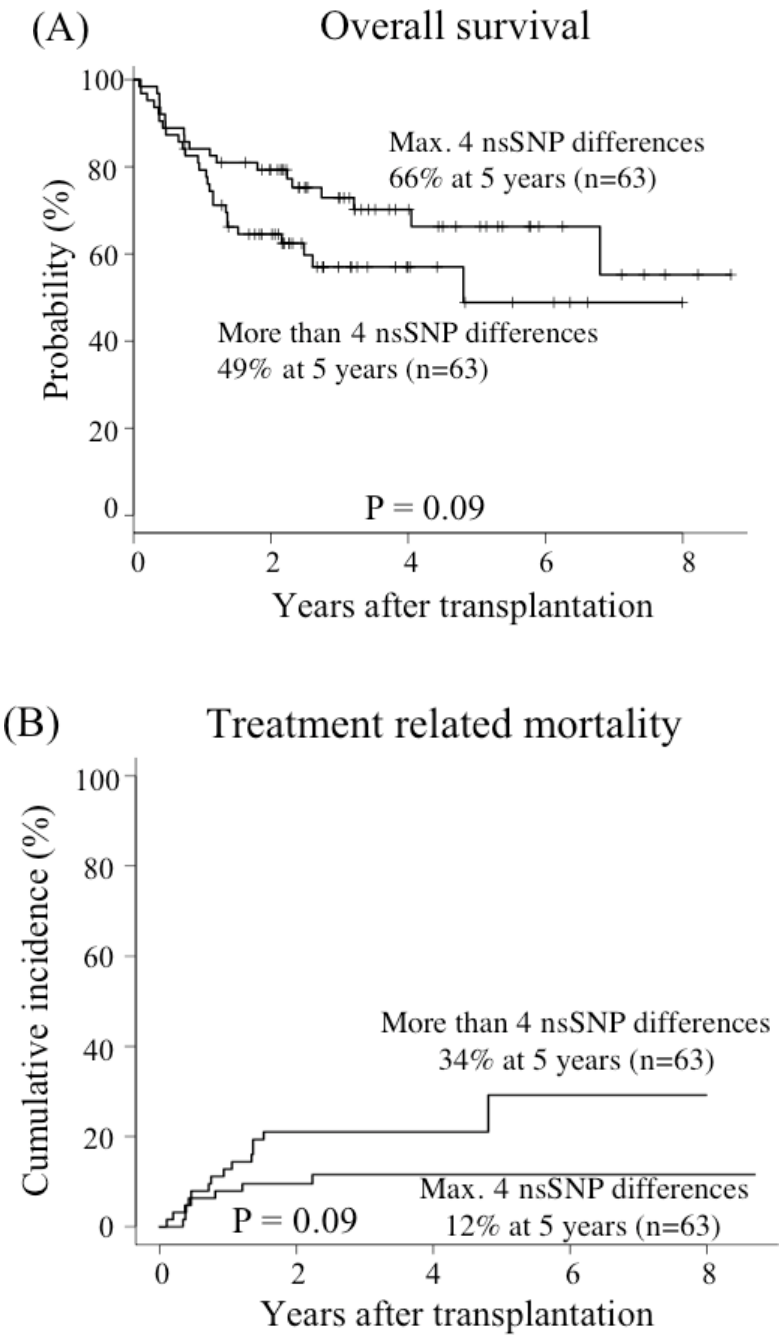


Figure 3.1: Probability of OS (A) and cumulative incidence of TRM (B) stratified according to the median number of nsSNP differences in the GVH direction.

3.3.4 Identification of potential mHags

A total of 26 nsSNPs within 6 of the 11 proteins showed variation in the GVH direction (see Table 3.3). Whenever a patient-donor pair had an nsSNP difference in the GVH direction, the binding of peptides containing the nsSNP to the patient's HLA-A and -B molecules was assessed using *NetMHCpan* (Nielsen et al., 2007). A binding strength threshold of 1% (binding strength falling within the top 1% compared with a large set of random natural peptides) was used in the analyses. Thresholds of 0.5% and 2% were tested without significantly altering the outcome of the analyses. In some cases, more than one peptide of 8-11 amino acids was a predicted binder, and thus each nsSNP in the GVH direction could result in 1, 2, or more potential mHags for a given patient. Significantly fewer nsSNP differences were present between patients and donors when the donor was matched related compared with matched unrelated (2 vs 7; $P < 10^{-5}$; Mann-Whitney U test). Similarly, related patient-donor pairs had fewer predicted mHags than unrelated pairs (1 vs 5; $P < 10^{-3}$; Mann-Whitney U test). Figure 3.2A shows the distribution of patient-donor pairs according to number of nsSNP differences in the GVH direction, and Figure 3.2B shows the distribution according to predicted mHags in the GVH direction.

3.3.5 Effect of number of predicted mHags in the GVH direction on outcome

There was a median of 3 predicted mHags per patient-donor pair (range, 0-16). A total of 215 mHags were predicted for the HLA types and nsSNP differences represented in our cohort (see Table 3.3), and 172 of these matched at least one patient. Patients with >3 predicted mismatched mHags in the GVH direction had a significantly lower probability of 5-year OS and PFS and a higher probability of 5-year TRM compared with patients with ≤ 3 predicted mHags (see Table 3.6 and Figures 3.3A and B). The presence of >3 predicted mHags also was a significant risk factor associated with 5-year OS, PFS, andTRM in both the unadjusted and adjusted Cox regression models (see Table 3.6 and Table 3.7).

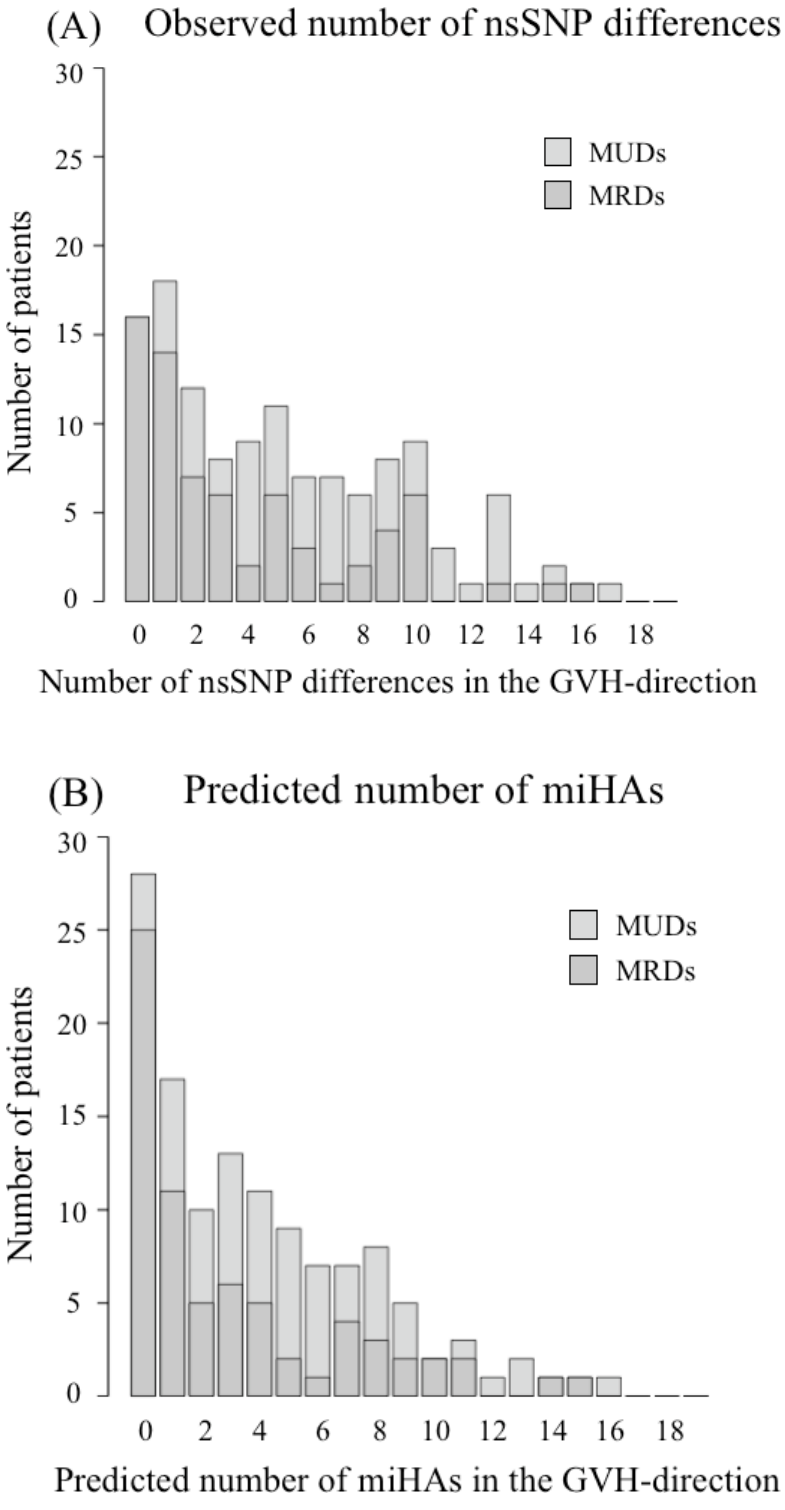


Figure 3.2: Histograms showing the distribution of patients by number of observed nsSNP differences in the GVH direction (A) and by number of predicted mHags in the GVH direction (B). MUD, matched unrelated donor; MRD, matched related donor.

Outcome	Covariate		HR	95% CI	*P*
OS	Number of predicted miHAs	≤3	Ref.		
		>3	2.2	1.2–4.0	**.014**
	Acute GVHD grade III-IV	Absence	Ref.		
		Presence	3.2	1.7–5.9	**<.001**
	Extensive chronic GVHD	Absence	Ref.		
		Presence	0.8	0.35–2.0	.67
PFS	Number of predicted miHAs	≤3	Ref.		
		>3	2.0	1.1–3.4	**.014**
	Acute GVHD grade III-IV	Absence	Ref.		
		Presence	2.7	1.5–4.8	**.001**
	Extensive chronic GVHD	Absence	Ref.		
		Presence	1.3	0.60–2.7	.525
TRM	Number of predicted miHAs	≤3	Ref.		
		>3	4.5	1.7–12.3	**.003**
	Patient age	≤40 years	Ref.		
		>40 years	3.8	0.5–29	.198
	Donor age	≤40 years	Ref.		
		>40 years	2.2	0.8–5.9	.126
	Acute GVHD grade III-IV	Absence	Ref.		
		Presence	4.4	1.9–10.6	**.001**
	Extensive chronic GVHD	Absence	Ref.		
		Presence	1.3	0.4–4.9	.669

OS indicates overall survival; PFS, progression-free survival; TRM, treatment-related mortality; GVHD, graft-versus-host disease; miHA, minor histocompatibility antigen.
Covariates were included in the final models only if they changed the estimate of the main variable by at least 10% or were significantly associated with outcome in pairwise analyses. *P* values <.05 are in bold type.

Table 3.7: Multivariate Cox regression analysis of the association of number of predicted mHags with 5-year transplantation outcome

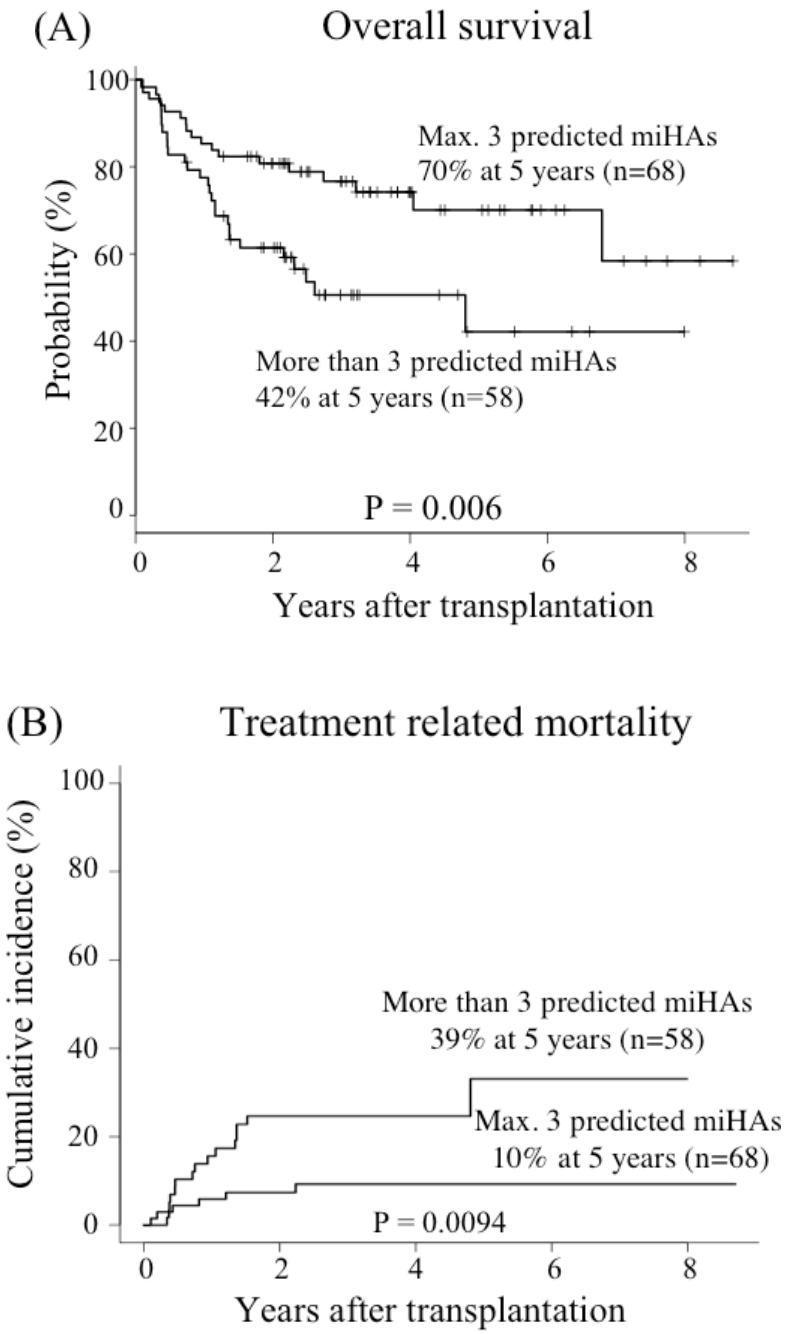


Figure 3.3: Probability of OS (A) and cumulative incidence of TRM (B) stratified according to the median number of predicted mHags in the GVH direction within a patient-donor pair.

No association between the number of mHags and any other clinical outcome parameter was observed. Other cutoffs besides the median of 3 predicted mHags were tested as well. The difference in OS and PFS was significant for all cutoffs between 0 and 6 predicted mHags (data not shown). The same was true for TRM, with the exception of a cutoff of 1 predicted mHag (P = .07) (data not shown). The probability of 5-year OS showed a successive decrease with 0, 1-2, and >2 predicted mHags per patient (hazard ratio [HR], 2.4; 95% confidence interval [CI], 1.5-4.0; P = .0005), implying a mHag dosage effect (see Figure 3.4). Patients with any predicted mHags had a significantly lower 5-year OS (46% vs 93%; HR, 8.1; 95% CI, 1.9-34; P = $.60 \times 10^{-3}$) than patients with 0 predicted mHags.

Apart from the association between the number of predicted mHags in a patient-donor pair and outcome, some protein-, nsSNP-, and predicted minor-specific associations with OS, PFS, or TRM were observed. The presence of any mHags in *SP110* and *AKAP13* (see Table 3.8), patient homozygosity for the minor allele of 3 nsSNPs in tight LD in *AKAP13* (rs2061821, rs2061822, rs4075254) (see Tables 3.5 and 3.8), and 4 predicted mHags were individually associated with outcome (Table 3.8). The multiple comparison penalty paid in these analyses increases the Bonferroni-corrected P values to well above the .05 threshold. According to dbMinor (Spierings et al., 2006), proteins *AKAP13* and *KIAA0020* are classified as broadly expressed, whereas *SP110*, *HMHB1*, *BCL2A1*, and *MYO1G* are classified as hematopoietically expressed. No tissue-specific effect was observed when dividing patients into those with predicted mHags only from hematopoietically expressed proteins, only from broadly expressed proteins, or from both kinds of proteins (5-year OS, 47% vs 49% vs 41%; P = .95).

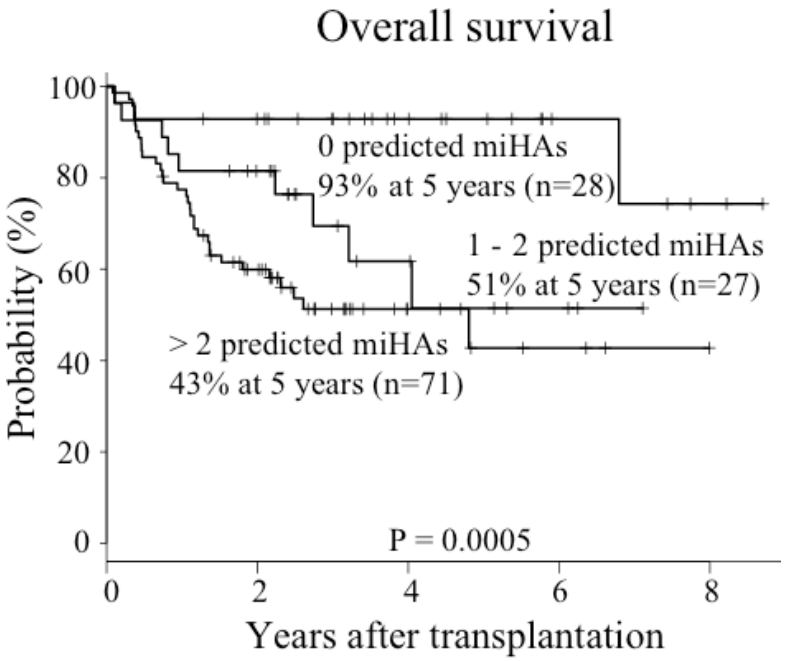


Figure 3.4: Probability of OS stratified according to number (0, 1-2, or >2) of predicted mHags in the GVH direction.

Protein	nsSNP Difference	Predicted miHA	Relevant Patients	*P* (OS)	*P* (PFS)	*P* (TRM)
SP110			45	**.025**	.058	.28
BCL2A1			29	.16	.11	.063
	rs1138357		25	.28	.30	.27
		YLQYVLQI*	13	.40	.90	**.025**
		RLAQDYLQYV	13	.40	.90	**.025**
	rs1138358		24	.17	.088	**.013**
		VLQKVAFSV	14	.51	.69	**.040**
AKAP13			49	**.041**	.062	.37
	rs2061821		26	.082	**.047**	.14
		LVMEPGTAQY†	13	**.0062**	**.0022**	**.0040**
	rs2061822		21	.41	.29	**.03**
	rs4075254		19	.20	.20	**.011**

OS indicates overall survival; PFS, progression-free survival; TRM, treatment-related mortality; GVHD, graft-versus-host disease; nsSNP, nonsynonymous single nucleotide polymorphism; miHA, minor histocompatibility antigen.
nsSNPs are listed under the protein in which they occur, and predicted miHAs are listed under the nsSNP that causes the miHA. Bold type denotes $P < .05$ (not corrected for multiple testing).
*Similar to the known miHA *ACC-1*:DYLQYVLQI.
†Similar to the known miHA *HA-3*:VTEPGTAY.

Table 3.8: Single predicted mHags, predicted mHags around a single nsSNP, and predicted mHags from a single protein with a significant impact on OS, PFS, or TRM

3.4 Discussion

To the best of our knowledge, the present study is the first to investigate the association between the number of predicted mHags in known mHag source proteins and clinical outcome after matched allogeneic HCT with NMA conditioning. By identifying nsSNP differences and using an Artificial Neural Network tool (*NetMHCpan*) 172 patient-donor specific mHags were predicted. Compared with the known HLA-A and -B binding mHags (n = 19; source dbMinor (Spierings et al., 2006)), this represents an almost 10-fold increase, suggesting that the investigated mHag source proteins contain additional mHags that have yet to be identified. Among the predicted mHags, 6 were already in the dbMinor database (Spierings et al., 2006): *HA-3* (VTEPGTAQY), *HA-8* (RTLDKVLEV), *HB-1H* (EEKRGSLHVW), *HB-1Y* (EEKRGSLY VW), *ACC-1* (DYLQYVLQI), and *ACC2* (KEFEDDIINW). The dbMinor database currently contains 29 mHags, of which 10 originate from the Y chromosome and thus were not considered in this study. We predicted only 6 of the remaining 19 previously identified HLA-A and -B binding mHags, because the corresponding nsSNP failed genotyping (6 mHags), the rs number was not listed in dbSNP (2 mHags), or the mHag was not caused by an nsSNP (5 mHags).

In line with the greater degree of genetic variation between unrelated individuals, significantly fewer nsSNP differences and predicted mHags in the GVH direction were observed with sibling patient-donor pairs compared with matched unrelated pairs. The patient-donor relationship did not significantly influence the transplantation outcome, however. When restricting the analysis to nsSNP differences in the GVH direction, it was possible to observe only a trend toward superior transplantation outcomes in patients with few nsSNP differences. But, when HLA restrictions were also taken into account by using mHag predictions, we were able to show that the presence of the median of $\leq$3 mHag disparities within a patient-donor pair, was a significant independent factor associated with a higher probability of both OS and PFS and lower risk of TRM. Although the group of patients with 0 predicted mHags had the highest OS compared with all other patients, the median was chosen to provide an unbiased level for dichotomization in the analyses, because the very few (2) events in the group of patients with 0 predicted mHags could affect the reliability of the statistical comparison when using 0 predicted mHags as the level of dichotomization.

These data suggest that the outcome of HCT depends on matching donor and recipient for HLA restricted mHags, rather than on the mere matching of nsSNPs. No association between the number of predicted mHags and aGVHD or cGVHD was observed. Although GVHD is considered one of the main causes of TRM (Ferrara et al., 2009), TRM also encompasses patients who succumbed to infection. Because it is unlikely that the number of predicted mHags is associated with the risk of infection without affecting the incidence of GVHD, the discrepancy between TRM and GVHD most likely results from insufficient study power. Given that no associations with relapse-related outcome measures were observed, our data suggest that the presence of many mHags confers an increased risk of death rather than inducing the beneficial GVT effect, implying that mismatching for most mHags results in decreased survival.

If the extent of interindividual genetic variation and HLA diversity is taken into account, then the current study assesses only a very limited subset of all possible predicted mHags. Because many of the predicted mHags likely will not initiate cytotoxic T cell responses because of immunodominance issues (Roopenian et al., 2002), it is of interest that the limited subset of mHags predicted in our study was associated with transplantation outcome. This could be explained by the assessed proteins being the source of most relevant mHags for transplantation (which we consider unlikely), or by the degree of mHag disparity in these proteins be-

ing a surrogate marker for the total genome-wide and HLA-wide mHag disparity within each patient-donor pair. If the degree of mHag disparity in our study only represents a proxy for the real patient-donor discrepancy, the exact level of dichotomization also becomes less important compared with making a distinction between few or many disparities. Apart from the impact of the predicted mHags on transplantation outcome, the possibility that factors such as the functional aspects of the nsSNPs cannot be excluded, and the general heterogeneity of the patient cohort also could influence the outcome in our cohort.

In several studies of single or very few patients, the identification and presence of mHag-specific T cells has been associated with remissions of chronic myelogenous leukemia (de Rijke et al., 2005; Marijt et al., 2003) or rejection (Voogt et al., 1990). However, larger studies of single or multiple mHag mismatches restricted to one or few HLA types in patients undergoing allogeneic HCT with sibling donors have uniformly been associated with GVHD without any association with RRM (Akatsuka et al., 2003b; Cavanagh et al., 2005; Goulmy et al., 1996; Grumet et al., 2001; Perez-Garcia et al., 2005; Tseng et al., 1999). In line with our study, this suggests that in general, mHag mismatch is not beneficial. In contrast, no association between mHag disparity and outcome was observed in a single study of 730 unrelated HLA-matched allogeneic HCTs (Spellman et al., 2009), possibly illustrating the impact of increased non-HLA genetic variation confounding the observations.

mHags have been classified into those with a restricted tissue expression encompassing tissues of hematopoietic origin and a broad tissue expression including non-hematopoietic tissues such as skin, gut, and liver (Bleakley and Riddell, 2004). It has been suggested that mHags with a restricted tissue expression would result in GVT effects without deleterious GVHD, because the GVHD elicited by such mHags would only result in the removal of normal recipient hematopoiesis. In contrast, mHags with a broad tissue expression would carry the risk of inducing potentially life-threatening GVHD. Among the 6 mHag source proteins showing nsSNP variation in the GVH direction, only *AKAP13* and *KIAA0020* (accounting for a total of 87 predicted mHags in our cohort) are classified as broadly expressed (source: dbMinor (Spierings et al., 2006)). The other 4 proteins - *SP110*, *HMHB1*, *BCL2A1*, and *MYO1G*, accounting for a total of 128 predicted mHags - have restricted tissue expression. However, a comparison of HCT outcome in patients with predicted mHags from broadly expressed proteins, proteins restricted to hematopoietic tissue, or both types of proteins showed no significant differences between the 3 patient groups, challenging either the experimental results on which the classification is based or the theoretical framework for separating GVHD and GVT effects (Bleakley and Riddell, 2004). Alternatively, in addition to creating mHags, the functional aspects of the nsSNPs also could decisively influence hematopoiesis and thus transplantation outcome. Therefore, it is possible that the current classification of mHags is simplistic and will require revision as our understanding grows.

Several limitations apply to the current study. Predictions were limited to HLA-A and -B molecules, because *NetMHCpan* is most accurate for these (Hoof et al., 2009). mHag predictions were planned for peptides surrounding 53 nsSNPs in 11 different non-Y chromosomal proteins. However, technical limitations because of both the genetic sequence surrounding the nsSNPs and the nature of the SNPstream genotyping platform limited the number of successfully genotyped nsSNPs to 31. Because the 53 nsSNPs at best only are surrogate markers for the common genetic variation between individuals, and because the 53 planned and 31 successfully genotyped nsSNPs probably represent similar fractions of the numerous potential nsSNPs in the entire genome, no further effort was made to pursue the genotype of the failed 22 nsSNPs. Although most genotypes adhered to HWE, 7 nsSNPs (6 of which were in strong LD) departed significantly. These observations are likely because of small sample size, be-

cause genotypes were confirmed by extensive resequencing, and because adherence to HWE was observed when a control population of 96 healthy blood donors was included. Furthermore, the significantly different distribution of rs2061821 and rs1135791 genotypes between patients and donors also was considered an artifact ascribed to the small study population, rather than a true association with disease susceptibility.

3.5 Conclusion

In conclusion, the current study presents a feasible method for large-scale *in silico* prediction of novel HLA-A- and -B-restricted mHags incorporating any patient-donor HLA types. Although the functional aspects of the predicted mHags are unknown and the study is purely descriptive, our findings suggest that the level of predicted mHag discrepancy between patient and donor could be associated with transplantation outcome. If these observations were to be validated in independent cohorts, mHag predictions specific for each patient-donor pair could have a place in future risk stratification and possibly in guiding donor selection and therapy.

With the current advancements in microarray-based genotyping, whole genome sequencing, and *in silico* modeling, a genome-wide and HLA-wide approach is within reach. mHag predictions could be expanded from encompassing only a few nsSNPs to all known nonsynonymous genetic variations and both class I and II HLA molecules, providing a unique mHag map of each patient-donor pair. This likely would enhance the prognostic value of the method in selecting the most optimal donor in those cases for which more than one 10/10 allele-matched donor was found using the current donor selection procedures.

Apart from the prognostic application, the large-scale mHag predictions also could function as a powerful tool in selecting candidate mHags involved in the GVT effect and GVHD for further evaluation in *in vitro* and *in vivo* experiments.

Chapter 4

Prediction of nsSNP derived mHags

The work described in this chapter is connected with the paper presented in Chapter 3. Whereas Chapter 3 concerns the correlation between predicted mHags and treatment outcome, this chapter describes the search for novel mHags around selected nsSNPs by predictions and subsequent experimental validations.

4.1 Introduction

The importance of mHags in relation to transplantation outcome has become well established through the last decade, as described in the previous chapters. Therefore, there is an ongoing effort to identify more mHags in general and mHags relevant to GVT in particular. Due to the vast number of HLA alleles and gene polymorphisms in the human population, there is an obvious advantage in using computerized methods to narrow down the mHag candidates to test experimentally. The scientific literature has seen several different computational approaches to mHag discovery including the use of GWAS, also described in Section 1.5.1. Kawase et al. (2008) used GWAS to identify a chromosomal region containing a novel mHag. Starting with an isolated CTL clone responding to lymphoblastoid cells of a patient but not of the donor, and therefore assumed to recognize an mHag, they divided a panel of lymphoblastoid cell lines into those recognized or not recognized by the CTL clones. After genotyping the cell lines with a SNP microarray, they used a GWAS to identify a chromosomal region differing between the two groups and showed that the mHag recognized by the CTL clone was located in this region. Another use of GWAS was illustrated by Ogawa et al. (2008) who correlated a number of SNPs and HLA alleles with GVHD in a cohort of ~1,600 transplanted patients indicating the exact location of potential mHags involved in GVHD.

MHC epitope prediction tools have also been used in the search for novel mHags. Schuler et al. have used the *SYFPEITHI* prediction algorithm (Rammensee et al., 1999) to create an online mHag predictor called *SNEP* (Schuler et al., 2005), which, for a limited number of HLA alleles, searches the *SWISS-PROT* database for 9mer epitopes around SNPs. Another online tool, similar to *SNEP*, called *SiPep* was developed by Halling-Brown et al. (2006) and is based on five different epitope predictors, namely *nHLAPred* (Bhasin and Raghava, 2007), *MHCPred* (Guan et al., 2003), *BIMAS* (Parker et al., 1994), *SYFPEITHI* (Rammensee et al., 1999), and *MMBPred* (Bhasin and Raghava, 2003). *SiPep* allows for advanced user queries taking tissue specific expression and cleavage prediction into account. The latest addition to the field of mHag predictors is called *PeptideCheck* (Deluca et al., 2009) and distinguishes itself by allowing high throughput analysis and integrating user-defined gene expression analysis.

The aim of this project was to apply *NetMHCpan* to identify potential mHags in selected proteins and verify these experimentally with ICS assays and tetramers using blood samples from transplanted patients. The strengths of using *NetMHCpan* compared to the specialized mHag predictors mentioned above are that *NetMHCpan* is more precise and can predict epitopes of 8-11 amino acids for all known HLA alleles, whereas some of the other predictors are limited to 9mers and do not include all HLA alleles. Due to limited resources, only proteins of special interest were selected for this study, comprising 16 proteins, in which mHags have previously been identified, and 14 proteins selected with the discovery of potential GVT mHags in mind.

The main focus of my part of this project, was to use *NetMHCpan* to identify the most promising mHag candidates in these selected proteins in relation to our patient cohort. As in Chapter 2, due to limited resources, the bioinformatical challenge of this task was to narrow down the number of raw predictions while keeping the ones, most likely to elicit a response in as many patients as possible.

4.2 Materials, methods and prediction results

4.2.1 Patients

This analysis includes data from 164 patients of which 126 with available follow-up data were used in the study described in Chapter 3. The patients were all treated with an allo-HCT with a peripheral blood graft from an HLA-identical related or 10/10 allele-matched unrelated donor after NMA conditioning between years 2000 and 2008 at the allo- HCT unit, Department of Hematology, Rigshospitalet, Copenhagen. For related donors, donor selection was based on serologic typing for HLA-A, -B, and -C, and on molecular typing for HLA class II. For unrelated donors, donor selection was based on molecular typing for HLA-A, -B, -C, -DRB1, and -DQB1. When available, HLA-identical siblings were preferred to matched unrelated donors, and cytomegalovirus serostatus and sex mismatch were taken into account when possible. All patients were treated for a malignant hematologic disease. Donor treatment, conditioning regimen, and supportive care were as described in (Kornblit et al., 2008).

4.2.2 Selected proteins

30 proteins that could be expected to contain mHags were selected for this small-scale study. Of these, 16 contain known mHags and 14 were selected due to their hematopoietic or cancer related expression. 5 of the 16 known mHag source proteins are located on the Y chromosome, while the 11 proteins also used in the study described in Chapter 3 are located on the autosomal chromosomes. Table 4.1 gives an overview of the selected proteins.

4.2.3 Prediction of mHags

NetMHCpan (Nielsen et al., 2007) was used for the prediction of potential mHags in the 30 selected proteins. *NetMHCpan* was run with all possible combinations of the 46 different HLA-A and -B alleles represented by the patients and all possible peptides of lengths 8-11 amino acids containing a SNP from one of the 30 proteins. A peptide with the reference amino acid at the SNP-position and the corresponding peptide with the missense amino acid constitute a peptide pair. Only those peptide pairs where at least one of the peptides was predicted to bind any of the HLA alleles with an affinity less than 500 nM were considered for further analysis.

Proteins with known mHags	Expression	Additional proteins	Expression
AKAP13	Broad	**BCL6**	B cell leukemia
SP110	Hematopoietic	**CD99**	T cell specific
BCL2A1	Hematopoietic	**TYR**	Melanoma
KIA0020	Broad	**MAGEA1**	Melanoma
MYO1G	Hematopoietic	CD3D	T cell specific
HMHB1	B cell specific	CD79B	B cell specific
USP9Y	Broad	CMRF35	Hematopoietic
DDX3Y	Broad	IL2	Hematopoietic
RPS4Y1	Broad	TP53	Tumor specific
SMCY	Broad	WT1	Tumor specific
UTY	Broad	TAL1	T cell Leukemia
HMHA1	Hematopoietic	MPL	Leukemia
CTSH	Broad	NOV	Broad/Cancer
ECGF1	Broad	PLAT	Endothelial cells
LHR1	Tumor specific		
TOR3A	Broad		

Table 4.1: **Proteins selected for mHag predictions.** In peptides marked in bold, nsSNP disparities were found in the patient cohort. Emphasized proteins are from the Y chromosome.

The number of such raw mHag predictions was 3,520 peptide/HLA combinations comprising 1,278 distinct peptide pairs around 173 nsSNPs.

4.2.4 Genotyping of patients

The patients were genotyped for the selected 173 nsSNPs with predicted mHags by means of the SNPstream genotyping system described in Chapter 3. For technical reasons, it was only possible to determine 120 of the 173 SNPs with this method. A nsSNP was considered to be of interest if it varied in the GVH direction in any patient-donor pair, such that the donor was homozygous for one allele and the patient was either heterozygous or homozygous for the other allele. In total, 33 of the successfully genotyped SNPs were found to vary in the GVH direction in this patient set corresponding to 28%. Failed or missing genotype values could in some cases be inferred from the 33 varying SNPs, in particular 3 SNPs that failed in all patients were inferred. The criterion for inferring genotypes in this way was complete LD (R^2=1) using the CEU population in the HapMap database (Consortium, 2003). Thus, 36 SNPs found in 252 of the predicted peptide pairs were selected for further analysis.

4.2.5 Patient subset

A subset of patients was selected due to a limited number of available patient blood samples. Thus, 105 out of 164 patients with a promising number of varying SNPs and peptides were selected for experimental validation. In detail, more than 5 predicted mHags, fitting the nsSNP variation and HLA alleles of the patient-donor pair, should be possible to test for each patient. Additionally, patients which were sex-mismatched were all included since they were already selected for the experimental validation of mHags from the Y chromosome as described in Chapter 2.

4.2.6 Ideal peptide selection

Ideally, all peptides matching any patient on both HLA allele and a SNP-variation in the GVH-direction should be tested. The ideal selection consists of 239 such potential mHags which can be found in Appendix C. In principle, it is only necessary to test the potentially immunogenic peptide of a peptide pair. However, often both peptides in a pair are included in the ideal selection, since they match different patients. If an immune reaction is observed, the non-immunogenic peptide should be acquired and tested as a negative control. An overview of the ideal selection is given in Table 4.2.

4.2.7 Submer filtering and final selection

Many of the predicted mHags in the ideal selection are very similar and differ only in length. To avoid purchasing almost identical peptides, we first chose those peptides in a family of similar peptides that were most likely to elicit a T cell response. This submer filtering works as follows:

- If the amino acid sequence of a peptide is a submer of the sequence of a longer peptide, only the peptide that matches most patients is kept. If the two peptides match almost (within 75 %) the same number of patients then the longest peptide is kept.
- All (4) known mHags are kept.
- 10mers or 11mers with at least 4 matching patients cannot be filtered out by 8mers or 9mers.

These filtering rules are intended to make sure that the peptides in the final selection are not too similar while optimizing the chance of observing a T cell response. The reason for this bias in the rules towards the longer version of the peptides is that peptide cleaving takes place in the ICS assays. Thus even though only the longer peptide version is mixed with PBMCs of a patient, the shorter versions are produced and can be recognized. Table 4.2 and Appendix C gives an overview of the final selection, which consists of 128 peptides.

4.3 Testing scheme

Although the peptide selection criteria relies on perfect matches with patient HLA alleles and nsSNP disparities, an immune response is possible even if the predicted affinity is above 500 nM. Therefore selected peptides, fitting a given patient's relevant nsSNPs, should still be tested even if they are not predicted to bind the patient's HLA alleles. Whenever an immune response is seen, the control peptide and possible peptide substrings should be tested as well. ICS and tetramer validations are now ongoing, using the same experimental methods as described in Chapter 2 and we await the validation results in the near future.

Protein	nsSNP	No. of peptides in the ideal peptide selection	No. of peptides in the final peptide selection
AKAP13	rs2061821	6	2
AKAP13	rs2061822	8	4
AKAP13	rs2061824	3	1
AKAP13	rs34434221	3	2
AKAP13	rs35624420	5	2
AKAP13	rs4075254	11	6
AKAP13	rs4075256	8	2
AKAP13	rs4843074	5	5
AKAP13	rs4843075	3	2
AKAP13	rs7177107	4	2
AKAP13	rs7162168	10	6
AKAP13	rs745191	5	4
Sum	**12**	**71**	**38**
SP110	rs9061	8	5
SP110	rs1135791	13	5
SP110	rs3948463	11	6
SP110	rs3948464	9	4
SP110	rs28930679	4	3
Sum	**5**	**44**	**23**
BCL2A1	rs1138358	16	7
BCL2A1	rs1138357	15	6
BCL2A1	rs3826007	3	3
Sum	**3**	**34**	**16**
KIA0020	rs2173904	15	8
KIA0020	rs2270891	7	5
KIA0020	rs10968457	4	3
Sum	**3**	**26**	**16**
MYO1G	rs7792760	7	5
MYO1G	rs3735485	10	5
Sum	**2**	**17**	**10**
HMHB1	rs161557	8	6
USP9Y	rs7067496	1	1
TYR	rs33955261	1	1
TYR	rs1042602	14	8
TYR	rs13312740	3	1
TYR	rs1126809	15	7
Sum	**4**	**33**	**17**
CD99	rs11556080	1	1
BCL6	rs2229362	2	1
MAGEA1	rs2008144	1	1
Total	**34**	**239**	**130**

Table 4.2: **Number of peptides in the ideal vs. the final peptide selection.** If variations are found for more nsSNPs within the same gene, sums are calculated for the gene. The last 4 genes do not contain known mHags.

4.4 Discussion and outlook

In this small-scale study aiming at identifying novel mHags, we focused on proteins known to contain mHags and proteins with a hematopoietically restricted or cancer related expression. nsSNPs within the selected proteins were genotyped in the patient cohort if mHag candidates could be predicted around the nsSNPs for any of the HLA alleles represented by the cohort. 36 nsSNPs were shown to be relevant, meaning that at least one patient was positive for a variant of the nsSNP not found in the corresponding donor.

As in the study described in Chapter 2, it was necessary to narrow down the number of mHag candidates to validate experimentally. The submer filtering used here was more complex than the one described in Chapter 2. The main reason for this was the fact that here, nsSNPs are the source of the predicted mHags, while for the Y chromosome they are caused by differences between genes on the Y chromosome and their homologues on the X chromosome. This means that all sex-mismatched patient/donor pairs represent the same H-Y peptide differences, assuming that all the male patients shared the same Y chromosome. nsSNP differences are, instead, unique for each patient/donor pair, meaning that each of the predicted mHags typically match fewer patients. Therefore, we chose to design the submer filtering, to optimize the number of patients matching each of the selected peptides. Nevertheless, the filtering procedure was given some bias towards the longer versions of the peptides, since peptide cleavage occurs in the ICS assays.

As a follow-up to this small-scale reverse immunology study, the patients have now been genotyped with the Illumina *HUMANOMNI1-QuAD v1 microarray*, which contains more than 1 million genomic markers including ~32,000 nsSNPs. The data will be used for the identification of novel mHags in a prediction-based approach similar to the one presented in this chapter, although at a much larger scale, and with particular focus on the identification of therapeutically relevant mHags. Additionally, analyses similar to those described in Chapter 3 will be carried out, to verify the correlations between number of mHags and transplantation outcome at a larger scale. The perspectives of the microarray data are described in greater detail in Chapter 6.

Chapter 5

HLArestrictor – a tool for patient-specific predictions of HLA restriction and epitopes

This chapter presents an online prediction method called *HLArestrictor*, which is designed for the prediction of optimal epitopes and corresponding HLA restriction within peptides or proteins using a patient-oriented approach.

The idea for the project emerged during our work on patient-specific prediction of mHags, described in Chapters 2, 3, and 4. We realized that *NetMHCpan* was not optimized to address the common scientific question of identifying the HLA restriction element and minimal epitope within a longer peptide, which had been observed to elicit an CTL response in a given patient.

HLArestrictor is based on *NetMHCpan* and offers a highly flexible overview of the prediction results. Compared to *NetMHCpan*, *HLArestrictor* is capable of predicting peptides of 8-11 amino acids simultaneously and offers different sorting options, allowing the user to tailor the output according to different needs and questions of interest.

HLArestrictor was benchmarked on a large dataset of HIV IFNγ ELIspot responses, where it was shown to identify HLA restrictions and minimal epitopes for ~90% of the peptide/patient pairs. Additionally, it was benchmarked on a smaller dataset of 18 tetramer validated responses, for all of which it correctly predicted both the HLA restriction element and minimal epitope. Furthermore, the ability of *HLArestrictor* to identify the correct restriction element was validated using a set of HLA restrictions identified through association studies.

My efforts in the work presented in the following paper involved the development of *HLArestrictor* and benchmark analyses of HIV ELIspot and tetramer data. I was main responsible for writing the manuscript, which was submitted to *Immunogenetics* in August 2010.

HLArestrictor – a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides or proteins

Malene Erup Larsen[1], Henrik Kløverpris[2], Anette Stryhn[3], Catherine K. Koofhethile[2], Stuart Sims[2], Thumbi Ndung'u[4,5], Philip Goulder[2], Søren Buus[3], Morten Nielsen[1]

[1]Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark
[2]University of Oxford, Peter Medawar Building for Pathogen Research, Oxford, England
[3]Laboratory of Experimental Immunology, University of Copenhagen, Denmark
[4]HIV Pathogenesis Programme, Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa
[5]Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology and Harvard University, USA

Abstract

Traditionally, T cell epitope discovery requires considerable amounts of tedious, slow and costly experimental work. During the last decade, prediction tools have emerged as essential tools allowing researchers to select a manageable list of epitope candidates to test from a larger peptide, protein or even proteome. However, no current tools addresses the complexity caused by the highly polymorphic nature of the restricting HLA molecules, which effectively individualizes T cell responses. To fill this gap, we here present an easy-to-use prediction tool named *HLArestrictor* (http://www.cbs.dtu.dk/services/HLArestrictor), which is based on the highly versatile and accurate *NetMHCpan* predictor, which here has been optimized for the identification of both the MHC restriction element and the corresponding minimal epitope of a T cell response in a given individual. As input, it requires high-resolution (i.e. 4-digit) HLA typing of the individual. *HLArestrictor* then predicts all 8-11mer peptide-binders within one or more larger peptides and provides an overview of the predicted HLA restrictions and minimal epitopes. The method was tested on a large dataset of HIV IFNγ ELIspot peptide responses, and was shown to identify HLA-restrictions and minimal epitopes for about 90% of the positive peptide/patient pairs, while rejecting more than 95% of the negative pairs. Furthermore, for 18 peptide/HLA tetramer validated responses, *HLArestrictor* in all cases predicted both the HLA restriction element and minimal epitope. Thus, *HLArestrictor* should be a valuable tool in any T cell epitope discovery process aimed at identifying new epitopes from infectious diseases and other disease models.

Keywords: HLA restriction · epitope prediction · MHC class I · peptide binding · T cell epitope validation · HLA tetramer validation

5.1 Introduction

CD8 positive cytotoxic T lymphocytes (CTLs) identify and eradicate host cells that have been infected with intracellular pathogens. They recognize protein antigen-derived peptides presented in complex with major histocompatibility complex (MHC) class I molecules. Prior to presentation, protein antigens are processed in a series of events beginning with the degradation of intracellular proteins by the proteasome (Rock et al., 2002), followed by transporter associated with antigen processing (TAP)-mediated peptide translocation into the endoplasmic reticulum (ER) (Townsend and Trowsdale, 1993; Uebel and Tampe, 1999), N-terminal shortening of longer peptides by ER-resident amino peptidases (Serwold et al., 2002), and eventually some of the resulting peptides are specifically bound to MHC class I molecules. Once a stable peptide/MHC complex has been formed, it is transported via the Golgi apparatus to the cell surface ready for inspection by circulating CTLs.

Considering the many different peptides that can be generated, even from a small target protein, and the extensive polymorphism of the presenting MHC molecules, identifying pathogen-specific, HLA-restricted T cell epitopes can be an immense experimental task. However, only a few percent of a random collection of peptides can bind with sufficient affinity to a particular MHC-molecule making this event the most selective step in the entire pathway of antigen presentation (Yewdell and Bennink, 1999), and a suitable starting point for T cell epitope discovery. Indeed, predictions of peptide/MHC interactions are widely used as an aid to identify T cell epitopes. *NetMHCpan* (Hoof et al., 2009; Nielsen et al., 2007) is one of the most precise publicly available predictors of peptide binding to MHC class I molecules (Lin et al., 2008a; Zhang et al., 2009), and it has the added advantage that it is capable of addressing any known HLA molecule.

As described above, T cell epitope discovery involves the concurrent identification of stimulating antigenic peptide and their restricting HLA elements. In human populations, several thousand allelic HLA-A, -B, and -C variants have already been registered (Robinson et al., 2001). For any given human individual, a complete CTL epitope discovery effort would have to consider up to six (three loci with two heterozygous alleles each) different restricting HLA class I molecules. Fortunately, current DNA sequencing-based technology allows high-resolution (i.e. 4-digit e.g. HLA-A*0201) typing of all HLA-A, -B, and -C alleles of any given individual. Thus, information on all the HLA class I types needed to perform *NetMHCpan* predictions for any given individual can readily be provided. The other piece of information needed for *NetMHCpan* predictions is the input proteome, protein or peptide. Whereas the number of HLA class I molecules can be limited to six, the number of peptides under consideration may be truly staggering; a problem, that is compounded by the ability of HLA class I molecules to bind peptides of different length (note, *NetMHCpan* can handle peptides of 8-11 amino acids in length). To reduce this complexity, one could conveniently exploit a commonly used approach of T cell epitope discovery: testing pools of overlapping peptides (OLP) with a length of 15-18 amino acids in IFNγ ELIspot or flow cytometry assays.

We here present a new online tool, *HLArestrictor*, aimed at identifying optimal peptides within one or several input peptides and corresponding HLA class I restriction elements targeted by CTL in given individuals. For a given individual, who has been fully typed for HLA-A, -B and -C, *HLArestrictor* is designed to identify all potential epitopes of length 8-11 amino acids that are predicted to bind to at least one of the individuals HLA restriction elements. A number of different sorting options are available for providing a user-friendly output.

We have benchmarked *HLArestrictor* with an HIV dataset of 5,145 18mer peptide IFNγ ELIspot responses from 694 treatment-naïve HIV infected individuals and could successfully

predict about 90% of the T cell epitopes. Using peptide/MHC class I tetramers, we furthermore demonstrated that *HLArestrictor* is able to correctly identify both the HLA restriction element and the optimal peptide length of a T cell epitope. The latter suggest that *HLArestrictor* could be an ideal design tool of HLA-tetramer for the T cell epitope discovery.

Input parameter	**Default value**	**Description**	**Letter in Figure 5.1**
Input / file	-	Input peptides in FASTA format/file.	A
List of HLAs	HLA-A0201	Comma separated list of HLA alleles.	B
Peptide lengths	8-11	Length(s) of submers to extract from input peptide.	C
Sort mode	HLA-oriented	Sort mode to use for showing the output: HLA-oriented, peptide oriented, or Descending prediction scores. Each mode is also available as pool version.	D
Sort method	%rank OR affinity	Specifies if sorting should be done using %rank or affinity values, and whether the binding criteria should be based on thresholds for %rank, affinity, %rank OR affinity, or %rank AND affinity.	E
Strong threshold	0.5 %rank or 50nM	Threshold for strong binding in %rank or affinity.	F
Weak threshold	2 %rank or 500nM	Threshold for weak binding in %rank or affinity.	F
Number of predictions per peptide	Off	Optional user defined maximal number of predictions to show per peptide. Predictions below weak binding threshold will, however, always be shown.	G
Non-binders factor	2	Show non-binders with a score of factor$\times$weak binding threshold.	H

Table 5.1: *HLArestrictor* features.

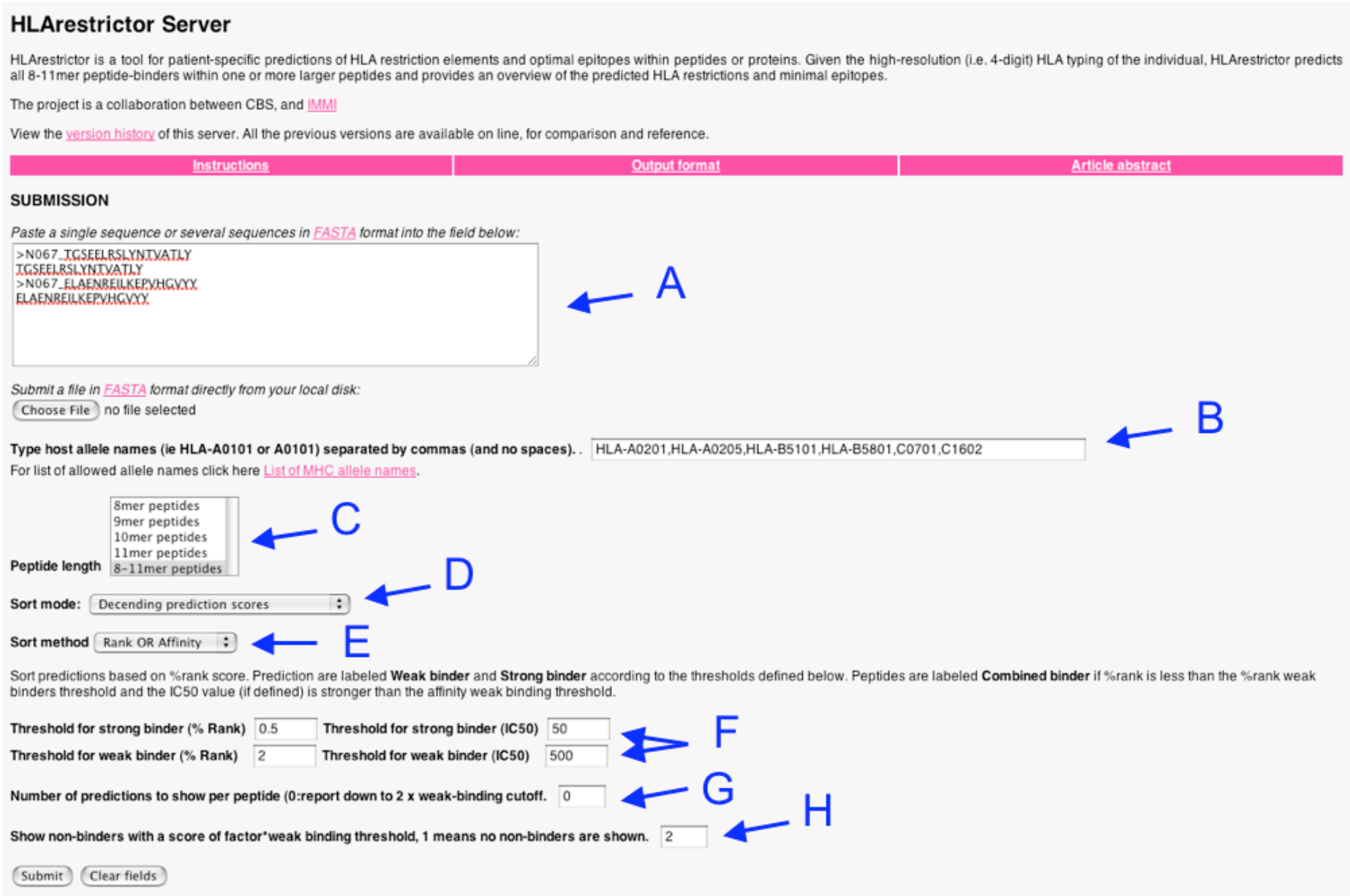


Figure 5.1: Screenshot of *HLArestrictor*. The input fields are marked in blue letters and described in Table 5.1 and in the text. Note, that multiple FASTA sequences can be given as input (see A), and that HLA-types can be given with or without the prefix "HLA-" (see B)

5.2 Materials and methods

5.2.1 *HLArestrictor* features

Using *NetMHCpan* version 2.2, *HLArestrictor* predicts all peptide binders of length 8-11 amino acids which are relevant for a given patient's HLA-types within an input peptide. A detailed overview of features of the *HLArestrictor* method is given in Table 5.1 and the corresponding input fields of the online *HLArestrictor* predictor are marked on Figure 5.1. Multiple input peptides can be given in the same file using the FASTA format. There is no principal limit neither for the length of the input peptide nor for the number of HLA alleles to predict for. Thus, *HLArestrictor* is not limited to its intended use: patient-specific predictions of epitopes within a peptide or a peptide pool, but can be applied to complete protein sequences or proteomes. Three different sort modes are available:

- *HLA-oriented*, which groups predicted peptide/HLA pairs by HLA allele and sorts by prediction score (see Figure 5.2)
- *Peptide oriented*, which groups predicted peptide/HLA pairs by peptide and sorts by prediction score (see Figure 5.3)
- *Descending prediction scores*, which sorts predicted peptide/HLA pairs by prediction score without any grouping (see Figure 5.4)

Each of the three sort modes is available as a standard version, where predictions are shown separately for each input peptide, and as a pool version, where predictions are shown together. This latter mode is useful when predicting responses for a larger peptide pools.

Finally, a "no-sort" option is available for optimal computational speed. Three binding thresholds are defined and labeled; Strong binder, Weak binder and Non-binder. A fourth label Combined binder is used to indicate predictions which do only qualify as binders if both %rank and affinity thresholds are considered as described below. Non-binders are shown up to a user-defined factor (default = 2) of the weak binding threshold. Additionally, the user can define a maximum number of predictions to show per input peptide, with the exception that all predicted strong or weak binders will always be shown.

Binding thresholds and sorting are based on either the predicted binding affinity (in nM) or a predicted %rank score, while both values are always shown in the output. The %rank score is defined as the rank score of a given candidate peptide relative to a set of 1 million random natural peptides for a given allele, such that a 2 %rank score means that only 2% of random peptides bind the allele with a predicted affinity stronger than the candidate epitope. If the default sort method Rank OR Affinity is chosen, a prediction need to fulfill either the affinity or %rank at the strong or weak binding thresholds to be labeled as Strong binder or Weak binder, respectively. This is a practical feature when the user wants to be alerted by predictions meeting either threshold and would consider subsequently testing all suggested epitopes. In contrast, if the option Rank AND Affinity is chosen, a prediction need to fulfill both the affinity and %rank at the strong or weak binding thresholds to be labeled as Strong binder or Weak binder, respectively.

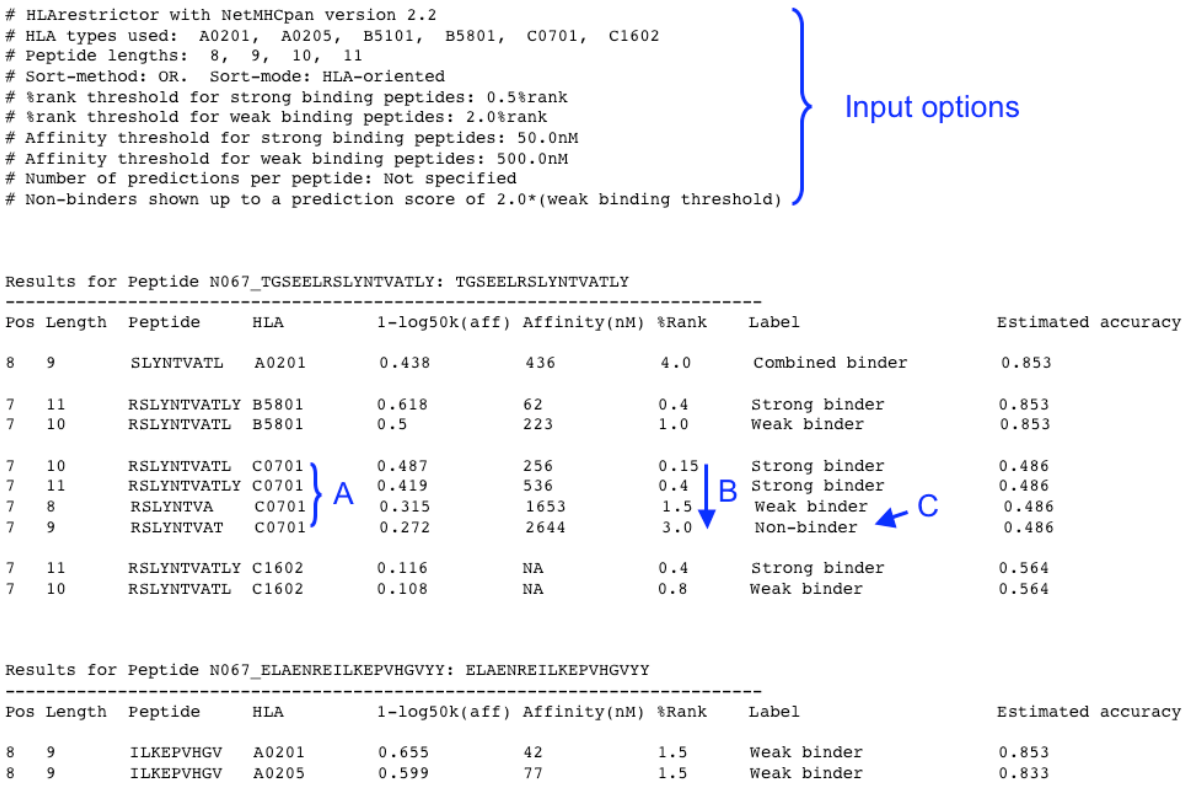


Figure 5.2: *HLArestrictor* example output using sort-mode *hla-oriented*, by which the predictions are grouped by HLA allele, then ranked by prediction score. In this example all predictions for HLA-C0701 are listed in the second group marked (A). Ranking within the group is marked with (B). Non-binders (C) are shown if their %rank value is less than 2 times the threshold for weak binding (2 %rank). All threshold values are user-defined.

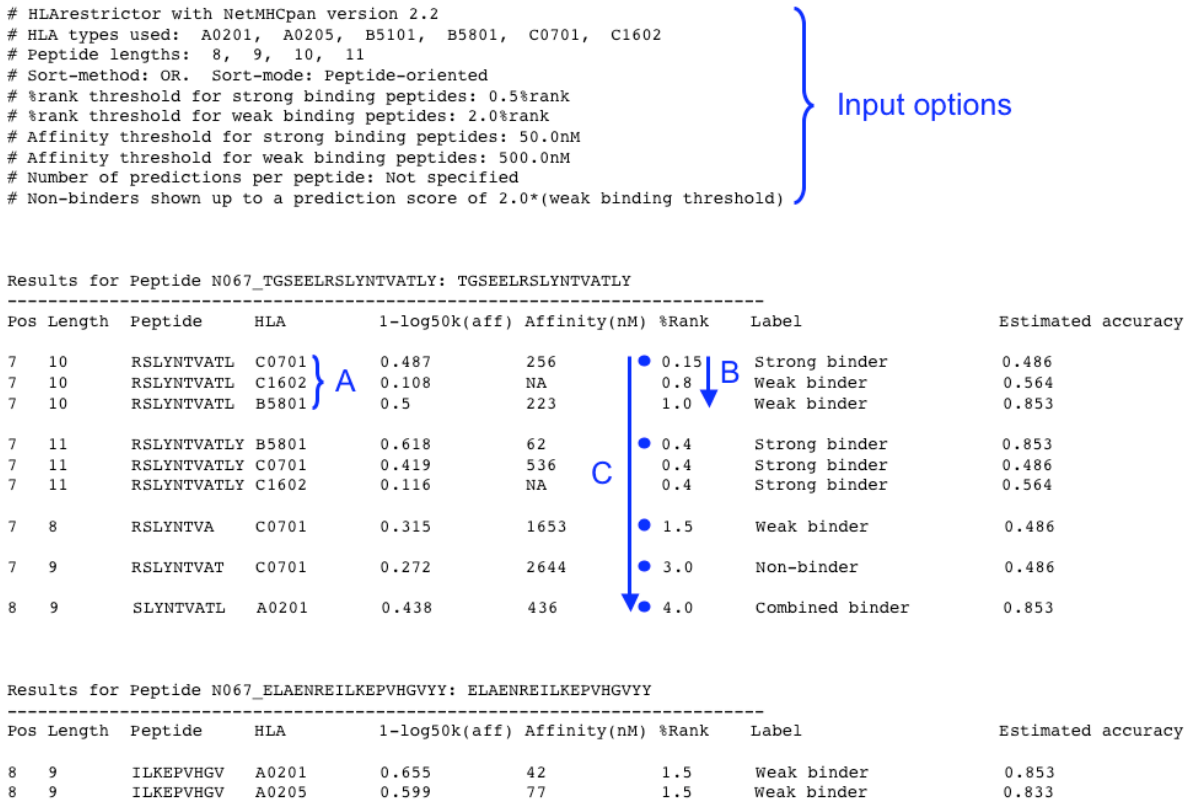


Figure 5.3: *HLArestrictor* example output using sort-mode *peptide-oriented*, by which the predictions are ordered such that multiple predictions of the same sub-peptide are grouped together, then ranked according to prediction score. In this example, all predictions for peptide RSLYNTVATL are listed in the group marked (A). Ranking within the group is marked with (B). Ranking between groups is marked with (C) and is based on the best scoring values marked with blue bullets.

```
# HLArestrictor with NetMHCpan version 2.2
# HLA types used:  A0201,  A0205,  B5101,  B5801,  C0701,  C1602
# Peptide lengths:  8,  9,  10,  11
# Sort-method: OR.  Sort-mode: Descending prediction scores
# %rank threshold for strong binding peptides: 0.5%rank
# %rank threshold for weak binding peptides: 2.0%rank
# Affinity threshold for strong binding peptides: 50.0nM
# Affinity threshold for weak binding peptides: 500.0nM
# Number of predictions per peptide: Not specified
# Non-binders shown up to a prediction score of 2.0*(weak binding threshold)


Results for Peptide N067_TGSEELRSLYNTVATLY: TGSEELRSLYNTVATLY
-------------------------------------------------------------------------
Pos Length  Peptide     HLA         1-log50k(aff) Affinity(nM) %Rank    Label                Estimated accuracy

7   10      RSLYNTVATL  C0701       0.487         256          0.15     Strong binder        0.486
7   11      RSLYNTVATLY B5801       0.618         62           0.4      Strong binder        0.853
7   11      RSLYNTVATLY C0701       0.419         536          0.4      Strong binder        0.486
7   11      RSLYNTVATLY C1602       0.116         NA           0.4      Strong binder        0.564
7   10      RSLYNTVATL  C1602       0.108         NA           0.8      Weak binder          0.564
7   10      RSLYNTVATL  B5801       0.5           223          1.0      Weak binder          0.853
7   8       RSLYNTVA    C0701       0.315         1653         1.5      Weak binder          0.486
7   9       RSLYNTVAT   C0701       0.272         2644         3.0      Non-binder           0.486
8   9       SLYNTVATL   A0201       0.438         436          4.0      Combined binder      0.853


Results for Peptide N067_ELAENREILKEPVHGVYY: ELAENREILKEPVHGVYY
-------------------------------------------------------------------------
Pos Length  Peptide     HLA         1-log50k(aff) Affinity(nM) %Rank    Label                Estimated accuracy

8   9       ILKEPVHGV   A0201       0.655         42           1.5      Weak binder          0.853
8   9       ILKEPVHGV   A0205       0.599         77           1.5      Weak binder          0.833
```

Figure 5.4: *HLArestrictor* example output using sort-mode *descending prediction score*, by which the predictions are ordered solely by their prediction score, in this case their %rank score. Note, that sorting by affinity is different than sorting by %rank, since the affinity distribution of ranked peptides is specific for each allele.

5.2.2 HIV benchmark set

We used a cohort of 1,000 antiretroviral naive HIV infected adults, of whom 864 were recruited from Durban, KwaZulu-Natal, South Africa (Kiepiela et al., 2007) and 136 recruited from Thames Valley, Oxfordshire, England. Informed consent was obtained from all participating individuals and institutional review boards at the University of KwaZulu-Natal, Massachusetts General Hospital, and the University of Oxford approved the study. Four-digit high resolution typing of HLA-A, HLA-B and HLA-C alleles was performed using the Dynal RELTIM reverse sequence-specific oligonucleotide (SSO) kits as previously described (Kiepiela et al., 2007). 694 patients were successfully typed on all alleles, and were included in the benchmark set.

5.2.3 IFNγ ELIspot

Comprehensive IFNγ ELIspot responses to a set of 420 overlapping peptides (OLPs) based on the 2004 C-clade consensus were used in a matrix system with 11-12 peptides in each pool. Responses to matrix pools were subsequently confirmed by stimulating with individual peptides as previously described (Kiepiela et al., 2004). A total of 5,145 ELIspot responses to a total of 294 different 18mer peptides were observed when measured in the 694 treatment-naïve HIV infected individuals.

5.2.4 Peptide/MHC class I tetramer synthesis and tetramer staining

Tetramers were generated in two different ways. The first method was described by Altman et al. (1996). Briefly, HLA-B*4201 heavy chain (HC) was expressed in Rosetta(DE3)pLysS (Novagen), purified and refolded around the peptide of interest in the presence of human β2M light chain. Unrefolded HC and peptide were separated from refolded peptide/MHC monomer complexes using FPLC prior to tetramerization of monomers and conjugation to R-phycoerythrin (Extravidin PE, Sigma) to obtain PE labeled HLA-B*4201 tetramers.

The second method was recently described by Stryhn and coworkers (Leisner et al., 2008). Briefly, HLA-class I heavy chain were expressed in E.coli BL21(DE3), which had been co-transfected with a vector encoding the BirA holoenzyme, leading to the expression of biotinylated HLA-class I heavy chain when the proteins were induced in the presence of biotin. Pre-oxidized, pre-biotinylated isomers were purified by column chromatography, and stored at -20°C until use. Peptide/HLA monomers were made by incubating these highly active isomers in a refolding buffer with excess b2m and peptide. Peptide/HLA tetramers were then generated by the addition of PE labeled streptavidin. PBMCs or CTL lines were thawed and stained with PE conjugated tetramer for 20 minutes, then washed and stained with the following extracellular antibodies CD3 AlexaFlour700 (BD) or CD3 Pacific Orange (Invitrogen), CD8 Qdot605 (invitrogen) or CD8 AlexaFlour750 (eBioscience) and Live/Dead marker Violet (Invitrogen) for another 20 minutes. Cells were washed, fixed and samples acquired within 24 hrs on a BD LSR II flowcytometer. Cells were gated on singlets, lymphocytes, live cells, CD3 and then evaluated for CD8+ T cells binding the peptide/MHC tetramer. Data were analyzed using FlowJo version 8.8.6.

5.3 Results

5.3.1 Benchmarking *HLArestrictor* on HIV data

The performance of the *HLArestrictor* at different %rank thresholds was benchmarked on a HIV dataset of 5,145 18mer peptide IFNγ ELIspot responses representing a total of 294 different 18mer peptides being recognized in one or more of 694 treatment naïve HIV infected individuals. To calculate the fraction of validated responses, which could be predicted at a given %rank threshold, predictions were carried out for each validated patient/peptide response for all six patient HLA molecules typed (see Figure 5.5). A response was considered correctly predicted if at least one of the patient's HLA molecules was predicted to bind an 8, 9, 10 or 11mer (a "submer") within the 18mer peptide with a binding strength stronger than the given threshold. At a 2 %rank threshold for instance, the method suggested at least one such epitope for 91% of the 5,145 responses observed, with an average of 5.3 predicted epitopes per response. Likewise, at a 1 %rank threshold, the method suggested at least one epitope for 78% of the responses, with an average of 3.5 predicted epitopes per response. Furthermore, the figure shows the fraction of responses with predicted HLA-A, HLA-B and HLA-C restrictions at each threshold. At the 2 %rank threshold, approximately 50% of the responses were predicted to be HLA-B restricted.

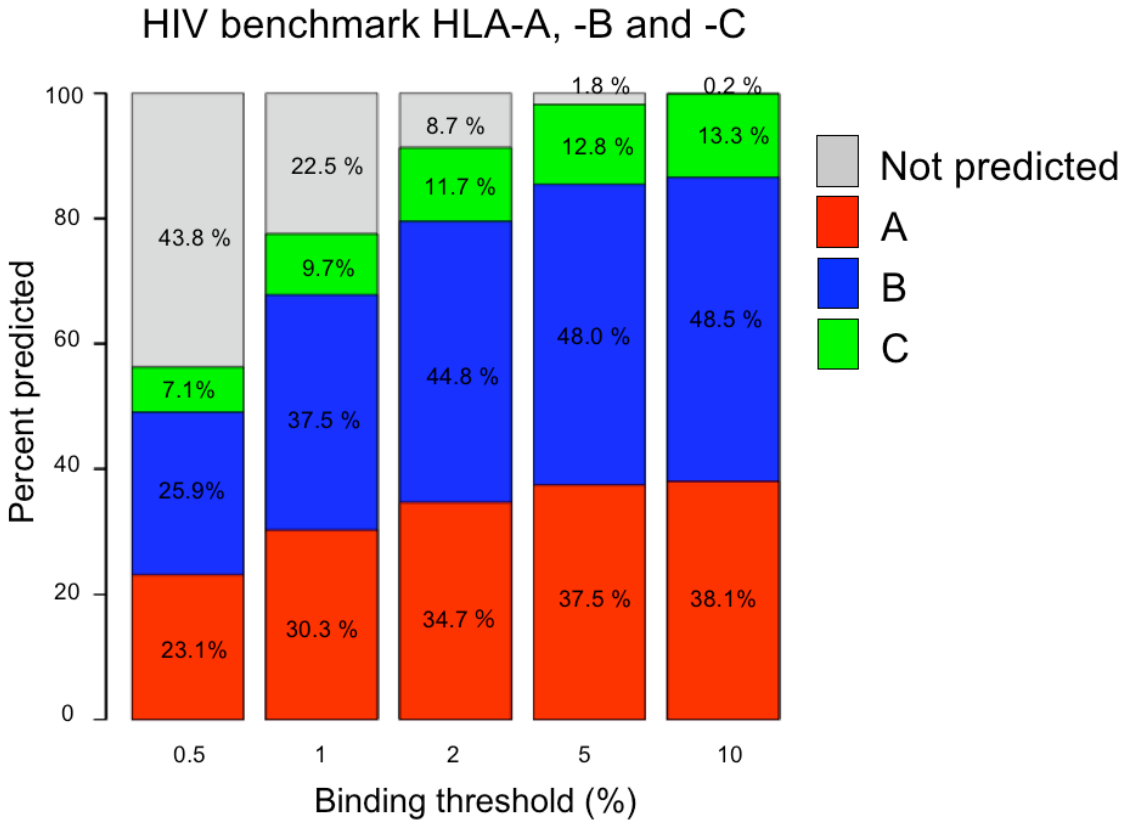


Figure 5.5: Benchmark on HIV peptide IFNγ ELIspot responses. The barplots are divided into colors indicating the number of cases in which a HLA-A, -B or -C type is predicted with the lowest %rank.

5.3.2 HLA restriction identification by association studies

We next investigated what prediction threshold should be applied to give the highest predictive performance of the *HLArestrictor* method. As illustrated in Figure 5.5, choosing a relative high %rank prediction score naturally led to a higher sensitivity of the predictions, however, this came at the price of a higher number of potential false-positive predictions that in real life would have to be analyzed in subsequent immuno-assay validations. To address the question of which prediction threshold would be optimal, we carried out a simple computational experiment. A commonly used method for assigning HLA restriction to immunogenic peptides is HLA association studies. In these studies, the HLA restriction of a given immunogenic peptide is assigned based on prevalence of an HLA allele in a large patient cohort that responds positive to the peptide. A set of 85 (35 HLA-A, and 50 HLA-B) HLA-peptide associations was identified using a Fisher's exact test based analysis that corrects for multiple comparisons. Briefly, a two-by-two contingency table is constructed for each peptide/HLA pair. P-values are then computed using Fisher's exact test for each table and exact q-values (Storey and Tibshirani, 2003) are computed by summing over the null space of all observed contingency tables, as previously described (Carlson et al., 2009). All associations had P-values less than 0.05.

We next applied the *HLArestrictor* method to validate these associations. We identified the patients in the HIV cohort data set expressing the HLA allele in question and that had responded positively to the given peptide. Next, a predicted binding value was assigned for the peptide to each of the patient's HLA alleles as the strongest %rank score for all 8-11submers contained with in the peptide. All peptide/HLA pairs matching the restriction element identified from the association studies were taken as positive, and all other suggested restriction elements as negative. This led to a set consisting of 1067 positive and 5115 negative data points (on average each peptide was tested in 12.5 patients expressing the given HLA allele).

The predictive performance of *HLArestrictor* was finally evaluated in terms of the Matthews correlation coefficient (MCC), sensitivity, and specificity for different values of the %rank threshold. The results of this analysis are shown in Figure 5.6 and demonstrate that the HLA restriction method achieved its highest predictive performance in terms of MCC for %rank threshold values in the range 0.5-2. If reducing the screening load is essential even at the expense of missing some of the epitopes (i.e. high specificity is a requirement and the concurrent loss in sensitivity is acceptable), then a threshold of 0.5 %rank is recommended. Here, the specificity is about 90% and the sensitivity is about 50%. If, on the other hand, identifying as many epitopes as possible is essential even at the expense of an increased screening load (i.e. high sensitivity is a requirement and the concurrent loss in specificity is acceptable), then a threshold of 2 %rank is suggested. Here the sensitivity is close to 90% and the specificity is about 50%.

The Matthews coefficient showed a highly significant ($P < 0.005$, exact permutation test) correlation between the physiological analysis of T cell responses and the prediction of the biochemical analysis of peptide/HLA interaction. In fact, we found that 73 of the 85 peptides (86%), contained an 8-11 submer peptide predicted to bind to the restriction element proposed by the association studies with a binding strength stronger than or equal to 2 %rank. In these cases, the two methods thus agree on the assignment of the HLA restriction element.

However, for 12 of the peptides, *HLArestrictor* failed to confirm the proposed HLA restriction using the suggested 2 %rank threshold. For these 12 peptides, we identified the patients with the proposed HLA allele that had responded positively to the peptide, and analyzed whether any of the other HLA alleles of these patients would predict alternative HLA restriction elements. This analysis allowed us to suggest alternative HLA restrictions for the majority of the positive patient/peptide pairs using the default threshold value of 2 %rank (see Table

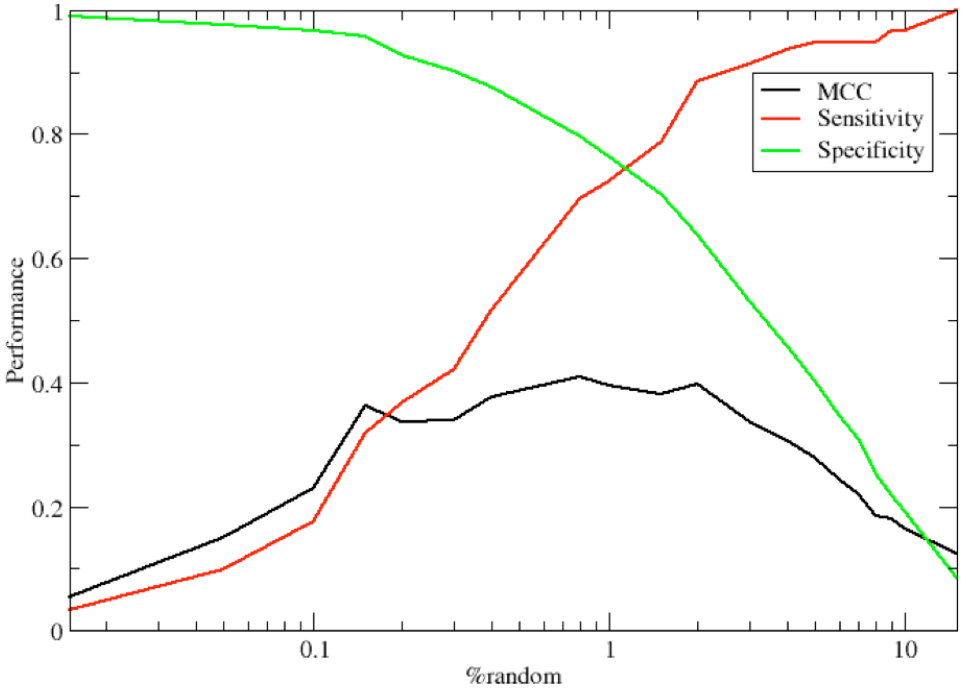


Figure 5.6: The predictive performance of *HLArestrictor* evaluated in terms of the Matthews correlation coefficient (MCC), sensitivity, and specificity for different values of the %rank threshold.

5.2). For instance, we find that all responses to the peptides GKKHYMLKHLVWASREL and EVGFPVRPQVPLRPMTFK by patients having the alleles HLA-A*3601 and HLA-A*0101, respectively, could be explained in terms of alternative HLA restriction elements. Note, HLA-A*3601 is in linkage disequilibrium with HLA-B*5301 (P=9.75×10^{-8}), and 5 of the 7 patients responding to the GKKHYMLKHLVWASREL peptide shared this allele. The peptide YMLKHLVW is predicted to bind HLA-B*5301 with a %rank of 0.8, strongly suggesting that this HLA is a dominant restriction element for this peptide response. Likewise, 21 of the 26 patients responding to the EVGFPVRPQVPLRPMTFK peptide shared both the HLA-A*0101 and HLA-B*8101 alleles. The peptide FPVRPQVPL is predicted to bind the HLA-B*8101 allele with a %rank of 0.01, strongly suggesting that this HLA is a dominant restriction element for this peptide response.

5.3.3 Validation of CD8+ T cell responses using peptide/MHC class I tetramers

To validate the optimal peptide and the corresponding HLA restriction element, we used a panel of 18 peptide/MHC class I tetramers across 8 different HLA alleles in 10 HIV infected individuals as shown in Table 5.3 and Figure 5.7. The optimal epitope within the 18mer was selected based on previously described epitopes combined with information about the binding motif of the restriction element of interest. All 18 epitopes are listed in the Los Alamos HIV molecular immunology database (http://www.hiv.lanl.gov/content/immunology) (Korber et al., 2009).

To confirm the optimal epitope and HLA class I restriction element, the corresponding peptide/MHC class I tetramer was produced and used to stain PBMCs from HIV infected IFNγ ELIspot responders or *in vitro* expanded CTLs as described in Section 5.2. In 16 of

Allele	Peptide	P-value	%rank	Patients	Fraction predicted
A*0101	**EVGFPVRPQVPLRPMTFK**	2.79×10^{-3}	15	26	1.00
A*0101	EWEFVNRPPLVKLWYQL	6.78×10^{-3}	4	12	1.00
B*3501	DEALLQAVRIIKILY	7.41×10^{-9}	3	10	1.00
A*3601	**GKKHYMLKHLVWASREL**	3.92×10^{-2}	15	6	1.00
B*3910	PPIVAKEIVASCDKCQLK	2.29×10^{-3}	15	3	0.67
A*3201	FRLPIQKETWETWWTDYW	2.35×10^{-4}	3	3	0.67
B*5301	AGRWPVKVIHTDNGSNF	3.79×10^{-49}	4	14	0.64
A*6801	FWEVQLGIPHPAGLKKKK	7.13×10^{-26}	5	11	0.64
B*8101	PPIVAKEIVASCDKCQLK	2.16×10^{-32}	9	21	0.62
B*3910	EVNIVTDSQYALGII	4.94×10^{-20}	3	10	0.60
A*0301	LVSIKVGGQIKEALL	1.57×10^{-2}	3	4	0.50
A*1101	IKKKDSTKWRKLVDFREL	2.46×10^{-3}	3	2	0.50

Table 5.2: Alternative HLA restrictions. Allele gives the restriction element predicted from the large cohort population studies. P-value gives the association study P-value for the HLA restriction prediction. %rank gives the predicted binding score in terms of the %rank score for the peptide to the proposed HLA restricting allele. Patients gives the number of patients in our study matching the proposed HLA restriction and responding to the given peptide, and fraction predicted gives the fraction of the responding HLA matched patients having alternative HLA restrictions identified by *HLArestrictor* with a %rank less than or equal to 2%. In bold are highlighted the two examples explained in the text.

18 cases, *HLArestrictor* successfully predicted the HLA restriction element and the minimal epitope with a %rank score below 2. In two cases, *HLArestrictor* failed to predict the validated HLA restriction and minimal epitope with a %rank score of 2. However, in the first case of VKVIEEKAF/HLA-B*1503, the predicted affinity was 155 nM and the predicted %rank was 6.0. In the second case of SLYNTVATL/HLA-A*0201 the predicted affinity was 436 nM and the predicted %rank was 4.0. Thus, both epitopes were predicted to bind stronger than the commonly used threshold of 500 nM (Lundegaard et al., 2007; Moutaftsi et al., 2006; Sette et al., 1994).

Further, the previously described epitope RSLYNTVATLY/HLA-B*58 predicted to bind to the B*5801 molecule with an affinity of 62 nM and a 0.4 %rank illustrating how *HLArestrictor* often will predict multiple restrictions, in this case both correct, within a given positive peptide. Additional known epitopes not tested in our patients were predicted in several other cases as well. For example within the 18mer WVKVIEEKAFSPEVIPMF, the known epitopes EEKAFSPEV/HLA-B*4501 and KAFSPEVI/HLA-B*5703 were predicted to bind at 0.3 %rank and 0.1 %rank, respectively.

Patient	HIV 18'mer with ELIspot response	Validated epitope	Valid. allele	Pred. affinity	Pred. %rank
N080	PRTLNAW**VKVIEEKAF**	VKVIEEKAF	B*1503	155 nM	6.0 %
N080	YHCLVC**FQTKGLGISY**GR	FQTKGLGISY	B*1503	8 nM	**0.8 %**
N080	VKAACWWAG**IQQEFGIPY**	IQQEFGIPY	B*1503	4 nM	**0.1 %**
N080	AVFIHN**FKRKGGIGGY**SA	FKRKGGIGGY	B*1503	24 nM	**1.5 %**
H044	WVKVIEE**KAFSPEVIPMF**	KAFSPEVIPMF	B*5703	NA	**0.4 %**
N021	ELKQEAVRH**FPRPWLHGL**	FPRPWLHGL	B*4201	49 nM	**0.05 %**
N012	ACQGVG**GPSHKARVL**AEA	GPSHKARVL	B*0702	36 nM	**0.3 %**
N012/ R050	CRAIRN**IPRRIRQGL**	IPRRIRQGL	B*0702	10 nM	**0.1 %**
R050	NY**TPGPGVRYPL**TFGWCF	TPGPGVRYPL	B*0702	45 nM	**0.3 %**
R050	QGWKG**SPAIFQSSM**TKIL	SPAIFQSSM	B*0702	10 nM	**0.1 %**
R050/ R039	WVKVIEE**KAFSPEVIPMF**	KAFSPEVIPMF	B*5701	67 nM	**0.1 %**
R039	PVGEIY**KRWIILGLNK**IV	KRWIILGLNK	B*2705	22 nM	**0.05 %**
R039	AVFIHNF**KRKGGIGGY**SA	KRKGGIGGY	B*2705	289 nM	**1.0 %**
R035	ELKNEA**VRHFPRIWL**HSL	VRHFPRIWL	B*2705	357 nM	**1.0 %**
R014	MASEFN**LPPIVAKEI**VA	LPPIVAKEI	B*4201	NA	**1.5 %**
N067	TGSEELR**SLYNTVATL**Y	SLYNTVATL	A*0201	436 nM	4.0 %
N067/ N096	ELAENRE**ILKEPVHGV**YY	ILKEPVHGV	A*0201	42 nM	**1.5 %**
N096	SDIAGT**TSTLQEQIAW**M	TSTLQEQIAW	B*5801	32 nM	**0.2 %**

Table 5.3: Tetramer validations on selected patients as exemplified in Figure 5.7. The validated epitope within each 18mer is underlined. The predicted affinity is listed when available and the predicted %rank score is marked in bold if it is below or equal to a 2% threshold. 16 of 18 peptide/MHC class I tetramer validated CD8+ T cell epitopes are successfully predicted by *HLArestrictor* at this threshold setting, while all 18 are predicted either below or equal to 2 %rank OR below 500 nM .

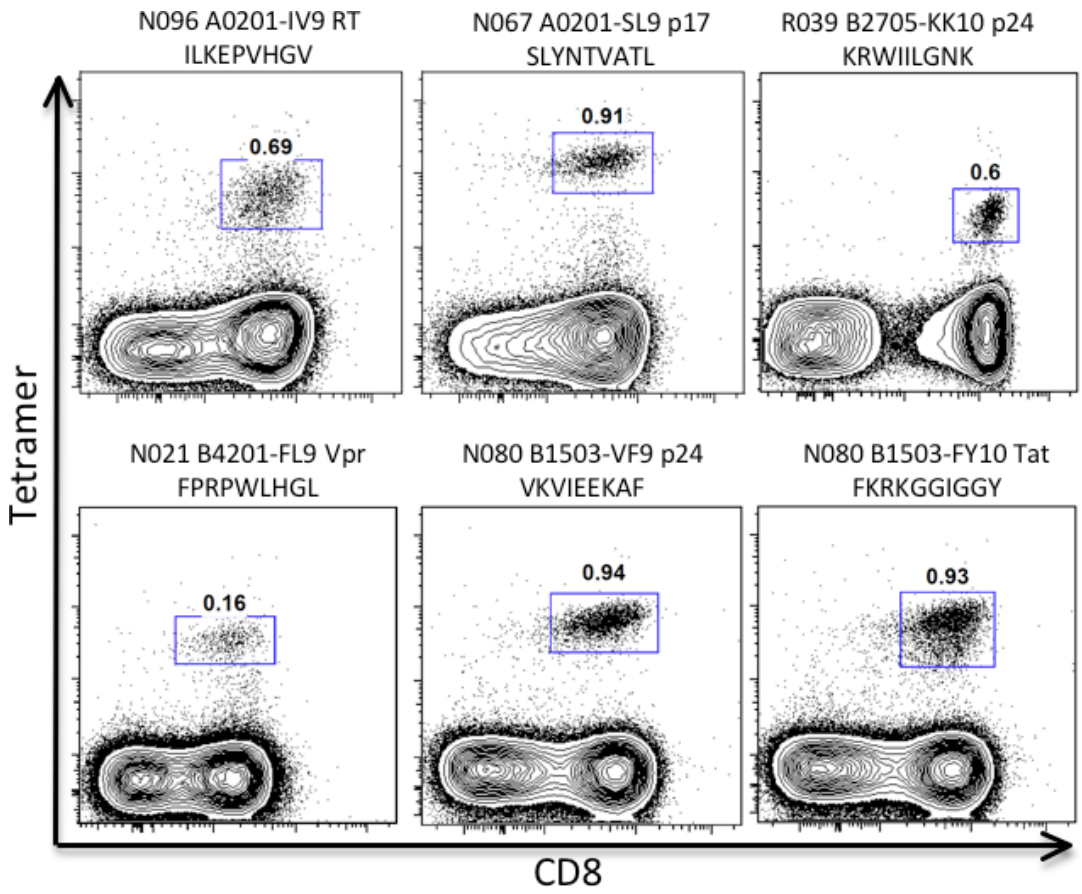


Figure 5.7: Examples of peptide/MHC class I tetramer stainings used to validate optimal epitopes and HLA restriction of CD8+ T cell responses in HIV infected individuals shown in Table 5.3. Cells are gated on singlets, lymphocytes, live CD3+ T cells and CD8+ T cells plotted against tetramer positive cells with the number indicating the percentage of tetramer positive in the CD3 gate. The patient ID, tetramer HLA, name of the peptide and the peptide sequence is shown above each plot.

5.4 Discussion

T cell epitopes consist of antigen-derived peptides presented in the context of HLA molecules; and the identification and validation of peptide/HLA complexes, which can stimulate T cell responses, is at the heart of any T cell epitope discovery process. Finding the stimulatory peptide and the presenting HLA restriction element is not a simple task. Here, we present an immunoinformatics method, *HLArestrictor*, which has been tailored to support T cell epitope discovery in individual subjects. As inputs, it needs the amino acid sequence of the target protein(s), and the HLA type of the individual in question (high-resolution HLA typing e.g. HLA-A*0101, and preferably for all relevant loci e.g. for HLA-A, -B, -C for HLA class I restricted CTL responses). Using these inputs, *HLArestrictor* creates all possible 8, 9, 10 and 11mer peptides from the target protein(s), predicts their binding to all the HLA molecules in question, and generates an output file consisting of the most likely peptide/HLA combination(s). Peptide/HLA tetramers is one of the most efficient means to validate T cell epitopes, and *HLArestrictor* can also be viewed as a tool for efficient design of specific peptide/HLA tetramers.

We have successfully applied *HLArestrictor* to the search for patient-specific HLA restriction elements and optimal epitopes. To this end, we have re-analyzed a major study of T cell epitopes within the 2004 C-clade consensus HIV sequence (Kiepiela et al., 2007). In this study, the consensus sequence was represented by 420 overlapping 18mer peptides and tested in 694 treatment naïve HIV infected individuals, which had been high-resolution typed for HLA-A, -B, and -C. A large set of HLA restrictions were identified from these data by association studies. Initially, we asked which %rank threshold should be used to generate T cell epitope predictions. At the 2 %rank and 1 %rank thresholds, the *HLArestrictor* method suggested HLA restriction elements and optimal peptides for 91% and 78%, respectively, of the 5145 identified HIV-specific IFNγ ELIspot peptide responses.

Next, we asked to what degree known T cell epitopes could be successfully predicted. A large set of HLA restrictions had been identified by association studies and could conveniently be used to further validate the predictive performance of the *HLArestrictor* method. Using this benchmark data set, the *HLArestrictor* was shown to achieve its optimal predictive performance for %rank score thresholds in the range from 0.5 to 2.0. At the 0.5 %rank threshold, *HLArestrictor* would correctly identify 50% of the known HLA restriction elements, while rejecting 90% of the non-restricting HLA alleles; whereas at the 2% threshold it would identify 90% of the known HLA restriction elements, while rejecting 50% of the non-restricting HLA alleles.

Note, that this type of analysis is very crude and simple, and that in particular the estimated specificity value of *HLArestrictor* should be interpreted with great caution. By way of example, we have only included the strongest association as the true positive HLA restriction element, and assigned all other possible HLA restriction element as being negative, This is not always correct as some of the less strongly associated HLA restriction elements may well be bona fide HLA restriction elements. Indeed, as pointed out above, some of the HLA restriction elements that are rejected in this way are well-known and experimentally characterized HLA restrictions. When *HLArestrictor* correctly identifies them as possible HLA restriction elements they inadvertently end up being considered false positives. None-the-less, the calculation is simple and un-biased and clearly demonstrates that the HLA restriction method achieves its highest predictive performance for %rank threshold values in the range 0.5-2. The benchmark demonstrated that here was a strong agreement between the HLA restriction identified by association study and the *HLArestrictor* predictions. However, in 12 of 85 cases, *HLArestrictor* failed to reproduce the restriction element suggested by the association study analysis. In these cases

HLArestrictor suggested alternative HLA restriction elements and minimal peptides for the majority of patients responding to the peptide in question. These findings strongly suggest that *HLArestrictor* is capable of providing information beyond what is obtainable using associations studies, and that the method is highly sensitive and specific when predicting potential peptide/HLA restrictions.

Further supporting the strong predictive power of the method and demonstrating that it goes beyond identification of the most likely HLA restriction element and also identifies the minimal peptide, we used a panel of 18 peptide MHC class I tetramers across 8 different HLA alleles in 10 HIV infected individuals to validate both the optimal epitope CD8+ T cell response and the corresponding HLA restriction. In 16 out of 18 cases, *HLArestrictor* successfully predicted the HLA restriction and minimal epitope with a %rank score below or equal to 2. If the settings of the *HLArestrictor* were changed so that they also included any predicted binding affinity below 500 nM then this figure changed to identifying 18 of 18 tetramer validated epitopes. These observations illustrate the strength of *HLArestrictor*, as it does not only predict a patient's IFNγ ELIspot response to an N-mer, but also the correct HLA restriction and optimal epitope. The method thus provides a valuable guidance for researchers designing tetramers to validate HLA restriction elements and minimal epitopes corresponding to a given cellular response.

During the development of the *HLArestrictor*, we preferred the %rank measure rather than the affinity measure since the %rank measure lends itself to the needed comparisons across different HLA molecules and HLA isotypes. Predicting the affinity measure is a more demanding task and not all HLA alleles are yet represented at a level that allows such quantitative predictions. It has further been suggested that not all MHC molecules present peptides at the same binding threshold (Rao et al., 2009; Stranzl et al., 2010). The %rank score removes this bias by placing binding scores for all MHC molecules on an equal scale. However, it has also been suggested that immunogenic peptides are characterized by an HLA-binding affinity threshold of 500 nM (Assarsson et al., 2007; Sette et al., 1994). As more peptide/HLA binding data becomes available, it will eventually become possible to use affinity measurements for more and more HLA molecules when interpreting the results.

It is even possible to estimate how reliable the prediction of affinity is for each HLA molecule (note that the output of the *HLArestrictor* includes this estimate of reliability). Whenever reliable, it is possible to include an affinity threshold in the interpretation of the output. For ease of operation, *HLArestrictor* includes an Rank OR Affinity setting (the default) that allows the selection of peptide/HLA combinations that meets either the %rank or the affinity thresholds. This allows that a predicted HLA restriction with e.g. a borderline %rank score might still be classified as an epitope due to a strong affinity. Indeed, this was the case for the two HLA tetramer validated epitopes, which the *HLArestrictor* failed to recognize when running in the %rank only setting. Both these peptides were predicted to bind to one of the patient's HLA molecules stronger than the commonly used affinity threshold of 500 nM even though they failed to be predicted below the 2 %rank threshold.

If running the *HLArestrictor* in the %rank $\leq$ 2% OR affinity $\leq$ 500 nM threshold setting for the definition of positive HLA restriction predictions and applying this setting to the 5145 IFNγ ELIspot peptide responses, as much as 94.0% of the responses were predicted with an average of 6.4 predicted epitope/HLA restrictions per peptide compared to 91.3% and 5.3 predicted epitope/HLA restrictions per peptide at a %rank $\leq$ 2%. This increase in sensitivity at a relative minor loss in specificity suggests that interpretations of minimal epitopes and HLA restrictions from the *HLArestrictor* predictions should be based on a combined evaluation of both the %rank and affinity prediction values.

The *HLArestrictor* is specifically aimed at T cell epitope discovery in individual subjects. Technically, it is possible to enter a collection of proteins up to an entire proteome although the response from the server might be rather slow due to the large number of calculations and subsequent output sorting. It was, however, developed with the analysis of shorter peptide sequences in mind. The use of overlapping peptides has emerged as a very powerful way to scan entire proteins for the presence of immunogenic epitopes. Conventionally, identifying HLA restriction elements and minimal epitopes within this approach are done by presenting peptides on partially HLA-matched B cells and using more or less systematic peptide truncations, respectively. This requires considerable resources (high-resolution typing, extensive peptide synthesis and extensive cellular testing).

HLArestrictor automates the bioinformatics analysis and avoids any bias inherent to a manual 'eye-balling' analysis, and should relieve the experimenter of much tedious and costly work. Validating all potential 8-11mer peptides to all the patient HLA alleles is clearly a highly costly and inefficient brute force approach. An 18mer peptide will for instance contain up to 38 distinct 8-11mer peptides leading to a total of 228 peptide/HLA pairs when tested against six HLA types of a patient. Furthermore, *HLArestrictor* was developed with the rapid identification and design of peptide/HLA tetramers in mind. These have emerged as the most direct and efficient method to detect peptide-specific, HLA restricted T cells. High-throughput methods are now available for peptide/HLA tetramer generation (Leisner et al., 2008; Toebes et al., 2006). It is therefore entirely feasible to use *HLArestrictor* as a rational guide to rapid and large-scale peptide/HLA tetramer formation for direct T cell epitope discovery.

5.5 Conclusion

In conclusion, *HLArestrictor* (http://www.cbs.dtu.dk/services/HLArestrictor) is a user-friendly tool for patient-specific epitope discovery within peptides. The user can adjust a number of parameters, predictions can be made for 8-11mers, and the different sort-modes provide the user with a flexible overview of the predictions. The large-scale benchmarking on experimental data of the method makes it one of the best validated prediction tools of its kind to date and proves how the method can be valuable tool to guide the rational identification of new epitopes from infectious diseases and other disease models. The method will be updated continuously as data becomes available for improving the underlying peptide/HLA predictors (e.g. the representation of binding data for HLA-C molecules is expected to improve significantly in the near future thereby improving HLA-C predictors in particular from an affinity measurement perspective). Another area of future development will be to include HLA class II predictions.

Chapter 6

Summary & perspectives

The work presented in this thesis concerns the molecular mechanisms involved in allo-HCT. Allo-HCT is, as other kinds of transplantations, a highly artificial situation. I find the following poetic description by Stevanovic (2005) quite incisive:

> When nature invented the adaptive immune system millions of years ago, her primary aim was resistance against small pathogens such as viruses or bacteria. At that time, transfusion or transplantation was unheard of.

Indeed, the evolution of the adaptive immune system was optimized for completely different purposes than the ability of being injected into another person, finding its own way into the bone marrow, replacing a malfunctioning hematopoietic system, and killing off the cancerous cells. Seen in this light, it is truly amazing that allo-HCT is such a successful means of curing diverse hematological diseases. That being said, it is still a risky treatment, with potentially lethal side effect. Currently, some patients are cured completely with few side effects, while others suffer from disease relapse, TRM, or detrimental GVHD. For instance, the first 100 danish patients receiving a NMC allo-HCT had a 49% prevalence of cGVHD, 25% RRM and 17% TRM (Kornblit et al., 2008). Several risk factors leading to GVHD have been identified, such as older patient age, unrelated donor, sex-mismatched donor, or HCT source (Filipovich et al., 2005). Even though general trends are observed in large cohorts, the outcome in a given patient, cannot be precisely predicted. Being able to better predict the outcome of a transplantation is thus one of the largest challenges faced by this field.

The focus of this thesis is the role that mHags play in transplantations. Chapter 2 concerns the special subset of mHags which are caused by the presence of proteins encoded by the Y chromosome in sex-mismatched allo-HCT. In a transplantation setting, where a male patient receive graft from a HLA-identical female donor, an additional mHag burden is present due to the fact that the female T cells have not been presented to Y chromosomal peptide fragments during their thymal training. As part of a larger research collaboration aiming at the identification of novel mHags, I here used *NetMHCpan* to predict candidate mHags encoded by the Y chromosome. As the number of predicted peptide binders was huge, it was necessary to select a more promising subset of peptides for subsequent validation experiments in our patient cohort. At the moment, the immunogenecity of these peptides is being tested with T cells from the patients, using ICS assays and tetramers.

Our approach at mHag discovery is untraditional, as we apply the concept of reverse immunology in a small scale. We begin with predictions and expect to end up with isolated T cells recognizing some of the predicted mHags. Usually, mHag identification works the other

way around, starting with an isolated T cell clone and, by various experimental and prediction methods, ending up with the mHag recognized by the T cells. Our preliminary results look promising and we hope that the results of our validation experiments will lead to the identification of novel mHags, thus proving our mHag identification method viable.

In Chapter 3, we demonstrated the correlation between the number of predicted mHags around selected nsSNPs and poorer clinical outcome after allo-HCT. This study was part of the pilot project, described in Chapter 4, the goal of which was to identify novel mHags caused by nsSNP-disparities between donor and patient. In Chapter 3 we saw that patients with many predicted mHags had a poorer OS and increased TRM than those with few predicted mHags. Interestingly, this effect was significant, even though we only considered nsSNPs from the relatively few proteins in which mHags have previously been identified. As follow-up to this small-scale study, the patients have now been genotyped with a SNP-microarray, and it shall prove interesting to see the results reproduced in a much larger scale. As the chip covers nsSNPs throughout most of the human genome, it will be necessary to restrict the analysis to proteins with a sufficiently high expression level, expressed in tissues relevant to GVHD or GVT. Should the result be reproducible in this large scale study, one could imagine that, in the future, a full analysis of the genomic differences between a patient/donor pair, followed by the *in silico* prediction of mHags, could be incorporated into the standard donor selection procedures.

Presently, the strength of microarray based genotyping lies within the search for novel mHags. In the pilot project described in Chapter 4, we use predictions to guide the search for mHags in proteins already known to encode mHags, as well as in selected proteins with a hematopoietically restricted expression. This study, too, could be expanded to encompass all relevant proteins, and using binding predictions, to select a set of mHags candidates to validate experimentally in our patient cohort. Narrowing down the number of mHag candidates to test experimentally would then be a huge bioinformatical challenge.

Alternatively, a more data-driven approach, similar to the work by Ogawa et al. (2008) described in Section 4.1, could be employed. Ogawa et al. used a GWAS to correlate GVHD with the presence of certain SNP-disparities coupled with the HLA alleles of patient/donor pairs without considering peptide binding predictions. By adding mHag predictions, it would be possible to search for predicted mHags with a statistically significant overrepresentation in patients with a successful vs. poor treatment outcome. The mHag candidates identified using such a data-driven approach could make up a more promising test set to validate experimentally than the ones identified by expanding the *ab initio* approach used in Chapter 4 to a genome-wide level.

One of the most promising perspectives of searching for novel mHags using microarray-based genotyping, is the possibility to search systematically for the therapeutically relevant mHags. To date, only around 13 immunotherapeutic mHags have been identified, several of which are only relevant to a limited number of hematological malignancies (Spaapen and Mutis, 2008). Due to the potentially curative effects of these mHags, there is now an increased focus on the identification of novel immunotherapeutic mHags, preferably represented by a significant fraction of the population and relevant to a broad range of malignancies.

The last part of this thesis, presented in Chapter 5 is more general. It concerns the development and benchmark of an online tool, *HLArestrictor*, based on *NetMHCpan* and designed to aid researchers in getting a quick, patient-specific overview of predicted binders from a peptide or protein. During our work on mHag prediction, especially while predicting mHags from the Y chromosome, it became clear to us that *NetMHCpan* would benefit from a more patient-oriented user interface. *HLArestrictor* solves this problem, by allowing predictions of

8-11mers simultaneously and by presenting the prediction results in a more flexible way. Although we developed *HLArestrictor* while working with mHag predictions it is just as useful in general, patient-specific T cell epitope prediction. We benchmarked *HLArestrictor* to investigate how well it solved the task of predicting HLA restriction elements and minimal epitopes within a large dataset of ELIspot responses. The benchmark results, presented in Chapter 5, clearly demonstrate the usefulness of *HLArestrictor* in such applications.

Bibliography

Y. Akatsuka, T. Nishida, E. Kondo, M. Miyazaki, H. Taji, H. Iida, K. Tsujimura, M. Yazaki, T. Naoe, Y. Morishima, Y. Kodera, K. Kuzushima, and T. Takahashi. Identification of a polymorphic gene, bcl2a1, encoding two novel hematopoietic lineage-specific minor histocompatibility antigens. *J Exp Med*, 197(11):1489--500, 2003a.

Y. Akatsuka, E. H. Warren, T. A. Gooley, A. G. Brickner, M. T. Lin, J. A. Hansen, P. J. Martin, D. K. Madtes, V. H. Engelhard, T. Takahashi, and S. R. Riddell. Disparity for a newly identified minor histocompatibility antigen, ha-8, correlates with acute graft-versus-host disease after haematopoietic stem cell transplantation from an hla-identical sibling. *Br J Haematol*, 123(4):671--5, 2003b.

Y. Akatsuka, Y. Morishima, K. Kuzushima, Y. Kodera, and T. Takahashi. Minor histocompatibility antigens as targets for immunotherapy using allogeneic immune reactions. *Cancer Sci*, 98(8):1139--46, 2007.

M. al Jurf, F. Aranha, C. Annassetti, JF. Apperley, R. Baynes, WI. Bensinger, D. Blaise, MA. Chaudhary, M. Clarke, JJ. Cornelissen, S. Couban, C. Cutler, B. Djulbegovic, M. Gyger, A. Gratwohl, D. Heldal, B. Van der Holt, I. Hozo, M. Kuentz, A. Kumar, J. Lipton, J. Matchamm, M. Mohty, J. Morton, T. Panzarella, R. Powles, SM. Richards, E. Sahovic, N. Schmitz, DR. Simpson, B. Sirohi, HP. Soares, CA. de Souza, AC. Vigorito, and Wheatley K. Allogeneic peripheral blood stem-cell compared with bone marrow transplantation in the management of hematologic malignancies: an individual patient data meta-analysis of nine randomized trials. *J Clin Oncol*, 23(22):5074--87, 2005.

D. G. Altman. *Practical statistics for medical research.* Chapman and Hall, first edition.

J. D. Altman, P. A. Moss, P. J. Goulder, D. H. Barouch, M. G. McHeyzer-Williams, J. I. Bell, A. J. McMichael, and M. M. Davis. Phenotypic analysis of antigen-specific t lymphocytes. *Science*, 274(5284):94--6, 1996.

S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389--402, 1997.

M. Aoudjhane, M. Labopin, N. C. Gorin, A. Shimoni, T. Ruutu, H. J. Kolb, F. Frassoni, J. M. Boiron, J. L. Yin, J. Finke, H. Shouten, D. Blaise, M. Falda, A. A. Fauser, J. Esteve, E. Polge, S. Slavin, D. Niederwieser, A. Nagler, and V. Rocha. Comparative outcome of reduced intensity and myeloablative conditioning regimen in hla identical sibling allogeneic haematopoietic stem cell transplantation for patients older than 50 years of age with

acute myeloblastic leukaemia: a retrospective survey from the acute leukemia working party (alwp) of the european group for blood and marrow transplantation (ebmt). *Leukemia*, 19 (12):2304--12, 2005.

E. Assarsson, J. Sidney, C. Oseroff, V. Pasquetto, H. H. Bui, N. Frahm, C. Brander, B. Peters, H. Grey, and A. Sette. A quantitative analysis of the variables affecting the repertoire of t cell specificities recognized after vaccinia virus infection. *J Immunol*, 178(12):7890--901, 2007.

P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412--24, 2000.

M. R. Barnes. *Bioinformatics for geneticists*. Wiley, second edition.

R. Barth, S. Counce, P. Smith, and G. D. Snell. Strong and weak histocompatibility gene differences in mice and their role in the rejection of homografts of tumors and skin. *Ann Surg*, 144(2):198--204, 1956. Journal Article Not Available.

M. Bhasin and G. P. Raghava. Prediction of promiscuous and high-affinity mutated mhc binders. *Hybrid Hybridomics*, 22(4):229--34, 2003.

M. Bhasin and G. P. Raghava. A hybrid approach for predicting promiscuous mhc class i restricted t cell epitopes. *J Biosci*, 32(1):31--42, 2007.

M. Bleakley and S. R. Riddell. Molecules and mechanisms of the graft-versus-leukaemia effect. *Nat Rev Cancer*, 4(5):371--80, 2004.

A. G. Brickner, E. H. Warren, J. A. Caldwell, Y. Akatsuka, T. N. Golovina, A. L. Zarling, J. Shabanowitz, L. C. Eisenlohr, D. F. Hunt, V. H. Engelhard, and S. R. Riddell. The immunogenicity of a new human minor histocompatibility antigen results from differential antigen processing. *J Exp Med*, 193(2):195--206, 2001.

A. G. Brickner, A. M. Evans, J. K. Mito, S. M. Xuereb, X. Feng, T. Nishida, L. Fairfull, R. E. Ferrell, K. A. Foon, D. F. Hunt, J. Shabanowitz, V. H. Engelhard, S. R. Riddell, and E. H. Warren. The pane1 gene encodes a novel human minor histocompatibility antigen that is selectively expressed in b-lymphoid cells and b-cll. *Blood*, 107(9):3779--86, 2006.

J.M. Carlson, D. Heckerman, and G. Shani. Estimating false discovery rates for contingency tables. *Microsoft Corporation TechReport*, 2009. http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/FalseDiscoveryRate/.

G. Cavanagh, C. E. Chapman, V. Carter, A. M. Dickinson, and P. G. Middleton. Donor cd31 genotype impacts on transplant complications after human leukocyte antigen-matched sibling allogeneic bone marrow transplantation. *Transplantation*, 79(5):602--5, 2005.

J. E. Christensen, J. P. Christensen, N. N. Kristensen, N. J. Hansen, A. Stryhn, and A. R. Thomsen. Role of cd28 co-stimulation in generation and maintenance of virus-specific t cells. *Int Immunol*, 14(7):701--11, 2002.

Jr. Collins, R. H., O. Shpilberg, W. R. Drobyski, D. L. Porter, S. Giralt, R. Champlin, S. A. Goodman, S. N. Wolff, W. Hu, C. Verfaillie, A. List, W. Dalton, N. Ognoskie, A. Chetrit, J. H. Antin, and J. Nemunaitis. Donor leukocyte infusions in 140 patients with relapsed malignancy after allogeneic bone marrow transplantation. *J Clin Oncol*, 15(2):433--44, 1997.

Consortium. The international hapmap project. *Nature*, 426(6968):789--96, 2003.

E. A. Copelan. Hematopoietic stem-cell transplantation. *N Engl J Med*, 354(17):1813--26, 2006.

C. Cutler, S. Giri, S. Jeyapalan, D. Paniagua, A. Viswanathan, and J. H. Antin. Acute and chronic graft-versus-host disease after allogeneic peripheral-blood stem-cell and bone marrow transplantation: a meta-analysis. *J Clin Oncol*, 19(16):3685--91, 2001.

C. C. Czerkinsky, L. A. Nilsson, H. Nygren, O. Ouchterlony, and A. Tarkowski. A solid-phase enzyme-linked immunospot (elispot) assay for enumeration of specific antibody-secreting cells. *J Immunol Methods*, 65(1-2):109--21, 1983.

B. de Rijke, A. van Horssen-Zoetbrood, J. M. Beekman, B. Otterud, F. Maas, R. Woestenenk, M. Kester, M. Leppert, A. V. Schattenberg, T. de Witte, E. van de Wiel-van Kemenade, and H. Dolstra. A frameshift polymorphism in p2x5 elicits an allogeneic cytotoxic t lymphocyte response associated with remission of chronic myeloid leukemia. *J Clin Invest*, 115(12): 3506--16, 2005.

D. S. Deluca, B. Eiz-Vesper, N. Ladas, B. A. Khattab, and R. Blasczyk. High throughput minor histocompatibility antigen prediction. *Bioinformatics*, 25:2411--17, 2009.

J. M. den Haan, N. E. Sherman, E. Blokland, E. Huczko, F. Koning, J. W. Drijfhout, J. Skipper, J. Shabanowitz, D. F. Hunt, V. H. Engelhard, and et al. Identification of a graft versus host disease-associated human minor histocompatibility antigen. *Science*, 268(5216):1476--80, 1995.

J. M. den Haan, L. M. Meadows, W. Wang, J. Pool, E. Blokland, T. L. Bishop, C. Reinhardus, J. Shabanowitz, R. Offringa, D. F. Hunt, V. H. Engelhard, and E. Goulmy. The minor histocompatibility antigen ha-1: a diallelic gene with a single amino acid polymorphism. *Science*, 279(5353):1054--7, 1998.

H. Dolstra, H. Fredrix, F. Preijers, E. Goulmy, C. G. Figdor, T. M. de Witte, and E. van de Wiel-van Kemenade. Recognition of a b cell leukemia-associated minor histocompatibility antigen by ctl. *J Immunol*, 158(2):560--5, 1997.

H. Dolstra, H. Fredrix, F. Maas, P. G. Coulie, F. Brasseur, E. Mensink, G. J. Adema, T. M. de Witte, C. G. Figdor, and E. van de Wiel-van Kemenade. A human minor histocompatibility antigen specific for b cell acute lymphoblastic leukemia. *J Exp Med*, 189(2):301--8, 1999.

M. E. Dudley, J. R. Wunderlich, P. F. Robbins, J. C. Yang, P. Hwu, D. J. Schwartzentruber, S. L. Topalian, R. Sherry, N. P. Restifo, A. M. Hubicki, M. R. Robinson, M. Raffeld, P. Duray, C. A. Seipp, L. Rogers-Freezer, K. E. Morton, S. A. Mavroukakis, D. E. White, and S. A. Rosenberg. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. *Science*, 298(5594):850--4, 2002.

J. L. Ferrara, J. E. Levine, P. Reddy, and E. Holler. Graft-versus-host disease. *Lancet*, 373 (9674):1550--61, 2009.

A. H. Filipovich, D. Weisdorf, S. Pavletic, G. Socie, J. R. Wingard, S. J. Lee, P. Martin, J. Chien, D. Przepiorka, D. Couriel, E. W. Cowen, P. Dinndorf, A. Farrell, R. Hartzman, J. Henslee-Downey, D. Jacobsohn, G. McDonald, B. Mittleman, J. D. Rizzo, M. Robinson, M. Schubert, K. Schultz, H. Shulman, M. Turner, G. Vogelsang, and M. E. Flowers. National institutes of health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. diagnosis and staging working group report. *Biol Blood Marrow Transplant*, 11(12):945--56, 2005.

P. Fontaine, G. Roy-Proulx, L. Knafo, C. Baron, D. C. Roy, and C. Perreault. Adoptive transfer of minor histocompatibility antigen-specific t lymphocytes eradicates leukemia cells without causing graft-versus-host disease. *Nat Med*, 7(7):789--94, 2001.

G. Gahrton. Risk assessment in haematopoietic stem cell transplantation: impact of donor-recipient sex combination in allogeneic transplantation. *Best Pract Res Clin Haematol*, 20 (2):219--29, 2007.

G. Gahrton, S. Iacobelli, J. Apperley, G. Bandini, B. Bjorkstrand, J. Blade, J. M. Boiron, M. Cavo, J. Cornelissen, P. Corradini, N. Kroger, P. Ljungman, M. Michallet, N. H. Russell, D. Samson, A. Schattenberg, B. Sirohi, L. F. Verdonck, L. Volin, A. Zander, and D. Niederwieser. The impact of donor gender on outcome of allogeneic hematopoietic stem cell transplantation for multiple myeloma: reduced relapse risk in female to male transplants. *Bone Marrow Transplant*, 35(6):609--17, 2005.

T. A. Gooley, W. Leisenring, J. Crowley, and B. E. Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med*, 18(6): 695--706, 1999.

E. Goulmy, A. Termijtelen, B. A. Bradley, and J. J. van Rood. Y-antigen killing by t cells of women is restricted by hla. *Nature*, 266(5602):544--5, 1977.

E. Goulmy, R. Schipper, J. Pool, E. Blokland, J. H. Falkenburg, J. Vossen, A. Gratwohl, G. B. Vogelsang, H. C. van Houwelingen, and J. J. van Rood. Mismatches of minor histocompatibility antigens between hla-identical donors and recipients and the development of graft-versus-host disease after bone marrow transplantation. *N Engl J Med*, 334(5):281--5, 1996.

R. J. Gray. A class of k-sample tests for comparing the cummulative incidence of a competing risk. *Annals of Statistics*, 16(3):1141--1154, 1988.

F. C. Grumet, D. D. Hiraki, B. W. M. Brown, J. L. Zehnder, E. S. Zacks, A. Draksharapu, J. Parnes, and R. S. Negrin. Cd31 mismatching affects marrow transplantation outcome. *Biol Blood Marrow Transplant*, 7(9):503--12, 2001.

P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower. Mhcpred: bringing a quantitative dimension to the online prediction of mhc binding. *Appl Bioinformatics*, 2(1):63--6, 2003.

M. Halling-Brown, R. Quartey-Papafio, P. J. Travers, and D. S. Moss. Sipep: a system for the prediction of tissue-specific minor histocompatibility antigens. *Int J Immunogenet*, 33(4): 289--95, 2006.

L. Hambach, E. Spierings, and E. Goulmy. Risk assessment in haematopoietic stem cell transplantation: minor histocompatibility antigens. *Best Pract Res Clin Haematol*, 20(2):171--87, 2007.

J. Hammer, E. Bono, F. Gallazzi, C. Belunis, Z. Nagy, and F. Sinigaglia. Precise prediction of major histocompatibility complex class ii-peptide interaction based on peptide side chain scanning. *J Exp Med*, 180(6):2353--8, 1994.

P. Hari, J. Carreras, M. J. Zhang, R. P. Gale, B. J. Bolwell, C. N. Bredeson, L. J. Burns, M. S. Cairo, C. O. Freytes, S. C. Goldstein, G. A. Hale, D. J. Inwards, C. F. Lemaistre, D. Maharaj, D. I. Marks, H. C. Schouten, S. Slavin, J. M. Vose, H. M. Lazarus, and K. van Besien. Allogeneic transplants in follicular lymphoma: higher risk of disease progression after reduced-intensity compared to myeloablative conditioning. *Biol Blood Marrow Transplant*, 14(2):236--45, 2008.

M. Harndahl, S. Justesen, K. Lamberth, G. Roder, M. Nielsen, and S. Buus. Peptide binding to hla class i molecules: homogenous, high-throughput screening, and affinity assays. *J Biomol Screen*, 14(2):173--80, 2009.

L. H. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, L. M. Silver, and R. C. Veres. *Genetics - From genes to genomes*. The McGraw-Hill Companies., first edition.

I. Hoof, B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, S. Buus, and M. Nielsen. Netmhcpan, a method for mhc class i binding prediction beyond humans. *Immunogenetics*, 61(1):1--13, 2009.

R. Ivanov, T. Aarts, S. Hol, A. Doornenbal, A. Hagenbeek, E. Petersen, and S. Ebeling. Identification of a 40s ribosomal protein s4-derived h-y epitope able to elicit a lymphoblast-specific cytotoxic t lymphocyte response. *Clin Cancer Res*, 11(5):1694--703, 2005.

A. S. Juncker, M. V. Larsen, N. Weinhold, M. Nielsen, S. Brunak, and O. Lund. Systematic characterisation of cellular localisation and expression profiles of proteins containing mhc ligands. *PLoS One*, 4(10):e7448, 2009.

C. Kahl, B. E. Storer, B. M. Sandmaier, M. Mielcarek, M. B. Maris, K. G. Blume, D. Niederwieser, T. R. Chauncey, S. J. Forman, E. Agura, J. F. Leis, B. Bruno, A. Langston, M. A. Pulsipher, P. A. McSweeney, J. C. Wade, E. Epner, F. Bo Petersen, W. A. Bethge, D. G. Maloney, and R. Storb. Relapse risk in patients with malignant diseases given allogeneic hematopoietic cell transplantation after nonmyeloablative conditioning. *Blood*, 110(7): 2744--8, 2007.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, second edition.

M. Kamei, Y. Nannya, H. Torikai, T. Kawase, K. Taura, Y. Inamoto, T. Takahashi, M. Yazaki, S. Morishima, K. Tsujimura, K. Miyamura, T. Ito, H. Togari, S. R. Riddell, Y. Kodera, Y. Morishima, T. Takahashi, K. Kuzushima, S. Ogawa, and Y. Akatsuka. Hapmap scanning of novel human minor histocompatibility antigens. *Blood*, 113(21):5041--8, 2009.

C. Karanes, G. O. Nelson, P. Chitphakdithai, E. Agura, K. K. Ballen, C. D. Bolan, D. L. Porter, J. P. Uberti, R. J. King, and D. L. Confer. Twenty years of unrelated donor hematopoietic cell transplantation for adult recipients facilitated by the national marrow donor program. *Biol Blood Marrow Transplant*, 14(9 Suppl):8--15, 2008.

T. Kawase, Y. Akatsuka, H. Torikai, S. Morishima, A. Oka, A. Tsujimura, M. Miyazaki, K. Tsujimura, K. Miyamura, S. Ogawa, H. Inoko, Y. Morishima, Y. Kodera, K. Kuzushima, and T. Takahashi. Alternative splicing due to an intronic snp in hmsd generates a novel minor histocompatibility antigen. *Blood*, 110(3):1055--63, 2007.

T. Kawase, Y. Nannya, H. Torikai, G. Yamamoto, M. Onizuka, S. Morishima, K. Tsujimura, K. Miyamura, Y. Kodera, Y. Morishima, T. Takahashi, K. Kuzushima, S. Ogawa, and Y. Akatsuka. Identification of human minor histocompatibility antigens based on genetic association with highly parallel genotyping of pooled dna. *Blood*, 111(6):3286--94, 2008.

J. H. Kessler and C. J. Melief. Identification of t-cell epitopes for cancer immunotherapy. *Leukemia*, 21(9):1859--74, 2007.

A. R. Khan, B. M. Baker, P. Ghosh, W. E. Biddison, and D. C. Wiley. The structure and stability of an hla-a*0201/octameric tax peptide complex with an empty conserved peptide-n-terminal binding site. *J Immunol*, 164(12):6398--405, 2000.

P. Kiepiela, 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, J. Szinger, C. Day, P. Klenerman, J. Mullins, B. Korber, H. M. Coovadia, B. D. Walker, and P. J. Goulder. Dominant influence of hla-b in mediating the potential co-evolution of hiv and hla. *Nature*, 432(7018): 769--75, 2004.

P. Kiepiela, K. Ngumbela, C. Thobakgale, D. Ramduth, I. Honeyborne, E. Moodley, S. Reddy, C. de Pierres, Z. Mncube, N. Mkhwanazi, K. Bishop, M. van der Stok, K. Nair, N. Khan, H. Crawford, R. Payne, A. Leslie, J. Prado, A. Prendergast, J. Frater, N. McCarthy, C. Brander, G. H. Learn, D. Nickle, C. Rousseau, H. Coovadia, J. I. Mullins, D. Heckerman, B. D. Walker, and P. Goulder. Cd8+ t-cell responses to different hiv proteins have discordant associations with viral load. *Nat Med*, 13(1):46--53, 2007.

I. Kim, S. S. Yoon, K. H. Lee, B. Keam, T. M. Kim, J. S. Kim, H. G. Kim, M. D. Oh, K. S. Han, M. H. Park, S. Park, and B. K. Kim. Comparative outcomes of reduced intensity and myeloablative allogeneic hematopoietic stem cell transplantation in patients under 50 with hematologic malignancies. *Clin Transplant*, 20(4):496--503, 2006.

L. P. Koh and N. Chao. Haploidentical hematopoietic cell transplantation. *Bone Marrow Transplant*, 42 Suppl 1:S60--S63, 2008.

H. J. Kolb, A. Schattenberg, J. M. Goldman, B. Hertenstein, N. Jacobsen, W. Arcese, P. Ljungman, A. Ferrant, L. Verdonck, D. Niederwieser, F. van Rhee, J. Mittermueller, T. de Witte, E. Holler, and H. Ansari. Graft-versus-leukemia effect of donor lymphocyte transfusions in marrow grafted patients. *Blood*, 86(5):2041--50, 1995.

B. T. M. Korber, C. Brander, B. F. Haynes, R. Koup, J. P. Moore, B. D. Walker, and D. I. Watkins. Hiv molecular immunology 2009. *Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 09-05941*, 2009. http://www.hiv.lanl.gov/content/immunology.

B. Kornblit, L. Munthe-Fog, S. L. Petersen, H. O. Madsen, L. Vindelov, and P. Garred. The genetic variation of the human hmgb1 gene. *Tissue Antigens*, 70(2):151--6, 2007.

B. Kornblit, T. Masmas, H. O. Madsen, L. P. Ryder, A. Svejgaard, B. Jakobsen, H. Sengelov, G. Olesen, C. Heilmann, E. Dickmeiss, S. L. Petersen, and L. Vindelov. Haematopoietic cell transplantation with non-myeloablative conditioning in denmark: disease-specific outcome, complications and hospitalization requirements of the first 100 transplants. *Bone Marrow Transplant*, 41(10):851--9, 2008.

M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, S. Brunak, O. Lund, and M. Nielsen. An integrative approach to ctl epitope prediction: a combined algorithm integrating mhc class i binding, tap transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*, 35 (8):2295--303, 2005.

S. J. Lee, G. Vogelsang, and M. E. Flowers. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant*, 9(4):215--33, 2003.

C. Leisner, N. Loeth, K. Lamberth, S. Justesen, C. Sylvester-Hvid, E. G. Schmidt, M. Claesson, S. Buus, and A. Stryhn. One-pot, mix-and-read peptide-mhc tetramers. *PLoS One*, 3(2): e1678, 2008.

H. H. Lin, S. Ray, S. Tongchusak, E. L. Reinherz, and V. Brusic. Evaluation of mhc class i peptide binding prediction servers: applications for vaccine research. *BMC Immunol*, 9:8, 2008a.

H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic. Evaluation of mhc-ii peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*, 9 Suppl 12:S22, 2008b.

C. Lundegaard, O. Lund, C. Kesmir, S. Brunak, and M. Nielsen. Modeling the adaptive immune system: predictions and simulations. *Bioinformatics*, 23(24):3265--75, 2007.

C. Lundegaard, O. Lund, S. Buus, and M. Nielsen. Major histocompatibility complex class i binding predictions as a tool in epitope discovery. *Immunology*, 130(3):309--18, 2010.

W. A. Marijt, M. H. Heemskerk, F. M. Kloosterboer, E. Goulmy, M. G. Kester, M. A. van der Hoorn, S. A. van Luxemburg-Heys, M. Hoogeboom, T. Mutis, J. W. Drijfhout, J. J. van Rood, R. Willemze, and J. H. Falkenburg. Hematopoiesis-restricted minor histocompatibility antigens ha-1- or ha-2-specific t cells can induce complete remissions of relapsed leukemia. *Proc Natl Acad Sci U S A*, 100(5):2742--7, 2003.

M. B. Maris, D. Niederwieser, B. M. Sandmaier, B. Storer, M. Stuart, D. Maloney, E. Petersdorf, P. McSweeney, M. Pulsipher, A. Woolfrey, T. Chauncey, E. Agura, S. Heimfeld, J. Slattery, U. Hegenbart, C. Anasetti, K. Blume, and R. Storb. Hla-matched unrelated donor hematopoietic cell transplantation after nonmyeloablative conditioning for patients with hematologic malignancies. *Blood*, 102(6):2021--30, 2003.

L. Meadows, W. Wang, J. M. den Haan, E. Blokland, C. Reinhardus, J. W. Drijfhout, J. Shabanowitz, R. Pierce, A. I. Agulnik, C. E. Bishop, D. F. Hunt, E. Goulmy, and V. H. Engelhard. The hla-a*0201-restricted h-y antigen contains a posttranslationally modified cysteine that significantly affects t cell recognition. *Immunity*, 6(3):273--81, 1997.

A. Mengarelli, A. Iori, C. Guglielmi, A. Romano, R. Cerretti, C. Torromeo, A. Micozzi, S. Fenu, L. Laurenti, V. Donato, L. De Felice, and W. Arcese. Standard versus alternative myeloablative conditioning regimens in allogeneic hematopoietic stem cell transplantation for high-risk acute leukemia. *Haematologica*, 87(1):52--8, 2002.

M. Mielcarek, P. J. Martin, W. Leisenring, M. E. Flowers, D. G. Maloney, B. M. Sandmaier, M. B. Maris, and R. Storb. Graft-versus-host disease after nonmyeloablative versus conventional hematopoietic stem cell transplantation. *Blood*, 102(2):756--62, 2003.

B. Mommaas, J. Kamp, J. W. Drijfhout, N. Beekman, F. Ossendorp, P. Van Veelen, J. Den Haan, E. Goulmy, and T. Mutis. Identification of a novel hla-b60-restricted t cell epitope of the minor histocompatibility antigen ha-1 locus. *J Immunol*, 169(6):3131--6, 2002.

M. Mora, D. Veggi, L. Santini, M. Pizza, and R. Rappuoli. Reverse vaccinology. *Drug Discov Today*, 8(10):459--64, 2003.

M. Moutaftsi, B. Peters, V. Pasquetto, D. C. Tscharke, J. Sidney, H. H. Bui, H. Grey, and A. Sette. A consensus epitope prediction approach identifies the breadth of murine t(cd8+)-cell responses to vaccinia virus. *Nat Biotechnol*, 24(7):817--9, 2006.

A. Mullally and J. Ritz. Beyond hla: the significance of genomic variation for allogeneic hematopoietic stem cell transplantation. *Blood*, 109(4):1355--62, 2007.

M. Murata, E. H. Warren, and S. R. Riddell. A human minor histocompatibility antigen resulting from differential expression due to a gene deletion. *J Exp Med*, 197(10):1279--89, 2003.

Kenneth Murphy, Paul Travers, and Mark Walport. *Janeway's Immunobiology*. Garland Science, Taylor & Francis Group, LCC, seventh edition.

M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, and O. Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, 12(5):1007--17, 2003.

M. Nielsen, C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Roder, B. Peters, A. Sette, O. Lund, and S. Buus. Netmhcpan, a method for quantitative predictions of peptide binding to any hla-a and -b locus protein of known sequence. *PLoS One*, 2(8): e796, 2007. http://www.cbs.dtu.dk/services/NetMHCpan.

M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, and O. Lund. Quantitative predictions of peptide binding to any hla-dr molecule of known sequence: Netmhciipan. *PLoS Comput Biol*, 4(7):e1000107, 2008.

M. Nielsen, O. Lund, S. Buus, and C. Lundegaard. Mhc class ii epitope predictive algorithms. *Immunology*, 130(3):319--28, 2010.

C. A. O'Callaghan and J. I. Bell. Structure and function of the human mhc class ib molecules hla-e, hla-f and hla-g. *Immunol Rev*, 163:129--38, 1998.

Y. Ofran, H. T. Kim, V. Brusic, L. Blake, M. Mandrell, C. J. Wu, S. Sarantopoulos, R. Bellucci, D. B. Keskin, R. J. Soiffer, J. H. Antin, and J. Ritz. Diverse patterns of t-cell response against multiple newly identified human y chromosome-encoded minor histocompatibility epitopes. *Clin Cancer Res*, 16(5):1642--51, 2010.

S. Ogawa, A. Matsubara, M. Onizuka, K. Kashiwase, M. Sanada, M. Kato, Y. Nannya, Y. Akatsuka, M. Satake, J. Takita, S. Chiba, H. Saji, E. Maruya, H. Inoko, Y. Morishima, Y. Kodera, and S. Takehiko. Exploration of the genetic basis of gvhd by genetic association studies. *Biol Blood Marrow Transplant*, 15(1 Suppl):39--41, 2008.

K. C. Parker, M. A. Bednarek, and J. E. Coligan. Scheme for ranking potential hla-a2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, 152 (1):163--75, 1994.

R. M. Pelletier and S. W. Byers. The blood-testis barrier and sertoli cell junctions: structural considerations. *Microsc Res Tech*, 20(1):3--33, 1992.

A. Perez-Garcia, R. De la Camara, A. Torres, M. Gonzalez, A. Jimenez, and D. Gallardo. Minor histocompatibility antigen ha-8 mismatch and clinical outcome after hla-identical sibling donor allogeneic stem cell transplantation. *Haematologica*, 90(12):1723--4, 2005.

R. A. Pierce, E. D. Field, J. M. den Haan, J. A. Caldwell, F. M. White, J. A. Marto, W. Wang, L. M. Frost, E. Blokland, C. Reinhardus, J. Shabanowitz, D. F. Hunt, E. Goulmy, and V. H. Engelhard. Cutting edge: the hla-a*0101-restricted hy minor histocompatibility antigen originates from dffry and contains a cysteinylated cysteine residue as identified by a novel mass spectrometric technique. *J Immunol*, 163(12):6360--4, 1999.

R. A. Pierce, E. D. Field, T. Mutis, T. N. Golovina, C. Von Kap-Herr, M. Wilke, J. Pool, J. Shabanowitz, M. J. Pettenati, L. C. Eisenlohr, D. F. Hunt, E. Goulmy, and V. H. Engelhard. The ha-2 minor histocompatibility antigen is derived from a diallelic gene encoding a novel human class i myosin protein. *J Immunol*, 167(6):3223--30, 2001.

M. Pizza, V. Scarlato, V. Masignani, M. M. Giuliani, B. Arico, M. Comanducci, G. T. Jennings, L. Baldi, E. Bartolini, B. Capecchi, C. L. Galeotti, E. Luzzi, R. Manetti, E. Marchetti, M. Mora, S. Nuti, G. Ratti, L. Santini, S. Savino, M. Scarselli, E. Storni, P. Zuo, M. Broeker, E. Hundt, B. Knapp, E. Blair, T. Mason, H. Tettelin, D. W. Hood, A. C. Jeffries, N. J. Saunders, D. M. Granoff, J. C. Venter, E. R. Moxon, G. Grandi, and R. Rappuoli. Identification of vaccine candidates against serogroup b meningococcus by whole-genome sequencing. *Science*, 287(5459):1816--20, 2000.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61--5, 2007.

D. Przepiorka, D. Weisdorf, P. Martin, H. G. Klingemann, P. Beatty, J. Hows, and E. D. Thomas. 1994 consensus conference on acute gvhd grading. *Bone Marrow Transplant*, 15(6):825--8, 1995.

H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. Syfpeithi: database for mhc ligands and peptide motifs. *Immunogenetics*, 50(3-4):213--9, 1999.

S. S. Randolph, T. A. Gooley, E. H. Warren, F. R. Appelbaum, and S. R. Riddell. Female donors contribute to a selective graft-versus-leukemia effect in male recipients of hla-matched, related hematopoietic stem cell transplants. *Blood*, 103(1):347--52, 2004.

X. Rao, A. I. Costa, D. van Baarle, and C. Kesmir. A comparative study of hla binding affinity and ligand diversity: implications for generating immunodominant cd8+ t cell responses. *J Immunol*, 182(3):1526--32, 2009.

N. Rapin, I. Hoof, O. Lund, and M. Nielsen. Mhc motif viewer. *Immunogenetics*, 60(12): 759--65, 2008.

R. Rappuoli. Reverse vaccinology. *Curr Opin Microbiol*, 3(5):445--50, 2000.

S. R. Riddell, M. Bleakley, T. Nishida, C. Berger, and E. H. Warren. Adoptive transfer of allogeneic antigen-specific t cells. *Biol Blood Marrow Transplant*, 12(1 Suppl 1):9--12, 2006.

J. Robinson, M. J. Waller, P. Parham, J. G. Bodmer, and S. G. Marsh. Imgt/hla database--a sequence database for the human major histocompatibility complex. *Nucleic Acids Res*, 29 (1):210--3, 2001.

J. Robinson, M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, and S. G. Marsh. Imgt/hla and imgt/mhc: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, 31(1):311--4, 2003.

J. Robinson, M. J. Waller, S. C. Fail, H. McWilliam, R. Lopez, P. Parham, and S. G. Marsh. The imgt/hla database. *Nucleic Acids Res*, 37(Database issue):D1013--7, 2009.

K. L. Rock, I. A. York, T. Saric, and A. L. Goldberg. Protein degradation and the generation of mhc class i-presented peptides. *Adv Immunol*, 80:1--70, 2002.

D. Roopenian, E. Y. Choi, and A. Brown. The immunogenomics of minor histocompatibility antigens. *Immunol Rev*, 190:86--94, 2002.

K. V. Rosinski, N. Fujii, J. K. Mito, K. K. Koo, S. M. Xuereb, O. Sala-Torra, J. S. Gibbs, J. P. Radich, Y. Akatsuka, B. J. Van den Eynde, S. R. Riddell, and E. H. Warren. Ddx3y encodes a class i mhc-restricted h-y antigen that is expressed in leukemic stem cells. *Blood*, 111(9): 4817--26, 2008.

J. M. Rowe, H. M. Lazarus, and A. M. Carella. *Handbook of Bone Marrow Transplantaion*. Martin Dunitz Ltd, first edition.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097--100, 1990.

M. M. Schuler, P. Donnes, M. D. Nastke, O. Kohlbacher, H. G. Rammensee, and S. Stevanovic. Snep: Snp-derived epitope prediction program for minor h antigens. *Immunogenetics*, 57 (11):816--20, 2005.

T. Serwold, F. Gonzalez, J. Kim, R. Jacob, and N. Shastri. Eraap customizes peptides for mhc class i molecules in the endoplasmic reticulum. *Nature*, 419(6906):480--3, 2002.

A. Sette, A. Vitiello, B. Reherman, P. Fowler, R. Nayersina, W. M. Kast, C. J. Melief, C. Oseroff, L. Yuan, J. Ruppert, J. Sidney, M. F. del Guercio, S. Southwood, R. T. Kubo, R. W. Chesnut, H. M. Grey, and F. V. Chisari. The relationship between class i binding affinity and immunogenicity of potential cytotoxic t cell epitopes. *J Immunol*, 153(12):5586--92, 1994.

H. M. Shulman, K. M. Sullivan, P. L. Weiden, G. B. McDonald, G. E. Striker, G. E. Sale, R. Hackman, M. S. Tsoi, R. Storb, and E. D. Thomas. Chronic graft-versus-host syndrome in man. a long-term clinicopathologic study of 20 seattle patients. *Am J Med*, 69(2):204--17, 1980.

H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825--37, 2003.

E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, 28(1):352--5, 2000.

R. J. Soiffer. *Hematopoietic Stem Cell Transplantation*. Humana Press, second edition.

R.J Soiffer. T-cell depletion to prevent graft-vs.-host disease. *In: Blume KG, Forman SJ, Appelbaum FR (eds). Thomas' Hematopoietic Cell Transplantation, 3rd edn. Blackwell: Malden, MA*, 2003.

R. Spaapen and T. Mutis. Targeting haematopoietic-specific minor histocompatibility antigens to distinguish graft-versus-tumour effects from graft-versus-host disease. *Best Pract Res Clin Haematol*, 21(3):543--57, 2008.

S. Spellman, M. B. Warden, M. Haagenson, B. C. Pietz, E. Goulmy, E. H. Warren, T. Wang, and T. M. Ellis. Effects of mismatching for minor histocompatibility antigens on clinical outcomes in hla-matched, unrelated hematopoietic stem cell transplants. *Biol Blood Marrow Transplant*, 15(7):856--63, 2009.

E. Spierings, A. G. Brickner, J. A. Caldwell, S. Zegveld, N. Tatsis, E. Blokland, J. Pool, R. A. Pierce, S. Mollah, J. Shabanowitz, L. C. Eisenlohr, P. van Veelen, F. Ossendorp, D. F. Hunt, E. Goulmy, and V. H. Engelhard. The minor histocompatibility antigen ha-3 arises from differential proteasome-mediated cleavage of the lymphoid blast crisis (lbc) oncoprotein. *Blood*, 102(2):621--9, 2003a.

E. Spierings, C. J. Vermeulen, M. H. Vogt, L. E. Doerner, J. H. Falkenburg, T. Mutis, and E. Goulmy. Identification of hla class ii-restricted h-y-specific t-helper epitope evoking cd4+ t-helper cells in h-y-mismatched transplantation. *Lancet*, 362(9384):610--5, 2003b.

E. Spierings, J. Drabbels, M. Hendriks, J. Pool, M. Spruyt-Gerritse, F. Claas, and E. Goulmy. A uniform genomic minor histocompatibility antigen typing methodology and database designed to facilitate clinical applications. *PLoS One*, 1:e42, 2006.

M. Stern, R. Brand, T. de Witte, A. Sureda, V. Rocha, J. Passweg, H. Baldomero, D. Niederwieser, and A. Gratwohl. Female-versus-male alloreactivity as a model for minor histocompatibility antigens in hematopoietic stem cell transplantation. *Am J Transplant*, 8(10): 2149--57, 2008.

S. Stevanovic. Antigen processing is predictable: From genes to t cell epitopes. *Transpl Immunol*, 14(3-4):171--4, 2005.

R. Storb. Can reduced-intensity allogeneic transplantation cure older adults with aml? *Best Pract Res Clin Haematol*, 20(1):85--90, 2007.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440--5, 2003.

T. Stranzl, M. V. Larsen, C. Lundegaard, and M. Nielsen. Netctlpan: pan-specific mhc class i pathway epitope predictions. *Immunogenetics*, 62(6):357--68, 2010.

A. N. Stumpf, E. D. van der Meijden, C. A. van Bergen, R. Willemze, J. H. Falkenburg, and M. Griffioen. Identification of 4 new hla-dr-restricted minor histocompatibility antigens as hematopoietic targets in antitumor immunity. *Blood*, 114(17):3684--92, 2009.

K. M. Sullivan. Graft-vs-host disease. *In: Blume KG, Forman SJ, Appelbaum FR (eds). Thomas' Hematopoietic Cell Transplantation, 3rd edn. Blackwell: Malden, MA*, pages pp 635--664, 2003.

S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P. M. Kloetzel, H. G. Rammensee, H. Schild, and H. G. Holzhutter. Modeling the mhc class i pathway by combining predictions of proteasomal cleavage, tap transport and mhc class i binding. *Cell Mol Life Sci*, 62(9):1025--37, 2005.

S. Terakura, M. Murata, E. H. Warren, A. Sette, J. Sidney, T. Naoe, and S. R. Riddell. A single minor histocompatibility antigen encoded by ugt2b17 and presented by human leukocyte antigen-a*2902 and -b*4403. *Transplantation*, 83(9):1242--8, 2007.

J. M. Tiercy, G. Nicoloso, J. Passweg, U. Schanz, R. Seger, Y. Chalandon, D. Heim, T. Gungor, P. Schneider, R. Schwabe, and A. Gratwohl. The probability of identifying a 10/10 hla allele-matched unrelated donor is highly predictable. *Bone Marrow Transplant*, 40(6):515--22, 2007.

M. Toebes, M. Coccoris, A. Bins, B. Rodenko, R. Gomez, N. J. Nieuwkoop, W. van de Kasteele, G. F. Rimmelzwaan, J. B. Haanen, H. Ovaa, and T. N. Schumacher. Design and use of conditional mhc class i ligands. *Nat Med*, 12(2):246--51, 2006.

A. Townsend and J. Trowsdale. The transporters associated with antigen presentation. *Semin Cell Biol*, 4(1):53--61, 1993.

L. H. Tseng, M. T. Lin, J. A. Hansen, T. Gooley, J. Pei, A. G. Smith, E. G. Martin, E. W. Petersdorf, and P. J. Martin. Correlation between disparity for the minor histocompatibility antigen ha-1 and the development of acute graft-versus-host disease after allogeneic marrow transplantation. *Blood*, 94(8):2911--4, 1999.

S. Uebel and R. Tampe. Specificity of the proteasome and the tap transporter. *Curr Opin Immunol*, 11(2):203--8, 1999.

UniProt. Human polymorphisms and disease mutations (humsavar, release 57.12, december 2009). 2009. http://www.uniprot.org/docs/humsavar.

D. Valcarcel, R. Martino, A. Sureda, C. Canals, A. Altes, J. Briones, M. A. Sanz, R. Parody, M. Constans, S. L. Villela, S. Brunet, and J. Sierra. Conventional versus reduced-intensity conditioning regimen for allogeneic stem cell transplantation in patients with hematological malignancies. *Eur J Haematol*, 74(2):144--51, 2005.

C. A. van Bergen, M. G. Kester, I. Jedema, M. H. Heemskerk, S. A. van Luxemburg-Heijs, F. M. Kloosterboer, W. A. Marijt, A. H. de Ru, M. R. Schaafsma, R. Willemze, P. A. van Veelen, and J. H. Falkenburg. Multiple myeloma-reactive t cells recognize an activation-induced minor histocompatibility antigen encoded by the atp-dependent interferon-responsive (adir) gene. *Blood*, 109(9):4089--96, 2007.

M. H. Vogt, R. A. de Paus, P. J. Voogt, R. Willemze, and J. H. Falkenburg. Dffry codes for a new human male-specific minor transplantation antigen involved in bone marrow graft rejection. *Blood*, 95(3):1100--5, 2000a.

M. H. Vogt, E. Goulmy, F. M. Kloosterboer, E. Blokland, R. A. de Paus, R. Willemze, and J. H. Falkenburg. Uty gene codes for an hla-b60-restricted human male-specific minor histocompatibility antigen involved in stem cell graft rejection: characterization of the critical polymorphic amino acid residues for t-cell recognition. *Blood*, 96(9):3126--32, 2000b.

M. H. Vogt, J. W. van den Muijsenberg, E. Goulmy, E. Spierings, P. Kluck, M. G. Kester, R. A. van Soest, J. W. Drijfhout, R. Willemze, and J. H. Falkenburg. The dby gene codes for an hla-dq5-restricted human male-specific minor histocompatibility antigen involved in graft-versus-host disease. *Blood*, 99(8):3027--32, 2002.

P. J. Voogt, W. E. Fibbe, W. A. Marijt, E. Goulmy, W. F. Veenhof, M. Hamilton, A. Brand, F. E. Zwann, R. Willemze, J. J. van Rood, and et al. Rejection of bone-marrow graft by recipient-derived cytotoxic t lymphocytes against minor histocompatibility antigens. *Lancet*, 335 (8682):131--4, 1990.

E. C. Walsh, K. A. Mather, S. F. Schaffner, L. Farwell, M. J. Daly, N. Patterson, M. Cullen, M. Carrington, T. L. Bugawan, H. Erlich, J. Campbell, J. Barrett, K. Miller, G. Thomson, E. S. Lander, and J. D. Rioux. An integrated haplotype map of the human major histocompatibility complex. *Am J Hum Genet*, 73(3):580--90, 2003.

W. Wang, L. R. Meadows, J. M. den Haan, N. E. Sherman, Y. Chen, E. Blokland, J. Shabanowitz, A. I. Agulnik, R. C. Hendrickson, C. E. Bishop, and et al. Human h-y: a male-specific histocompatibility antigen derived from the smcy protein. *Science*, 269(5230):1588--90, 1995.

E. H. Warren, M. A. Gavin, E. Simpson, P. Chandler, D. C. Page, C. Disteche, K. A. Stankey, P. D. Greenberg, and S. R. Riddell. The human uty gene encodes a novel hla-b8-restricted h-y antigen. *J Immunol*, 164(5):2807--14, 2000.

E. H. Warren, N. J. Vigneron, M. A. Gavin, P. G. Coulie, V. Stroobant, A. Dalet, S. S. Tykodi, S. M. Xuereb, J. K. Mito, S. R. Riddell, and B. J. Van den Eynde. An antigen produced by splicing of noncontiguous peptides in the reverse order. *Science*, 313(5792):1444--7, 2006.

E. H. Warren, N. Fujii, Y. Akatsuka, C. N. Chaney, J. K. Mito, K. R. Loeb, T. A. Gooley, M. L. Brown, K. K. Koo, K. V. Rosinski, S. Ogawa, A. Matsubara, F. R. Appelbaum, and S. R. Riddell. Therapy of relapsed leukemia after allogeneic hematopoietic cell transplantation with t cells specific for minor histocompatibility antigens. *Blood*, 115(19):3869--78, 2010.

L. A. Welniak, B. R. Blazar, and W. J. Murphy. Immunobiology of allogeneic hematopoietic stem cell transplantation. *Annu Rev Immunol*, 25:139--70, 2007.

J. W. Yewdell and J. R. Bennink. Immunodominance in major histocompatibility complex class i-restricted t lymphocyte responses. *Annu Rev Immunol*, 17:51--88, 1999.

J. W. Yewdell, E. Reits, and J. Neefjes. Making sense of mass destruction: quantitating mhc class i antigen presentation. *Nat Rev Immunol*, 3(12):952--61, 2003.

H. Zhang, C. Lundegaard, and M. Nielsen. Pan-specific mhc class i predictors: a benchmark of hla class i pan-specific prediction methods. *Bioinformatics*, 25(1):83--9, 2009.

Appendices

Appendix A

Selection of candidate H-Y mHags

The final peptide selection of 324 candidate H-Y mHags, as explained in Chapter 2, can be found online at www.gersborg.dk/PhD

Below is an example of the information found in the Excel file.

Protein	Peptide	Position	Predictions
UTY	LPAFARVVSA	1053	B0702 (38nM), B3502(74nM), ….
	LPAFARVVS	1053	B0702 (52nM), B3501 (460nM), B3502(99nM), ….
	LPAFARVV	1053	B0702 (22nM), B3501 (481nM), B3502(31nM), ….
….	….	….	….

Figure A.1: **Example of predicted mHag from the final H-Y selection.** The first peptide is one of the 324 peptides, which have been bought for experimental validation, whereas the two next peptides are submers which also have significant predictions. The Excel file has different sheets listing all peptides, peptides per HLA allele, and peptides per patient.

Appendix B

CD4+ responses to peptides from the Y chromosome

As mentioned in Chapter 2, some CD4+ responses were observed, when running ICS assays with the peptides selected for their predicted binding to HLA class I molecules.

patient	Sekvens	protein	HLA II					
289	YFYYNAFHWAI	UTY	DRB1*0701	DRB1*1501	DQB1*0202	DQB1*0602		
257	GSSKMFNY	NLGN4Y	**DRB1*0301**	DRB1*1302	**DQB1*0201**	DQB1*0604		
	NEYKFYVPENL	PCDH11Y	**DRB1*0301**	DRB1*1302	**DQB1*0201**	DQB1*0604		
297	YPAGFIDVISI	RPS4Y2	**DRB1*0101**	DRB1*1501	DQB1*0501	DQB1*0602		
627	RVLAIQLKR	USP9Y	DRB1*1302	DRB1*1501	**DRB3*0301**	**DRB5*0101**	DQB1*0602	DQB1*0604
	APSAHRGSLVI	JARID1D	DRB1*1302	DRB1*1501	**DRB3*0301**	**DRB5*0101**	DQB1*0602	DQB1*0604
AET	VYLQYLRSGEL	USP9Y	**DRB1*0101**	**DRB1*0401**	DQB1*0602	DQB1*0604		
	IEIVPHLL	USP9Y	**DRB1*0101**	**DRB1*0401**	DQB1*0602	DQB1*0604		
283	VYFYYNAFHW	UTY	DRB1*0804	DRB1*1101	DRB3*0202	DQB1*0301	DQB1*0402	DPB1*03 or *78,02
	VALFSSCPVAY	USP9Y	DRB1*0804	DRB1*1101	DRB3*0202	DQB1*0301	DQB1*0402	DPB1*03 or *78,02
	KVADVDLAVPV	CYorf15B	DRB1*0804	DRB1*1101	DRB3*0202	DQB1*0301	DQB1*0402	DPB1*03 or *78,02
	SLMPLLQLSY	JARID1D	DRB1*0804	DRB1*1101	DRB3*0202	DQB1*0301	DQB1*0402	DPB1*03 or *78,02
611	WEEKAHFCL	JARID1D	DRB1*0701	DRB1*1101	DRB3*0202	DRB4*0101	DQB1*0202	DQB1*0301
	RVLAIQLKR	USP9Y	DRB1*0701	DRB1*1101	DRB3*0202	DRB4*0101	DQB1*0202	DQB1*0301
	RTIRYPDPVIK	RPS4Y1	DRB1*0701	DRB1*1101	DRB3*0202	DRB4*0101	DQB1*0202	DQB1*0301

Figure B.1: **CD4+ responses to peptides from the Y chromosome.** All the peptides listed here were found to elicit a CD4+ response. The (few) HLA alleles marked in blue, have been produced in Laboratory of Experimental Immunology. Binding assays are planned for these. Tetramers are not yet available for class II alleles.

Appendix C

Selection of peptides for nsSNP derived mHags

The ideal peptide selection as explained in Chapter 4, can be found online at www.gersborg.dk/PhD

Below is an example of peptides listed in the Excel file.

Protein	nsSNP	Peptide	Matching patients	Predictions
SP110	rs1135791	G**M**TLGELLK	8	A0301 (174nM), A1101 (207nM)….
		G**T**TLGELLK	11	A0301 (387nM), A1101 (42nM) ….
		MTLGELLKRK	10	A0301 (159nM), A1101 (15nM) ….
		TTLGELLKRK	7	A0301 (555nM), A1101 (21nM) ….
….	….	….	….	….

Figure C.1: **Example of predicted mHags from the nsSNP selection.** The example shows 4 predicted mHags around the same nsSNP. Note that both versions of a peptide can be selected as candidate mHags for different patients. The Excel file has different sheets listing all peptides, peptides that have been bought, and relevant peptides per patient.