# Analysis of cytotoxic T cell epitopes in relation to cancer

**Stranzl, Thomas; Brunak, Søren; Larsen, Mette Voldby**

*Publication date:*
2012

*Document Version*
Publisher's PDF, also known as Version of record

# Analysis of cytotoxic T cell epitopes in relation to cancer

Thomas Stranzl

October 31, 2011

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

## Preface

This thesis was prepared at the Department of Systems Biology, the Technical University of Denmark, in partial fulfillment of the requirements for acquiring the Ph.D. degree.

# Contents

## Summary

The human immune system is a highly adaptable system, defending our bodies against pathogens and tumor cells. Cytotoxic T cells (CTL) are cells of the adaptive immune system, capable of inducing a programmed cell death and thus able to eliminate infected or tumor cells. CTLs discriminate between healthy and infected cells based on peptide fragments presented on the cells surface. All nucleated cells present these peptide fragments in complex with Major Histocompatibility Complex (MHC) class I molecules. Peptides that are recognized by CTLs are called epitopes and induce the CTLs to subsequently kill the infected cells.

The focus of my PhD project has been on improving a method for CTL epitope pathway prediction, on analyzing the epitope density in the alternative cancer exome, and on a study investigating minor histocompatibility antigens (mHags) associated with leukemia.

Part I is an introduction to the fields covered in the thesis. Part II describes a pan-specific, integrative approach for the prediction of CTL epitopes. The presented method, NetCTLpan, an improved and extended version of NetCTL, performs predictions for all MHC class I molecules with known protein sequence and allows predictions for 8, 9, 10 and 11-mer epitopes. One of the major benefits of the method is its optimization to achieve high specificity. Its low false positive rate is especially useful in rational reverse immunogenetic epitope discovery approaches. When this method is compared to the NetMHCpan and NetCTL methods, the experimental effort to identify 90% of new epitopes can be reduced by 15% and 40%, respectively.

Part III reports the results of an analysis investigating how the alternatively spliced cancer exome differs from the exome of normal tissue in terms of containing predicted MHC class I binding epitopes. We show that peptides unique to cancer splice variants comprise significantly fewer predicted HLA class I epitopes than peptides unique to spliced transcripts in normal tissue. We furthermore find that hydrophilic amino acids are significantly enriched in the unique carcinoma sequences, which contribute to the lower likelihood of carcinoma-specific peptides to be predicted epitopes. Carcinoma is known to have developed mechanisms for evading the host's immune system. Here, we show for the first time that carcinoma has a bias towards fewer possible epitopes already at the step of mRNA splicing.

Part IV of the thesis deals with the analysis of 93 patient-donor pairs that underwent hematopoietic stem cell transplantation (HCT). HCT is a standard treatment for a variety of hematological diseases. Graft-versus-host disease is a possible complication after an HCT, where the recipient´s cells are perceived as foreign and the target of an immune response mediated by the donor´s transplanted immune cells. The immune response is provoked by epitopes unique to the patient, so-called mHags. Here, a gene-specific association between the number of SNPs or predicted mHags and the possible clinical outcome following an HCT is presented.

## Dansk resumé

Det humane immunsystem er et meget tilpasningsdygtigt system, som forsvarer vores krop mod patogener og tumorceller. Cytotoksiske T-celler (CTL) er celler fra vores adaptive immunsystem, som er i stand til at forårsage programmeret celledød og dermed eliminere inficerede celler eller tumorceller. CTLs kan skelne mellem raske og inficerede celler på baggrund af peptidfragmenter præsenteret på cellernes overflade. Alle kerneholdige celler præsenterer disse peptidfragmenter i kompleks med Major Histocompatibility Complex (MHC) klasse I molekyler. Peptidfragmenter, der genkendes af CTLs, kaldes epitoper og inducerer efterfølgende CTLs til at dræbe de inficerede celler.

Fokus i mit PhD projekt har været på at forbedre en metode til CTL epitop pathway forudsigelse, ved at analysere epitop-tætheden i det alternative cancer exom og ved et studie af minor histocompatibility antigener (mHags) associeret med leukæmi.

Del I er en introduktion til de områder, der bliver dækket i denne afhandling. Del II beskriver en pan-specifik integrativ tilgang til forudsigelsen af CTL epitoper. Den præsenterede metode, NetCTLpan, en forbedret og udvidet version af NetCTL, kan forudsige alle MHC klasse I molekyler med kendte protein sekvenser og tillader forudsigelser for 8, 9, 10 og 11-mer epitoper. En af de store fordele ved denne metode er, at den er optimeret til at opnå høj specificitet. Dens lave falsk positive rate er især brugbar i forbindelse med rationel omvendt immunogenetisk epitop opdagelsestilgange. Ved at sammenligne denne metode med NetMHCpan og NetCTL metoderne kan den eksperimentelle indsats, der er nødvendig for at identificere 90% af nye epitoper, reduceres med henholdsvis 15% og 40%.

Del III beskriver resultaterne af en analyse, hvor det undersøges, hvordan det alternativt splejsede cancer exom afviger fra exomet i normalt væv, i forbindelse med indholdet af forudsagte MHC klasse I epitoper. Vi viser, at peptider, som er unikke for cancer splejsningsvarianter, indeholder væsentligt færre forudsagte HLA klasse I epitoper end peptider, der er unikke for splejsede transkripter i normalt væv. Vi konstaterer ydermere, at hydrofile aminosyrer er signifikant beriget i de unikke karcinom sekvenser, hvilket bidrager til den lavere sandsynlighed for at forudsige epitoper i karcinomspecifikke peptider. Karcinoma er kendt for at have udviklet mekanismer til at undvige værtens immunsystem. Her viser vi for første gang, at karcinoma har en bias mod færre mulige epitoper allerede ved mRNA splejsningen.

Del IV af afhandlingen beskæftiger sig med analyse af 93 patient-donor par, der har fået foretaget en hæmatopoietisk stamcelle transplantation (HCT). HCT er en standard behandling for en lang række hæmatologiske sygdomme. Graft-versus-host sygdom er en mulig komplikation efter en HCT, hvor visse af modtagerens celler opfattes som fremmede, og donorens transplanterede immunceller medierer et immunrespons mod dem. Immunresponset er fremprovokeret af epitoper, der er unikke for patienten, såkaldte mHags. Her præsenteres en gen-specifik forbindelse mellem antallet af SNPs eller forudsagte mHags og mulige kliniske forløb efter en HCT.

## Acknowledgements

My time spent being a PhD student, under the supervision of Søren Brunak, at the Center for Biological Sequence Analysis (CBS) has been an invaluable experience. I am very thankful to Søren for providing such a stimulating environment, and especially for all of his fabulous ideas, inspiration and his broad knowledge on many subjects.

A big thank you to my co-supervisor Mette Voldby Larsen. As my day-to-day supervisor, she always had an open ear and provided support whenever needed. I am further grateful for her commenting and proofreading of the thesis.

Thanks to all current and past members of the Immunological Bioinformatics group, for providing a scientific framework and some great nights out. A special thanks to Morten Nielsen for at the same time bright and clear ideas, as well as to Ole Lund for scientific input and optimism.

I would like to thank all members of the Systems Biology group, especially Daniel, Konrad and Nils for insightful discussions, both at lunch or in the office, and further my officemates Sonny and Kirstine, for additionally extending my Danish language skills.

I have been very fortunate to collaborate with experts in the field of hematology at Rigshospitalet and Panum. Thank you all for valuable insights, especially Lars Vindeløv for bringing us all together.

My thanks go to Thomas A. Gerds at the Department of Biostatistics, KU for statistical guidance and R-code .

I thank all members of the CBS administration and the system administration for always smiling, taking care of practicalities and a good spirit.

I enjoyed the time with my colleagues at CBS. An extra thank you to the friends I found and to the friends I have.

## Papers included in the thesis

The following two papers are presented in this thesis. Additionally, work presented in Chapter 4 will be used in a manuscript.

- **Thomas Stranzl**, Mette V. Larsen, Claus Lundegaard, Morten Nielsen.
  NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics, Vol. 62, no. 6, June 2010, pp. 357-68*

- **Thomas Stranzl**, Mette V. Larsen, Ole Lund, Morten Nielsen, Søren Brunak.
  The cancer exome generated by alternative mRNA splicing dilutes predicted HLA class I epitope density. *Submitted to Cancer Research*

# Chapter 1

# Introduction

## 1.1 From DNA to protein

"Once information has got into a protein it can´t get out again". The central dogma of molecular biology enunciated by Francis Crick in 1958 is a framework describing the sequential transfer of information from DNA to protein [15]. Crick concluded that the flow of information is from nucleic acid (DNA or RNA) to protein. In general the dogma is covering three principles: DNA replication, a biological process where DNA is copied; Transcription, a step where DNA is copied to messenger RNA (mRNA); and translation, where mRNA is decoded to amino acids and further folded into a protein. More and more exceptions to the "Central Dogma" are described. RNA can make copies of itself and it is possible to go back to DNA from RNA. However, there is no known mechanism for proteins to make copies of themselves, nor is it known to be possible to go back to DNA or RNA from proteins.

### Alternative splicing

When the first euchromatic sequence of humans was sequenced and assembled in 2001 by Venter et al. they provided a major surprise: They found that the number of human genes is far lower (26,000 to 38,000) than earlier molecular predictions ranging from 50,000 to over 140,000 genes [118]. Further, it was shown that the human genome encodes only 20,000 to 25,000 protein-coding genes [12]. While one-third of the human genome would be transcribed as genes, only about 1.5% of the human genome codes for proteins [50].

Compared to other organisms, the amount of genes in human is nothing spectacular. We have approximately the same amount of protein-coding genes as flies and mice, the number of protein-coding genes for a roundworm (13,000) is more than half compared to humans, and rice was found

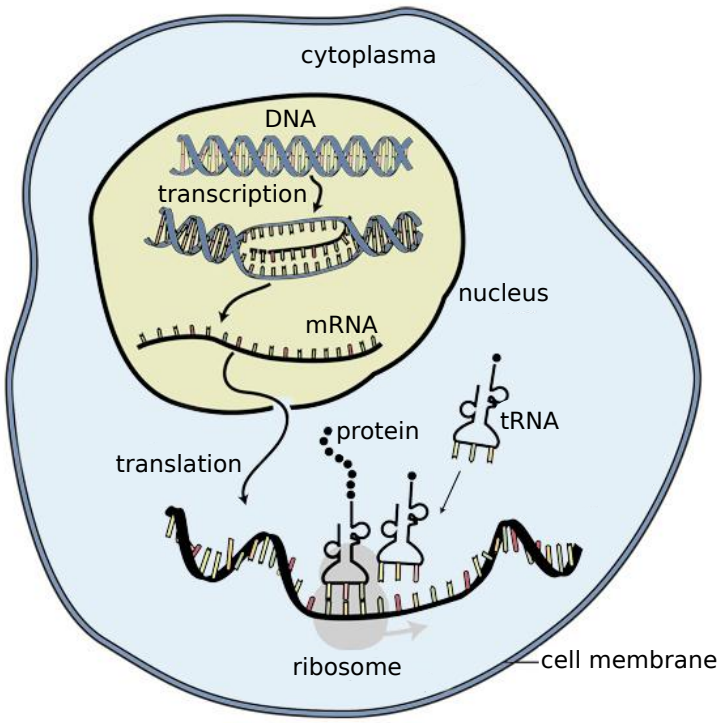


**Figure 1.1.** Gene transcription and translation. Double-stranded DNA unwinds and its triplet code is transcribed to mRNA. Translation is the process of protein synthesis, accomplished by mRNA along with ribosomes and tRNA.

to have more than 46,000 genes [131]. These findings raised the question for the source of organism complexity. Alternative splicing, which is the process of inclusion or exclusion of regions of the pre-mRNA, was discovered as one of the major mechanisms for increasing transcript diversity. It changes, by inclusion or skipping of exons, the structure of mRNA and further their encoded proteins. This may lead to affected function, stability or binding properties of encoded proteins.

Studies have shown that there are other, previously unknown mechanisms, like antisense transcription, where a large proportion of the genome can produce transcripts from both strands [41]. This shows that there are different mechanisms for increasing genomic diversity, but alternatively splicing, which has been shown to occur in 95% of multiexonic human genes, is still found to be a major driving force [75].

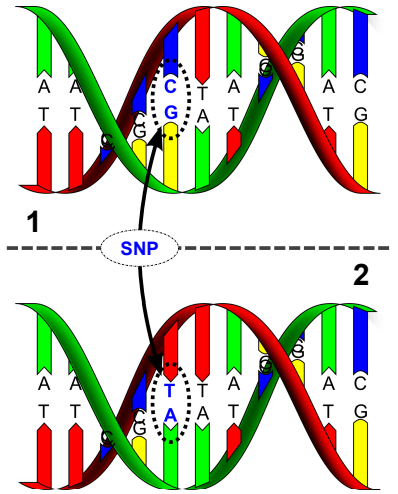


**Figure 1.2.** Two DNA fragments contain a difference in a single nucleotide. A C/T SNP is shown. (David Hall, Creative Commons License)

### Single nucleotide polymorphisms

A single-nucleotide polymorphism (SNP) is a single nucleotide variation (A, T, C or G) in the genome differing between members of a species or between homologue chromosomes within an individual. A SNP may, for example, replace the nucleotide cystine (C) with the nucleotide thymine (T) at a specific position (see Figure 1.2).

Several projects are genotyping SNPs with the aim of providing public resources for genetic research. Approximately 10 million SNPs in humans were identified by the Human Genome Project, the SNP Consortium and the International HapMap Project [1].

There are different types of SNPs, depending on which location within the DNA sequence the SNP occurs. A SNP is a non-coding SNP, if the SNP is falling within intergenic regions or non-coding regions of genes. These SNPs are not translated to proteins, some of them might, however, have an influence on the level of gene expression, transcription factor binding or affect gene splicing. If a SNP is located within a region coding for a protein, it is called a coding SNP. Due to the redundancy of the genetic code, some of the nucleotides coding for amino acids can be exchanged without changing the amino acid that the triplet codes for, not all coding polymorphisms result in a change in the amino acid sequence of a protein. This type of SNPs, where both alleles result in the same final amino acid sequence, are called synonymous SNPs. Nonsynonymous SNPs are, on the other hand, SNPs where the polymorphism leads to a change in the resulting protein. Nonsynonymous SNPs can be divided into missense mutation, where translation results into a different amino acid, or nonsense mutation, which results in the introduction of a stop codon and truncation of the final protein.

The human genome is 3-billion bases long - every 100 to 300 bases a SNP occurs. This variation makes up 90% of all genetic variation found in humans [50]. SNP variations are correlated to diseases and functional variations, even allowing to assign phenotypic characteristics based on the genome sequence of an extinct ancient human [83].

## 1.2 The adaptive immune system

The immune system is a protection system against infectious disease, pathogens and tumor cells. It consists of two parts: The innate immune system as the first line of defense and the highly diverse, but slower, adaptive immune system. Innate immune responses are not specific to a particular pathogen and have no memory if encountering the same pathogen. In contrast, if the adaptive immune system encounters the same antigen again, the second response will be much more rapid and stronger than the primary response. Both systems cooperate with each other, but from the point of view of personalized medicine and transplantation medicine, the highly specialized adaptive immune system is a more interesting target than the innate immune system.

The adaptive immune system is highly specific. Its antigenic specificity allows antibodies to recognize subtle differences between proteins only differing by a single amino acid. It has further a high diversity in recognizing billions of different structures. Due to the immunologic memory, after an initial encounter, it offers a lifelong protections against some infectious diseases. Further, because of its self-nonself recognition, the adaptive immune system is normally capable of only reacting to foreign antigens.

Lymphocytes and antigen-presenting cells are the two major groups of cells involved in an adaptive immune response. Lymphocytes are white blood cells circulating in the lymphatic systems and in the blood. The two major lymphocytic cell types are B and T lymphocytes. The main role of the B lymphocytes, also called B cells, is the creation of antibodies for identifying and neutralizing foreign objects. There is a huge variation in the antigen binding site of different B cells, enabling the immune system to detect a vast amount of different antigens (for example pathogens). A B cell encountering an antigen matching its antibodies the first time causes the cell to divide rapidly. B cells differentiate into memory B cells plasma cells. Memory B cells have a long lifespan and enable the immune system to react faster if the host gets infected by the same antigen again. Further, accumulating their amount enables a strong immune response. Plasma cells produce antibodies with the same specificity as their parent B cells, but in a secretable form. These secreted antibodies bind to and inactivate antigens. In humans, secreted antibodies are the major effector of the immune system, per second up to thousands of antibodies can be secreted by a single plasma cell.

The other major group of cells, which is part of the adaptive immune system, is the T lymphocytes. In contrast to B cells, which are able to bind to free antigen, T lymphocyte receptors usually bind to antigen in complex

with a major histocompatibility complex (MHC) molecule. If a T lymphocyte encounters an antigen combined with MHC, the T lymphocyte proliferates into various effector T lymphocytes and memory T lymphocytes. One type of effector cells are cytotoxic T lymphocytes (CTL). This group of lymphocytes is known to induce death of infected somatic cells and tumor cells. Further, CTLs are capable of eliminating cells of a foreign tissue graft.

### Cytotoxic T cells

CTLs are a key player for the effector function of the adaptive immune system [3]. Due to their ability to destroy cells posing a threat to the organism, it is crucial that these cells are capable of distinguishing between a potential threat and harmless cells originating from self proteins. CTLs are also known as CD8+ T cells, since they express a CD8 co-receptor at the cell´s surface.

T cells are educated in the thymus to distinguish between self and non-self. T lymphocytes arise in the bone marrow and subsequently migrate to the thymus, an organ of the immune system, for maturation. The somatic rearrangement during this process leads to the expression of a unique T cell receptor (TCR) [3]. In 95% of all T cells. the TCR is composed by an $\alpha$ and a $\beta$ protein chain. Each chain is composed of different gene segments. Functional TCR genes are produced by rearranging variable (V) and joining (J) gene segments for the $\alpha$ chain and by rearranging V, J and an additional diversity (D) gene segments for the $\beta$ chain. The rearrangement of gene segments and a further addition of random nucleotides results in $10^{18}$ possible combinations and therefore unique TCRs. This diversity is the key for the detection and subsequent combating of pathogens.

This huge repertoire of potential T cells undergoes a selection process. The selection process consists of two parts: the *positive selection* and the *negative selection*. *Positive selection* ensures that a potential T cell is capable of binding to self-MHC molecules. T cell precursors have to interact with self-MHC molecules, cells that fail to bind are eliminated by apoptosis. *Positive selection* results in MHC restriction and ensures that only T cells capable of binding to self-MHC molecules survive.

The second selection process, *negative selection*, ensures that T cells are not binding too strongly with self-MHC or self-MHC in complex with self-peptides. *Negative selection* results in self-tolerance. This is crucial, as T cells should not induce cell death to the host´s cells. A partial failing of this mechanism is a potential cause for autoimmune diseases. A graphical representation is shown in Figure 1.3.

### Class I antigen processing

All proteins in eukaryotic cells are continuously degraded into peptide fragments and most of these peptides are further degraded into their constituent amino acids. A selection of these peptides, composed of 8-11 amino acids, escape complete destruction and are displayed on the cell (or ‘s surface by

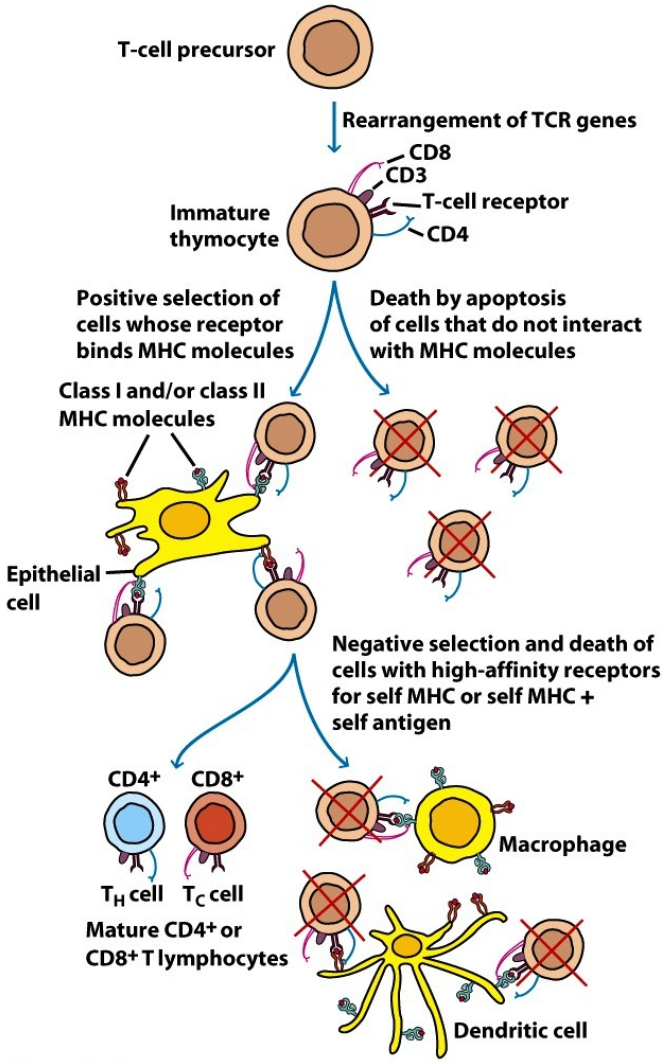


**Figure 1.3.** Positive and negative selection of potential T cells in the thymus. Positive selection results in MHC restriction; negative selection results in self-tolerance. From Kuby Immmunology [45].

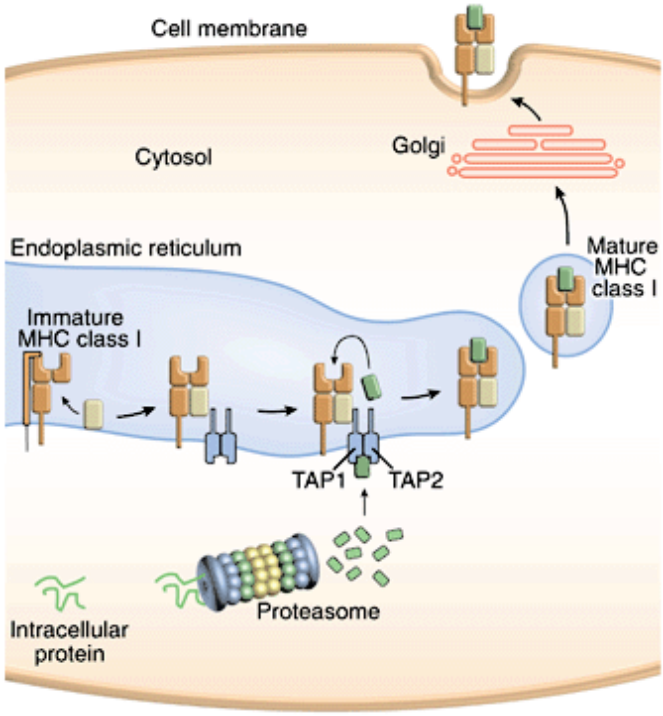


**Figure 1.4.** MHC class I antigen presentation pathway. Intracellular proteins are degraded by the proteasome into peptides. The peptides are transported into the ER by TAP. In the ER, an immature MHC class I complex binds to TAP; a stable peptide/MHC complex is formed with a suitable peptide. This complex is transported via the Golgi apparatus to the cell surface, where it is presented for interaction with T cells. From Andersen et al. [4]

MHC class I molecules [89]. By this mechanism, the cell is presenting its internal world to the outside. T cells are able of recognizing the presented complex and distinguish between self and foreign peptides.

There are three essential steps involved in the expression of a peptide/MHC class I complex at the cells surface: proteasomal cleavage of proteins, translocation by the transporter associated with antigen processing (TAP) molecule to the endoplasmatic reticulum (ER), and the assembly of the MHC with the antigenic peptide, following the transport of the the peptide/MHC complex to the cell surface [4]. The antigen processing pathway is shown in Figure 1.4.

The major protease for cutting proteins into peptides is the proteasome. Presentation of peptides on the cell surface is decreased by as much as 90% by proteasome inhibitors, whereas at the same time some specific peptides are shown to increase their surface expression. This indicates that other proteases are also involved in the degradation of proteins [56]. The proteasome

is required to generate the C-terminal but not the N-terminal ends of peptides presented in the context of MHC class I [63, 14]. N-terminal trimmed peptides are transported by the TAP molecule to the ER. TAP consists of two subunits, TAP-1 and TAP-2, forming a binding pocket. TAP preferentially binds to peptides of size 10-18 amino acids. These peptides are larger than peptides presented by MHC class I, additional trimming of the peptides occurs at a later stage in the ER. A model for predicting TAP affinity highlights that the C-terminal and the three outmost N-terminal amino acids are the key residues in defining binding affinities to TAP [78]. Once translocated to the ER, additional trimming of the peptides to a length of approximately 8-11 amino acids occurs. Within the ER, peptides are trimmed from the N-terminal side mainly; C-terminal trimming in the ER was shown to occur at a much lower frequency by several studies [103, 21]. This inefficiency of the ER to trim peptides at the N-terminus supports the idea of protease being the main workhorse for N-terminal trimming. At this step there are still a vast amount of possible peptides to choose from for binding to MHC class I. The binding of peptides to MHC molecules is the most stringent factor limiting presentation of possible epitopes on the cell surface. It is estimated that only 1 out of 200 potential peptides binds to a particular MHC class I molecule and that only half of these are immunogenic due to limitations in the T cell repertoire. Taking all steps of the antigen processing pathway into account, only 1 out of 2,000 possible epitopes is able to elicit a T cell response [129].

## 1.3 Hematopoietic cell transplantation

Hematopoietic stem-cell transplantation (HCT) is a standard treatment for a variety of malignant diseases and hematological malignancies [32]. It consists of an intravenous infusion of hematopoietic stem cells, with the goal of reestablishing marrow functions in patients with defective bone marrow or immune systems. In a report from 1939, an intravenous marrow infusion for treating aplastic anemia is described for the first time [74]. Over the years, with the discovery of human leukocyte antigen (HLA) and the later use of immunosuppressive drugs for minimizing Graft-versus-host disease (GVHD), HCT has become an effective treatment for different diseases, applied to thousands of patients every year.

Depending on the source of the graft, distinctions are made between two types of HCT: If the patients own marrow is used to reestablish hematopoietic function, it is called an autologous HCT. One of the benefits of this method is a low occurrence of GVHD, since the transplant comes from the patient himself. For some types of hematologic diseases, however, autologous HCT leads to lower survival than allogeneic HCT due to disease related mortality [8]. Allogeneic HCT is the other possible approach; it involves the transfer of marrow from another person, the donor, to a recipient. Patients undergo an immunosuppressive therapy and the patients immune system cells are replaced by the transplant [130]. The absence of malignant cells in the graft

and a possible Graft-versus-tumor (GVT) effect are the major advantages of allogeneic HCT. Main disadvantages are the risk of GVHD and difficulties in finding an appropriate donor [64].

### Hematologic malignant diseases

Hematological malignancies are cancers of the blood, the bone marrow or lymph nodes. The classification of hematological malignant diseases is based on their main occurrence: It is defined as leukemia if the malignancy is mainly located in blood, and lymphoma, if it is mainly affecting the lymph nodes. A more specific categorization including diagnostic criteria, associated genetic alterations and pathological features, is regularly published by the World Health Organization, currently in its $4^{\text{th}}$ edition [73]. A rough classification is further possible by the cell lineage. First, there are hematological malignancies derived from myeloid cell lines. Myeloid leukemias are further divided into acute (AML) and chronic (CML) myelogenous leukemia. Second, lymphoid leukemia and lymphoma is derived from lymphoid cell lines. Lymphoma is usually a solid tumor, whereas lymphoid leukemia is affecting lymphocytic cells in the blood.

### Donor selection

The selection of an appropriate donor is a major factor for the success of an HCT. Each HLA mismatch does not only lead to a difference of the specific HLA molecule, but also in the vast amount of peptides each HLA molecule is able to present to T cells at the cell surface, with the possibility of leading to a strong immune response.

While transplantation of graft from an HLA-matched sibling shows the best results, only 30%of the patients have the possibility for such a donor [64]. For finding an optimal donor, histocompatibility testing is done by high resolution typing to identify differences in nucleotides for the HLA-A, -B, -C, DRB1 and DQB1 alleles. There are 10 possible variations in a given patient, as humans have two homologous copies of each chromosome. If all alleles for a recipient and a donor are matching, they are defined as 10/10 matched. Accordingly, a single HLA locus disparity would be a 9/10 match and a multi-locus mismatch with two disparities would be a 8/10 match [120]. In the 1980s, national donor registries were started as a consequence of risen demand for unrelated donors. In an effort to enable international searches, *Bone Marrow Donors Worldwide* started to connect national registries and organizations [10]. Established in 1988, the database is now providing centralized access to almost 18 million donors. In recent years, availability of these databases and advances in HLA typing have greatly improved donor matching. A study including more than 11,000 patients reports a significant increase of 10/10 matched patient-donor pairs. From 1987-1998, only 28% of donor-patient pairs had no identified HLA mismatch, whereas this was increased to more than half from 1999-2002 and to 65% from 2003-2006 [40].

### Major Histocompatibility Complex

The major histocompatibility complex (MHC) is a gene family whose products are presenting intercellular products to the cell´s outside. All known mammalian species have an MHC complex. In humans they are called human leukocyte antigens (HLAs). In humans, the MHC is organized into three regions: Class I, II, and III. Class I type MHCs are present on the surface of nearly all cells. They are presenting peptides from the cell´s inside to the cell´s outside. Class II MHCs are only expressed by a subset of somatic cells. They are mainly found on B cells, macrophages and dendritic cells. Peptides presented by MHC class II are, in contrast to peptides presented by MHC class I, derived from extracellular proteins. MHC class III molecules do not share a similar function with MHC class I and II. They are located on chromosome 6 between the other MHC molecules and code for immune-related proteins.

### Minor Histocompatibility Antigens

Minor histocompatibility antigens (mHags) are a possible source for the rejection of MHC-matched transplants [88]. Even in a perfectly MHC matched allogeneic HCT, small variations in other proteins can cause the rejection of a grafted tissue. First found in mice, mHags were later recognized as being additional histocompatibility loci in human by rejection of skin grafts exchanged between HLA-identical siblings [11]. Later still, it was suggested that typing for some mHags prior to hematopoietic cell transplantation may identify patients at high risk for graft-versus-host disease and improve donor selection [30]. Minor histocompatibility antigens are peptides, derived from cellular proteins and presented at the cells surface, where they are recognized by MHC-restricted T lymphocytes and further raise an immune response [99]. It has been shown that both $CD8^+$ (class I restricted T cells) and $CD4^+$ (MHC class II restricted T cells) respond to mHag epitopes, albeit by different mechanisms [91].

### Graft-versus-host disease

Graft-versus-host disease (GVHD) is a complication after a hematopoietic cell transplantation, where healthy cells of the recipient are attacked. The recipient´s cells are seen as foreign and an immune response is mediated by the donor´s transplanted immune cells. An HLA mismatch between donor and recipient is a possible source for GVHD. In addition to an HLA mismatch, mHags may raise an attack by the immune system. A one amino acid difference in a protein presented by MHC can be enough to be perceived as foreign by the donor´s T cells and to trigger an immune response.

GVHD is divided into acute GVHD (aGVHD) and chronic GVHD (cGVHD). GVHD occuring within the first 100 days after HCT is called aGVHD, whereas the chronic form of GVHD normally occurs after 100 [29]. Tissues typically affected by aGVHD are liver, skin and the gastrointestinal

tract. By definition, each of these tissues and the overall grade of aGVHD are divided into grades from I-IV, where no treatment is required for grade I and grade IV is fatal [79]. Chronic GVHD is usually appearing at a later stage than aGVHD and involves more immune related cell types. It is further affecting a broader range of tissues. The classification system used for staging of chronic GVHD, originally proposed by the Seattle Group and based on 20 patients, differentiates between "limited" and "extensive" [54]. Several additional classification scales were developed allowing a finer grading of patients. However, the "limited/extensive" classification is still the most widely used.

### Graft-versus-tumor effect

The graft-versus-tumor (GVT) effect is a beneficial effect, based on the same principles that lead to GVHD. Immunological non-identity between recipient and donor, as induced by mHags, are responsible for GVHD, but they may also support tumor eradication [124]. The GVT effect was shown to reduce the risk of relapse for leukemia patients following an allogeneic transplant. Malignant target cells are recognized as foreign by the donor's immune cells and a response is initiated by the donor's CTLs and natural killer cells [85]. As the GVT effect is relying on the same principles as GVHD, one of the challenges of HCT is the prevention of undesirable GVHD without loosing the favorable GVT effect. Recent studies have shown that immunotherapy using donor lymphocytes can produce a GVT effect without leading to GVHD [47]. Several mHags exclusively expressed in hematopoietic tissues have been described [122]. Because of their hematopoietic cell-restricted cell damage, these mHags can be specifically used to eliminate a hematologic malignant disease, such as leukemia. These mHags are associated with a low risk of GVHD, as GVHD is targeting other organs such as skin or liver cells [65].

# Chapter 2

# MHC pathway epitope prediction

**Research Article**

## NetCTLpan: pan-specific MHC class I pathway epitope predictions

Thomas Stranzl[1], Mette V. Larsen[1], Claus Lundegaard[1], Morten Nielsen[1]

[1] Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark

## 2.1 Abstract

Reliable predictions of immunogenic peptides are essential in rational vaccine design and can minimize the experimental effort needed to identify epitopes. In this work, we describe a pan-specific major histocompatibility complex (MHC) class I epitope predictor, NetCTLpan. The method integrates predictions of proteasomal cleavage, transporter associated with antigen processing (TAP) transport efficiency, and MHC class I binding affinity into a MHC class I pathway likelihood score and is an improved and extended version of NetCTL. The NetCTLpan method performs predictions for all MHC class I molecules with known protein sequence and allows predictions for 8-, 9-, 10-, and 11-mer peptides. In order to meet the need for a low false positive rate, the method is optimized to achieve high specificity. The method was trained and validated on large datasets of experimentally identified MHC class I ligands and cytotoxic T lymphocyte (CTL) epitopes. It has been reported that MHC molecules are differentially dependent on TAP transport and proteasomal cleavage. Here, we did not find any consistent signs of such MHC dependencies, and the NetCTLpan method is implemented with fixed weights for proteasomal cleavage and TAP transport for all MHC molecules. The predictive performance of the NetCTLpan method was shown to outperform other state-of-the-art CTL epitope prediction methods. Our results further confirm the importance of using full-type human leukocyte antigen restriction information when identifying MHC class I epitopes. Using the NetCTLpan method, the experimental effort to identify 90% of new epitopes can be reduced by 15% and 40%, respectively, when compared to the NetMHCpan and NetCTL methods. The method and benchmark datasets are available at http://www.cbs.dtu.dk/services/NetCTLpan/.

## 2.2 Introduction

Cytotoxic T lymphocytes (CTLs) are a subgroup of T cells able to induce cell death of other cells. CTLs kill only infected or otherwise damaged cells. In

order to discriminate between infected and healthy cells, all nucleated cells present host cell peptide fragments on the cell surface in complex with major histocompatibility complex class I molecules (MHC class I). Not all possible peptides originating from cell proteins will be presented by MHC class I. In fact, it is estimated that only one out of 2,000 potential peptides will be immunodominant [129]. One of the first steps involved in MHC class I antigen presentation is the degradation of intracellular proteins, including proteins from the cytoplasm and nucleus, by the proteasome [52, 76, 14, 2, 63, 107, 38]. These peptides may be trimmed at the N-terminal end by cytosolic exopeptidases [55]. A subset of the peptides is transported by transporter associated with antigen processing (TAP) complex into the endoplasmatic reticulum (ER), where further N-terminal trimming occurs [87, 46, 116, 94]. Inside the ER, a peptide may bind to an MHC class I molecule and the peptide–MHC complex will be transported to the cell surface, where it subsequently may be recognized by CTLs. These successive steps from protein to ligand presented on the cell surface are limiting the number of possible epitopes. The most restricting step in antigen presentation is peptide binding to MHC class I molecule [129].

Reliable predictions of immunogenic peptides can minimize the experimental effort needed to identify epitopes. We have previously described a method, NetCTL [52, 53], integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions to an overall prediction of CTL epitopes. The NetCTL method has proven successful in identification of CTL epitopes from, for instance influenza [121], HIV [77], and Orthopoxvirus [110]. Several other groups have developed methods for CTL epitope identification by integrating steps of the MHC class I pathway (MAPPP, [31]; WAPP, [17]; EpiJen, [18]; MHC-pathway, [111]). All these methods are limited by the fact that they only allow for prediction of peptide binding to a highly limited set of different MHC molecules. In a large-scale benchmark evaluation of publicly available server of MHC class I pathway presentation prediction, Larsen et al. [52] showed that the NetCTL method significantly outperformed all these methods, closely followed by MHC-pathway. The MHC-pathway method has recently been updated to include more accurate predictions of MHC binding and a broader allelic coverage (close to 60 human leukocyte antigen (HLA)-A and HLA-B alleles are covered by the default MHC-pathway method in the 2009-09-01 release). In contrast to this, the NetCTL method has not been updated since 2007, and the MHC binding prediction remains limited to the 12 common HLA supertypes [57]. In the following, we describe an improved and extended version of NetCTL, called NetCTLpan, which is able to make predictions for all MHC class I molecules with known protein sequence. In addition, NetCTLpan can identify 8-, 9-, 10-, and 11-mer epitopes, as opposed to NetCTL, which only allowed for prediction of 9-mer epitopes. The method has been trained on a large data set of experimentally identified MHC ligands from the SYPFEITHI database [80]).

Choosing a performance measure for evaluating a prediction method is a nontrivial task, and the definition of performance measure will often influence

the benchmark outcome and subsequent choice of best method. A commonly used measure for predictive performance is the area under the receiver operating characteristic (ROC) curve, the AUC value. This measure integrates the sensitivity curve as a function of specificity for the range of sensitivity from one to zero. This measure might not be optimal if a prediction method is required to have a very high specificity in order to lower the false positive rate for subsequent experimental validation. In such situations, it could be beneficial to use only the high specificity part of the ROC curve to calculate the predictive performance. To match such requirements for a low false positive rate, we have therefore in this work focused on optimizing the method to achieve high specificity at a potential loss in sensitivity.

The predictive performance of the NetCTLpan method is validated on large and MHC diverse data sets derived from the SYFPEITHI [80] and Los Alamos HIV databases (http://www.hiv.lanl.gov/), and its performance has been compared to other state-of-the-art CTL epitope prediction methods.

It has been suggested that supertype-specific differences exist in how dependent MHC class I presentation of peptides is on transport via TAP molecules [9, 5, 34, 102] and proteasomal cleavage [126]. Likewise, it has been suggested that the rescaling procedure commonly used to correct for possible discrepancies between the allelic predictors [108, 53, 52] could mask genuine biological difference between MHC molecules and potentially lower the epitope predictive performance [60]. In the context of the NetCTLpan method, we investigate to what extend such differences are observed in large data sets that are diverse with regard to both MHC restriction and CTL epitopes.

## 2.3 Materials

### SYF data set

The SYFPEITHI database [80] was used as the source of MHC class I ligands. MHC class I binding peptides classified as ligands were downloaded in August 2009. Altogether, the database contained 2,966 HLA class I ligand pairs. Considering only ligands with length of 8 to 11 amino acids (the lengths for which the MHC class I binding predictor NetMHCpan can perform predictions), the data set consists of 2,752 unique HLA class I ligand pairs. Data used for training the individual MHC class I pathway predictors—MHC binding [68, 35], proteasomal cleavage [69], and TAP transport efficiency [78]—was removed from the data set, downsizing it to 2,309 unique HLA class I ligand pairs.

Peptides in the data set with only serotypic HLA assignment were assigned to the most common HLA allele in the European population for this serotype (e.g., the serotype HLA-A*01 was assigned to the specific allele HLA-A*0101). The HLA allele frequencies were obtained from the dbMHC database (http://www.ncbi.nlm.nih.gv/mhc/). Subsequently, for every peptide, the source protein was found in the UniProtKB/Swiss-Prot

database [13]. If more than one matching protein was a possible source for a peptide, the protein was selected with preference for human and long protein sequences. Peptides without corresponding source protein in UniProtKB/Swiss-Prot were searched against NCBI NR protein database (http://www.ncbi.nlm.nih.gov). These steps consequently resulted in the SYF data set consisting of 2,267 HLA class I ligand pairs with corresponding source proteins, where 226 ligands are 8-mers, 1,443 are 9-mers, 430 are 10-mers, and 168 ligands belong to the group of 11-mers. Note, that HLA-C ligands are included in these numbers. In the evaluation, HLA-C ligands are merged to a separate test set.

### HIV data set

The same HIV data set has been used as for the paper describing the original NetCTL method [52]. For comparison reasons, the data set has not been updated. The data set is derived from the Los Alamos HIV database (http://www.hiv.lanl.gov/). It consists of 216 HLA class I ligand pairs with corresponding source proteins covering the 12 supertypes [57].

### Training and test sets

Each of the HLA alleles in the SYF data set was assigned a supertype association using the distance measure described by Nielsen et al. [68]. In short, an HLA allele was associated to the most similar supertype defined in terms of the correlation coefficient between NetMHCpan prediction scores for 1,000,000 random natural 9-mer peptides for the HLA allele in question and any of the 12 supertype representatives [53]. In a few cases (less than ten), the supertype association was ambiguous. In these cases, the association was assigned by applying the classification from the work by [98]. The associated supertypes for each HLA class I allele are shown in Supplementary Table S1. Some supertypes in the 9-mer SYF data set contain more HLA class I ligand pairs than others. Only four out of the 12 supertypes had more than 100 HLA class I ligand pairs assigned. In order to minimize bias toward only a few supertypes, a training data set with maximum 50 randomly selected ligands per supertype was generated. For seven supertypes, it was possible to select 50 ligands for the training set, while the selection for the five remaining supertypes consisted of between 19 and 47 ligands. This results in a training set of 504 HLA class I ligand pairs. Remaining HLA-A and HLA-B ligands not included in the training data were assigned to a separate set used for evaluation. This evaluation set covers seven supertypes and consists of 889 9-mers. All HLA-A and HLA-B 8-, 10-, and 11-mer ligands were merged into another evaluation set, resulting in a total of 806 ligands. The HIV data set was used as a third independent evaluation set. The numbers of ligands per supertypes for the training and test sets are listed in Table 2.1. Finally, a set of 65 HLA-C ligands from the SYFPEITHI database of length 8–11 amino acids was used as a fourth evaluation set.

**Table 2.1.** Numbers of ligands per supertype in the training and test set.

| Supertype | Train | Test 9-mer | Test 8/9/10/11-mer | HIV |
|---|---|---|---|---|
| A1 | 36 | 0 | 29 | 5 |
| A2 | 50 | 208 | 94 | 82 |
| A3 | 50 | 49 | 75 | 41 |
| A24 | 19 | 0 | 5 | 9 |
| A26 | 50 | 43 | 74 | 2 |
| B7 | 50 | 8 | 57 | 32 |
| B8 | 28 | 0 | 19 | 5 |
| B62 | 47 | 0 | 27 | 10 |
| B27 | 50 | 224 | 141 | 3 |
| B39 | 50 | 21 | 36 | 1 |
| B44 | 50 | 336 | 227 | 16 |
| B58 | 24 | 0 | 22 | 10 |
| Total | 504 | 889 | 806 | 216 |

## 2.4 Methods

### MHC class I affinity prediction

The current version of the pan-specific MHC class I binding prediction method, NetMHCpan-2.2 [35], is an updated version of the original NetMHCpan method [68]. It has been evaluated as the best pan-specific method in large benchmark study [132] and is now including the extension to perform predictions for 8-, 10-, and 11-mer peptides [59]. NetMHCpan-2.2 was trained on a data set of 102,146 quantitative peptide–MHC affinity data points covering more than 100 distinct MHC molecules. The prediction server is available at http://www.cbs.dtu.dk/services/NetMHCpan-2.2/.

### TAP transport efficiency prediction

The prediction of TAP transport efficiency is based on the matrix method described in Peters et al. [78]. The method predicts TAP transport efficiency of peptides by a scoring method using only the C terminus and the tree N-terminal residues of a peptide. The contribution to the prediction score of the N-terminal residues is down-weighted by a factor of 0.2 in comparison with the score of the C terminus. In the original publication, the TAP transport efficiency score was computed as the average of the values for the 9-mer and its 10-meric precursor. Here, we extend this approach and predict the TAP transport efficiency score for peptides of length from 8 to 11 amino acids, as the average of the values for the original peptide and its precursor extended by one amino acid N-terminally. The matrix published in Peters et al. [78] was modified as all values in the TAP scoring matrix were multiplied by

a factor of −1, in order to have a high predicted value corresponding to high transport efficiency. This way the interpretation is consistent with the prediction of proteasomal cleavage and MHC class I binding affinity.

### Proteasomal cleavage prediction

NetChop C-term 3.0 [69] was used for predicting cleavage sites. As in the original NetCTL publication, only the C-terminal cleavage score of a peptide was included.

### Combined class I pathway presentation prediction—NetCTLpan

The NetCTLpan prediction value is defined as a weighted sum of the three individual prediction values for MHC class I affinity, TAP transport efficiency, and C-terminal proteasomal cleavage. Optimal relative weights on TAP transport efficiency and proteasomal cleavage were estimated using the training data set and based on the average AUC value per HLA class I ligand pair.

The AUC measure is a commonly used measure for quantitative tests and model comparison. AUC is the area under the ROC curve, summarizing the sensitivity as a function of 1—the specificity. The specificity is given as 1—the false positive ratio defined as the fraction of the number of correctly predicted nonligands relative to the total number of nonligands in the dataset [58]. A specificity of 100% is interpreted as all nonligands are actually classified as nonligands. The sensitivity is the true positive rate (TPR) and is defined as the number of correctly predicted ligands relative to the total number of ligands in the dataset. The higher the TPR, the more actual positives are recognized. The AUC measure might not be optimal if a prediction method is required to have very high specificity in order to lower the false positive rate in subsequent experimental validations. In such situations, it is beneficial to use only the high specificity part of the ROC curve to calculate the predictive performance. Therefore, a search optimizing the AUC value integrated for specificities from 1 to x (AUCx), where x [0:1] was performed to optimize the method to achieve high specificity. High values of x will focus the method toward high specificity at a potential loss in sensitivity, whereas low values of x will result in equal focus on sensitivity and specificity.

When calculating the AUC value, the source protein was divided into overlapping peptides of the size of the given ligand. All peptides, except those annotated as ligands in either the complete SYFPEITHI or Los Alamos HIV databases, were taken as negative peptides (nonligands) and the given ligand was taken as positive. A perfect AUC value of 1.0 corresponds to the ligand having the highest combined score (NetCTLpan score) compared to all other possible peptides originating from the source protein.

Another important issue to resolve is how to calculate AUC values. Should it have been done per protein, where an AUC value is calculated for each ligand–HLA–protein triplet and the performance reported as the average AUC

value over all triplets or should it have been made in a pooled way, where all peptide data for the different source proteins and HLA alleles are merged together before calculating the AUC value? Here, we suggest using the per-protein measure, since pooling data from different proteins and HLA alleles will place ligands in a nonbiological competition for presentation. The source proteins in the SYF ligand data sets have a length distribution varying from 36 to more than 8,000 amino acids. Applying the NetCTLpan method to our training set (most homogenous data set) shows a tendency for shorter proteins having a lower $AUC_{0.1}$ than longer proteins. Proteins from our training set with length of 0–200 have a mean $AUC_{0.1}$ of 0.817, whereas proteins longer than 200 AA have a mean $AUC_{0.1}$ of 0.876. The Spearman's rank correlation between the protein length and $AUC_{0.1}$ values for the training data set is 0.15. This value is significantly different from random ($p < 0.001$, exact permutation test). In a pooled evaluation, where source protein data are merged, the predictive performance would predominantly reflect the performance for the longer protein. Further, not all proteins are expressed in equal amounts within the cell and the presentation of peptides in complex with HLA molecules happens in competition with the four most different HLA-A and HLA-B molecules within a given host and not 46, as it would be the case, when all the HLA alleles from the SYF training data set are pooled. Finally, it is becoming apparent that not all MHC molecules present peptides at the same binding threshold [81]. This observation would make an evaluation, where data for different HLA alleles is pooled, highly problematic, as illustrated in Fig. 2.1. Here, a ROC curve is shown for a pooled set of 29 HLA-A*0101, 50 HLA-B*4402, and 31 HLA-B*5101 ligands using the NetCTLpan method. In addition, the allele-specific sensitivity (fraction of ligands identified) for each allele is shown as a function of the pooled specificity. The figure clearly demonstrates that different alleles dominate the ROC curve in different specificity ranges. At a specificity of 0.0025, for instance, 60% (66) of the 110 ligands are identified. Of these are 25 (86% of 29) HLA-A*0101, 32 (62% of 50) are HLA-B*4401, and only nine (29% of 31) are HLA-B*5101 restricted. At very high specificities, the ROC curve is thus predominantly shaped by the HLA-A*0101 data, at intermediate specificities values the curve is shaped by the HLA-B*4402 data, and finally at low specificity values, the HLA-B*5101 data defines the curve. This is clearly not an optimal way of evaluating an overall predictive performance of a prediction method that is aimed at achieving uniform prediction accuracy across a broad range of HLA alleles. To conclude, we find that the proposed triplet evaluation per ligand–HLA–protein evaluation constitutes the least biased approach to evaluate a prediction method with broad allelic coverage.

## 2.5 Results

### The NetCTLpan method

The optimal weights on proteasomal cleavage and TAP transport efficiency were calculated for AUC fractions (AUCx) varying x from 0.05 to 1, with a

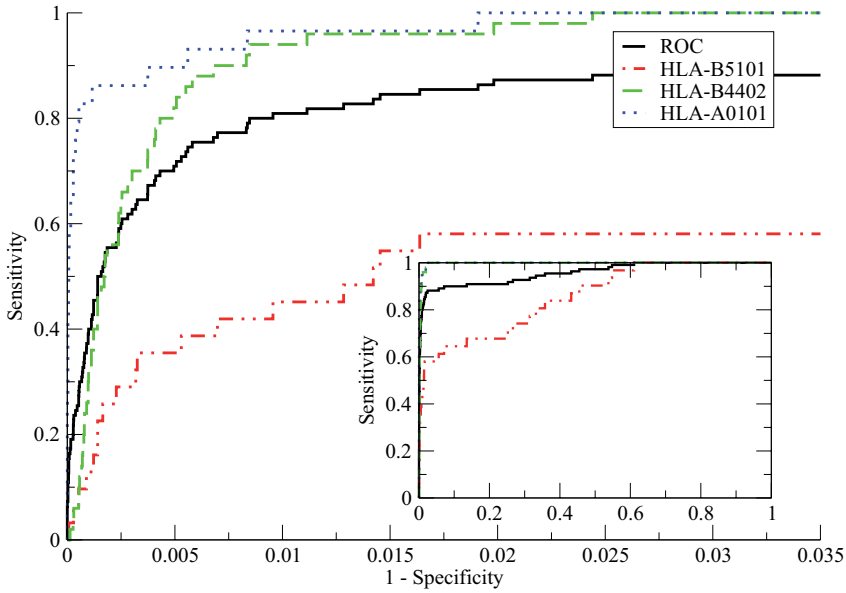


**Figure 2.1.** ROC curves for a pooled data set from the HLA-A*0101, HLA-B*4402 and HLA-B*5101 alleles. The source proteins for all three alleles were cut into overlapping peptides of the size of the given ligand and all peptides except the given ligands were taken as negative. The data set contained 31 HLA-A*0101, 50 HLA-B*4402, and 29 HLA-B*5101 ligands and the predictions were made using the NetCTLpan method. The black curve shows the ROC curve for the combined data set. The other three curves show the allele-specific sensitivity (fraction of ligands identified) as a function of the overall specificity for each of the three alleles. The insert shows the curves for the full range of specificities.

step size of 0.05. With x equal to 1, this corresponds to the conventional AUC value calculation and the way of selecting optimal weights for the original NetCTL method. The result of this analysis is shown in Fig. 2.2. For an AUC fraction of 1, the optimal weights were zero on both proteasomal cleavage and TAP transport. This implies that NetMHCpan 2.2, the method used for predicting MHC class I binding affinity, has a very high performance and that adding predictions for proteasomal cleavage or TAP transport decreased the overall performance. Figure 2.2 illustrates that the more the method is focused on high specificity (low values of x), the higher the weights and thus importance of proteasomal cleavage and TAP transport predictions become. This is, however, achieved at a loss in sensitivity at low specificity values. Based on this observation, the best performing weights on proteasomal cleavage and TAP transport were selected using an AUC fraction of 0.1 as benchmark measure and were found to be 0.225 for cleavage and 0.025 for TAP. This selection of weights defines the NetCTLpan method. When

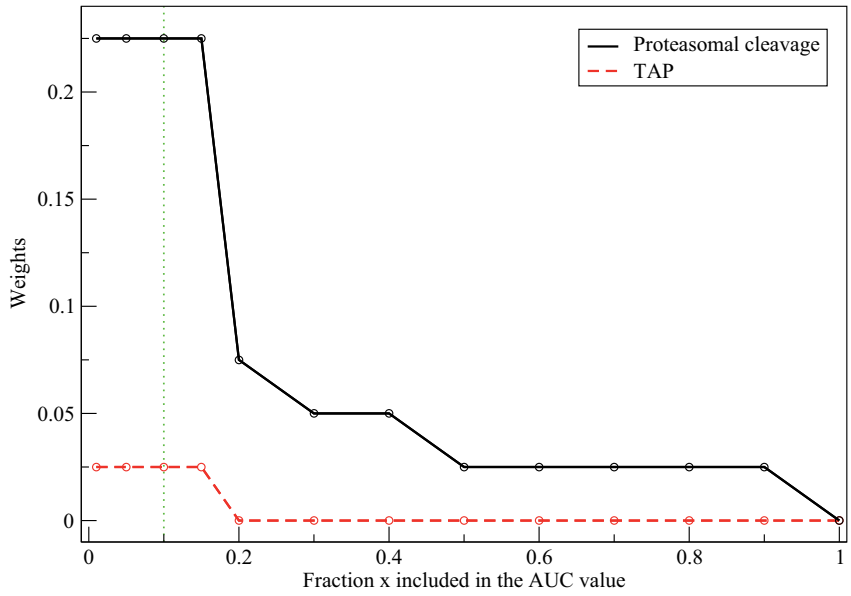


**Figure 2.2.** Weights on proteasomal cleavage and TAP transport efficiency related to AUCx fraction. The smaller the included fraction, the higher the contribution of proteasomal cleavage and TAP transport efficiency to a high performance. Optimal weights on proteasomal cleavage and TAP were found by optimizing the average AUCx value on the SYF training data set. The dotted line indicates the $AUC_{0.1}$ fraction.

interpreting the weights for cleavage and TAP, keep in mind that the contribution of the different prediction methods is not directly reflecting their relative biological contribution in the pathway.

A comparison of the ROC curves for NetMHCpan and our described method NetCTLpan is shown in Fig. 2.2. The overall AUC value for the NetMHCpan method is 0.980 and the corresponding $AUC_{0.1}$ value is 0.852. For the NetCTLpan method, the overall AUC value is 0.976 and the corresponding $AUC_{0.1}$ value is 0.869. These numbers and the graphs in Fig. 2.2 illustrate the improved specificity of the NetCTLpan method compared to NetMHCpan. Up to a specificity of 0.85, the ROC curve for NetCTLpan has a higher sensitivity than NetMHCpan, indicating that this method will identify more true ligands at a given specificity threshold. On the other hand, below a specificity of 0.85, the two ROC curves cross and the NetMHCpan method achieves the highest sensitivity. This crossover, however, happens at a very low specificity corresponding to a false positive rate of 0.15 (15% of the negative peptides are falsely classified as positive) and is of limited use when doing actual epitope discovery work, underlining the importance of optimizing the methods on high specificity.

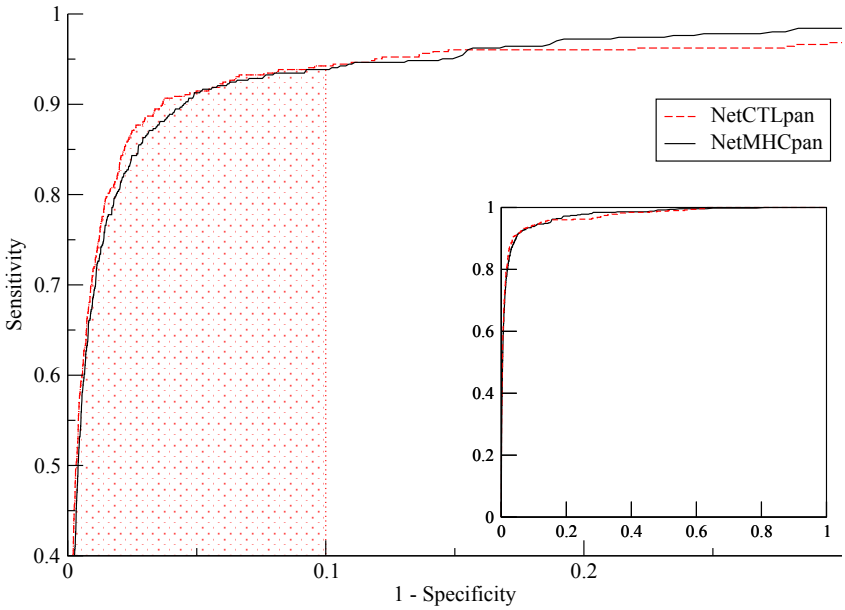


**Figure 2.3.** Performance comparison in terms of ROC curves for NetCTLpan and NetMHCpan. The true positive rate is shown as a function of the false positive rate. The figure is based on the SYF training set. The shaded area shows the area under the curve used to calculate the $AUC_{0.1}$. The insert shows the complete curves.

Table 2.2 displays the comparison between NetCTLpan and NetMHCpan for the different data sets using both the overall AUC and $AUC_{0.1}$ benchmark measures. Using the $AUC_{0.1}$ measure, the NetCTLpan method has a significantly higher performance compared to NetMHCpan for all data sets. On the other hand, when comparing the overall AUC value, the two methods show comparable performance. Here, for the SYF data set, the NetMHCpan method has the highest performance, while for the HIV data set and the HLA-C test set, NetCTLpan performs best. So, if high sensitivity is essential (even at a cost in specificity), the NetMHCpan method should be preferred. In more common situations, where specificity is the more important issue, NetCTLpan should be the choice.

Results displayed in Table 2.2 are mean AUC and $AUC_{0.1}$ values over all ligand–HLA–protein triplets in each data set. Paired tests were used for comparing performance between different prediction methods. In Supplementary Table S2 are given the AUC and $AUC_{0.1}$ values for each ligand–HLA–protein triple in the SYFPEITHI data sets. From this table, it is clear that the predictive performance does not only vary between supertypes, but also within supertypes. For the training data set, the difference between HLA-B*5101 and HLA-B*0702 (both B7 supertype alleles) for the NetCTLpan method is

**Table 2.2.** AUC and fractional AUC value comparison between NetCTLpan and NetMHCpan. The performance values are calculated as average per protein AUC values over the corresponding data sets. P-values are calculated by a paired t-test excluding ties. The best performing method is, for each data set and performance measure, high-lighted in bold.

| Data | Measure | NetCTLpan | NetMHCpan | P-value |
|---|---|---|---|---|
| Train (9) | AUC | 0.976 | **0.980** | 0.056 |
| | $AUC_{0.1}$ | **0.869** | 0.852 | 0.002 |
| Test (8/9/10/11) | AUC | 0.977 | **0.979** | 0.273 |
| | $AUC_{0.1}$ | **0.863** | 0.855 | 0.002 |
| Test (HIV) | AUC | **0.933** | 0.920 | 0.028 |
| | $AUC_{0.1}$ | **0.612** | 0.593 | 0.106 |
| Test (HLA-C) | AUC | **0.920** | 0.866 | <0.001 |
| | $AUC_{0.1}$ | **0.495** | 0.307 | <0.001 |
| Total | 504 | 889 | 806 | 216 |

thus 0.374 in terms of the $AUC_{0.1}$ measure. These performance variations demonstrate the need for large-scale HLA diverse benchmark data set to evaluate differences in performance between prediction methods, as the performance difference between similar (supertype-wise) alleles often is as high as the difference for individual alleles between two prediction methods within a given data set.

### Data redundancy

Several ligands appear in the SYFPEITHI ligand data sets as duplicates restricted to multiple HLA class I alleles. One might be worried that the potential peptide similarity/redundancy could influence the performance estimates of the NetCTLpan method. The training data set, for instance, consists of 504 HLA ligand pairs, but only 492 of these are unique peptides. The 9-mer test set consists of 889 9mer HLA ligand pairs, of which 802 are unique peptides. The training and 9-mer test sets share 42 identical ligands and three ligands with one mismatch, all coupled to different alleles. The training set contains four ligands identical with one mismatch. To investigate the impact on this data redundancy within the training data set and between the training and test data sets, we calculated the performance on redundancy-reduced data sets. The performance on the training set was calculated by removing duplicates and ligands with one mismatch and for the test set by excluding duplicates and ligands with one mismatch to ligands in the training data. Predictive performance was shown to be close to identical for both training and test set, suggesting that peptide redundancy plays a negligible role in our performance evaluation (see data in Supplementary Table S3).

### MHC affinity rescaling

In contrast to the NetCTL method, the NetCTLpan method does not use rescaling of predicted MHC class I affinities. Previously, rescaling has been used to make prediction values comparable between MHC class I molecules. It has been suggested that such a rescaling might remove genuine biological differences between MHC molecules and potentially lowers the epitopes predictive performance [60]. To investigate, if the predictive performance of the NetCTLpan method is influenced when including rescaling, we defined a rescaling factor for each MHC allele and used that factor to rescale all MHC binding affinity values before integrating with proteasomal cleavage and TAP scores. For each allele, the rescaling factor was determined as the 1 percentile score of the NetMHCpan method for a set of 1,000,000 random natural 9-mer peptides. An overall performance gain using rescaling as compared to not applying rescaling was observed if focusing on the overall AUC value (no rescaling AUC 0.976 versus rescaling AUC 0.978, p value 0.006, paired t test). For high specificity predictions ($AUC_{0.1}$), however, the method without rescaling performed similar ($AUC_{0.1}$ 0.869) to the method using rescaling ($AUC_{0.1}$ 0.868) with a p value of 0.835. From these results, and to maintain potential biological differences in specificity between MHC molecules, we chose not to include rescaling in the NetCTLpan method. One might argue that rescaling versus nonrescaling cannot influence the performance of the NetCTLpan method, when the performance is calculated per ligand–HLA allele, as it is the case in this study. When focusing on MHC binding predictions alone, this is true and both methods give identical results. However, when integrated with proteasomal cleavage and TAP transport efficiency, this situation changes. Rescaling places all MHC binding predictions on a similar scale and hence also places the relative weights on TAP and proteasomal cleavage on a similar scale across the set of MHC alleles. This is no longer the case if rescaling is left out. Here, alleles with low (predicted) binding affinity preference will have higher relative weights on TAP and proteasomal cleavage as compared to alleles with high binding affinity preference.

### Supertype-specific weights on proteasomal cleavage and TAP scores

As mentioned earlier, previous work has suggested that different MHC molecules have different dependencies on TAP transport efficiency and proteasomal cleavage. Based on these observations, it seems natural to find allele-specific weights for TAP transport and proteasomal cleavage. Due to the small size of the training data set, we limited ourselves to a search for supertype-specific weights. For each supertype, we estimated the weights on proteasomal cleavage and TAP transport that give optimal average $AUC_{0.1}$ values. Optimal weights per supertype and performance values for the different data sets can be seen in Table 2.3. It shows that relative large differences

exist between the optimal weights across the different supertypes. Naturally, the average $AUC_{0.1}$ for the training set is higher with supertype-specific weights as compared to the fixed weights (estimate for the complete training data set). Applying these weights resulted in an inconsistent pattern in performance gain across the different supertypes for the different test sets when compared to fixed weights. Only three supertypes (A24, B8, and B58) showed a consistent performance gain for the SYFPHITHI and HIV test sets using supertype-specific weights. This result strongly indicates that optimal weights per supertype are not reflecting biological differences but occur most likely due to overfitting. Note that we are not stating that proteasomal cleavage and TAP transport dependency could not vary between MHC molecules; we only state that based on our data, we cannot consistently reproduce such a differentiated dependency.

### Comparison to NetCTL

The comparison of the performance between NetCTLpan and NetCTL is based on the 9-mer data sets, since NetCTL is only capable of predicting 9-meric epitopes. Table 2.4 shows the performance for NetCTLpan and NetCTL on the different data sets. For both SYF data sets, the NetCTLpan method significantly outperforms NetCTL. The HIV test set does not show NetCTLpan being significantly better than NetCTL. The HIV test set is supertype based, and the HLA restriction for each HIV epitope is assigned to the corresponding HLA supertype. This is in contrast to the SYF ligand data sets, where full typing HLA restriction is available for most ligands. One hundred nineteen out of 216 HIV peptide supertype pairs are, however, annotated in the Los Alamos HIV database with full typing for the HLA restriction. Using this additional information about the HLA restriction improves the mean $AUC_{0.1}$ from 0.612 to 0.745 and the overall AUC from 0.933 to 0.959. Both measurements thus testify NetCTLpan as having a significantly better performance (both p values $<0.001$, paired t test) compared to NetCTL. These results clearly confirm earlier findings [77, 35] of the importance of going beyond HLA supertypes and the use of full-type HLA restriction information when identifying MHC class I epitopes.

To determine the source of the strong gain in predictive performance between the NetCTL and NetCTLpan methods, we compared the predictive performance of the NetCTLpan method to that of NetCTL using the supertype representative for each HLA allele also for the NetCTLpan method. This analysis clearly shows (see Table 2.5) that the shift from supertype to allele-specific predictions is the main driving force behind the gain in predictive performance between NetCTL and NetCTLpan. In all benchmarks has the NetCTLpan_ST (supertype-specific NetCTLpan method) a similar predictive performance to that of NetCTL.

**Table 2.3.** Supertype-specific weights benchmark. Optimal weights per supertype are shown. Performance is given as the average $AUC_{0.1}$ value for each data set. Fixed weights for proteasomal cleavage and TAP transport efficiency are 0.225 and 0.025, respectively. The higher $AUC_{0.1}$ value is highlighted in bold for each data set and supertype.

| **Supertype** | **Weights** | | **Train** | | | **Test (8/9/10/11)** | | | **Test (HIV)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Cleavage** | **TAP** | **Fixed** | **Specific** | **P-value** | **Fixed** | **Specific** | **P-value** | **Fixed** | **Specific** | **P-value** |
| A1 | 0.050 | 0.075 | 0.942 | 0.950 | 0.294 | 0.937 | 0.936 | 0.326 | 0.381 | 0.455 | 0.610 |
| A2 | 0.550 | 0.000 | 0.808 | 0.822 | 0.133 | 0.776 | 0.758 | 0.008 | 0.681 | 0.657 | 0.104 |
| A3 | 0.225 | 0.025 | 0.890 | 0.890 | 0.598 | 0.872 | 0.872 | (*) | 0.648 | 0.648 | (*) |
| A24 | 0.000 | 0.000 | 0.917 | 0.942 | 0.257 | 0.783 | 0.895 | 0.389 | 0.636 | 0.636 | 0.960 |
| A26 | 0.275 | 0.025 | 0.885 | 0.885 | 0.476 | 0.873 | 0.864 | 0.006 | 0.761 | 0.771 | 0.500 |
| B7 | 0.000 | 0.000 | 0.710 | 0.736 | 0.378 | 0.765 | 0.765 | 0.998 | 0.493 | 0.437 | 0.064 |
| B8 | 0.725 | 0.000 | 0.916 | 0.920 | 0.231 | 0.858 | 0.870 | 0.517 | 0.132 | 0.144 | 0.374 |
| B62 | 0.275 | 0.200 | 0.889 | 0.902 | 0.014 | 0.751 | 0.727 | 0.345 | 0.440 | 0.496 | 0.303 |
| B27 | 0.475 | 0.025 | 0.911 | 0.921 | 0.014 | 0.921 | 0.902 | 0.001 | 0.370 | 0.299 | 0.390 |
| B39 | 0.175 | 0.025 | 0.859 | 0.860 | 0.896 | 0.853 | 0.849 | 0.362 | 0.739 | 0.711 | (**) |
| B44 | 0.100 | 0.025 | 0.859 | 0.868 | 0.127 | 0.885 | 0.896 | 0.001 | 0.636 | 0.631 | 0.845 |
| B58 | 0.025 | 0.025 | 0.959 | 0.963 | 0.399 | 0.820 | 0.887 | 0.161 | 0.774 | 0.882 | 0.128 |
| All | 0.225 | 0.025 | 0.869 | 0.878 | 0.016 | 0.863 | 0.860 | 0.143 | 0.612 | 0.603 | 0.300 |

**Table 2.4.** Benchmark comparison of the NetCTLpan and the NetCTL methods. Average AUC and $AUC_{0.1}$ values for the NetCTLpan and NetCTL methods calculated for the SYF train set and the SYF and HIV test sets. For each data set and performance measure, the best performing method is shown in bold. P-values are calculated by a paired t-test excluding ties.

| Data | Measure | NetCTLpan | NetCTL | P-value |
|---|---|---|---|---|
| Train (9) | AUC | **0.976** | 0.971 | 0.018 |
| | $AUC_{0.1}$ | **0.869** | 0.816 | <0.001 |
| Test (9) | AUC | **0.982** | 0.975 | <0.001 |
| | $AUC_{0.1}$ | **0.877** | 0.802 | <0.001 |
| Test (HIV) | AUC | 0.933 | **0.936** | (*) 0.366 |
| | $AUC_{0.1}$ | **0.612** | 0.606 | 0.600 |

**Table 2.5.** Benchmark comparison of NetCTL, NetCTLpan and NetMHCpan_ST (supertype-specific version of NetCTLpan). The performance values are calculated as average per protein AUC values for the training and test data sets.

| Data | Measure | NetCTL | NetCTLpan | NetCTLpan ST |
|---|---|---|---|---|
| Train (9) | AUC | 0.971 | 0.976 | 0.971 |
| | $AUC_{0.1}$ | 0.816 | 0.869 | 0.830 |
| Test (9) | AUC | 0.975 | 0.982 | 0.971 |
| | $AUC_{0.1}$ | 0.802 | 0.877 | 0.805 |
| Test (8/10/11) | AUC | NA | 0.972 | 0.961 |
| | $AUC_{0.1}$ | NA | 0.848 | 0.770 |

### Comparison to state-of-the-art MHC class I pathway prediction methods

Next, we compared the performance of the NetCTLpan method to the MHC-pathway method [111]. This method has earlier been shown to be a state-of-the-art MHC class I pathway predictor [53]. Like the NetCTLpan method, this method integrates predictions of MHC binding, C-terminal proteasomal cleavage, and TAP transport into a combined pathway presentation score. Here, we use the method with default parameters via the link http://tools.immuneepitope.org/analyze/html/mhc_processing.html. The MHC-pathway method is not pan-specific and hence does not allow predictions for all HLA class I alleles used in our benchmark data. Further, it does not allow for predictions of 8- and 11-mer epitopes and only allows 10-mer epitope predictions for a subset of the included alleles. To allow for a fair comparison, we therefore only included ligands from the SYF data set restricted to HLA alleles covered by the MHC-pathway method. The results

of the benchmark calculation are shown in Table 2.6 and clearly show that NetCTLpan outperforms the MHC-pathway method for all three data sets. The improved performance is maintained for both the AUC and $AUC_{0.1}$ measure. Further, the table shows that the MHC binding predictors for the two methods have close to identical performance (NetMHCpan versus MHC). The cleavage method employed by the NetCTLpan method is performing consistently better than the immunoproteasome prediction method used by MHC-pathway (NetChop versus Immu). The TAP prediction method is identical between the two methods. These results suggest that the integration method employed by MHC-pathway is not optimal either due to the relative low performance of the immunoproteasome predictor or as a consequence of how the three prediction scores have been integrated in the MHC-pathway method.

## 2.6 Discussion

Earlier work has demonstrated the benefit of integrating proteasomal cleavage, TAP transport efficiency, and MHC binding predictions when using reverse immunology to identify potential CTL epitopes. However, to the best of our knowledge, none of the publicly available methods providing this integration are pan-specific and hence do not allow for prediction of CTL epitopes restricted to any MHC allele.

Here, we have developed a pan-specific MHC class I epitope predictor, NetCTLpan. The method integrates prediction of proteasomal cleavage, TAP transport efficiency, and MHC binding into a MHC class I pathway presentation likelihood score. In large-scale benchmarks comprising more than 1,000 MHC class I ligands and CTL epitopes restricted by close to 60 different HLA alleles, the method was shown to outperform both the original NetCTL method, as well as MHC-pathway, another state-of-the-art class I presentation pathway prediction method.

NetCTLpan was optimized to achieve high specificity in order to meet the need for a low false positive rate when using the method for large-scale epitope discovery. If focusing on optimal sensitivity, it was shown that the optimal prediction method should exclude both cleavage and TAP predictions reducing the method to MHC binding prediction alone. This is in contrast to earlier work, where proteasomal cleavage and TAP transport efficiency consistently have been reported to improve the predictive performance. Whether this observation reflects true biological aspects of the specificity overlap between the three pathway players (see for instance Nielsen et al. [69]) or it simply occurs because the prediction of MHC class I affinity has gained accuracy during the recent years, whereas predictors for TAP transport efficiency and proteasomal cleavage have not changed or been updated, remains to be seen.

Recent publications have suggested that some MHC molecules are, compared to others, more or less dependent on TAP transport and proteasomal

**Table 2.6.** Benchmark comparison of the NetCTLpan and MHC-pathway methods. The performance values are calculated as average per protein AUC values for the training and test data sets. The benchmark is made on the subset of the SYF ligand data sets covered by the MHC-pathway method. [1]MHC prediction score from MHC-pathway method. [2]Immuno-Proteasomal cleavage score from MHC-pathway predictions. The TAP prediction method is identical between the two methods. P-value for the comparison of NetCTLpan to MHC-pathway are calculated by a paired t-test excluding ties.

| **Data** | **Measure** | **NetCTLpan** | **MHC-pathway** | **p-value** | **NetMHCpan** | **MHC[1]** | **TAP** | **NetChop** | **Immu[2]** | **N** |
|---|---|---|---|---|---|---|---|---|---|---|
| Train (9) | AUC | 0.978 | 0.972 | <0.001 | 0.983 | 0.981 | 0.839 | 0.881 | 0.803 | 438 |
| | $AUC_{0.1}$ | 0.874 | 0.854 | 0.01 | 0.858 | 0.862 | 0.278 | 0.360 | 0.260 | 438 |
| Test (9) | AUC | 0.978 | 0.974 | <0.001 | 0.978 | 0.977 | 0.809 | 0.870 | 0.774 | 615 |
| | $AUC_{0.1}$ | 0.871 | 0.847 | <0.001 | 0.864 | 0.870 | 0.204 | 0.362 | 0.215 | 615 |
| Test (10) | AUC | 0.966 | 0.957 | <0.005 | 0.964 | 0.966 | 0.810 | 0.817 | 0.734 | 291 |
| | $AUC_{0.1}$ | 0.842 | 0.800 | <0.005 | 0.835 | 0.824 | 0.272 | 0.238 | 0.180 | 291 |

cleavage. Using the NetCTLpan method in large-scale benchmarks, we however find no consistent signal of such an HLA allele differentiated dependency of proteasomal cleavage and TAP transport efficiency. A performance gain using supertype-specific weights could only be observed for the training set. Applying these weights to the test sets resulted in an inconsistent pattern in performance gain for the different supertypes when compared to fixed weights, indicating that optimal weights per supertype are not reflecting biological differences but most likely are a result of overfitting.

NetCTL, the ancestor of NetCTLpan, uses a rescaling of MHC binding affinity values to make prediction values comparable between MHC class I molecules. It has been suggested that such a rescaling might remove genuine biological differences between MHC molecules and potentially lower the method's predictive performance. Here, we show that rescaling has no significant impact on the overall predictive performance of the NetCTLpan method. Further, we observed a tendency of different MHC molecules presenting ligands at different (predicted) binding thresholds. Based on these observations, the NetCTLpan method is implemented without use of rescaling, thus maintaining potential genuine biological differences between MHC molecules. To allow comparison between presentation likelihood scores for different MHC molecules, we include a rank-score for each prediction. The rank-score is calculated as the percent rank of a given NetCTLpan likelihood score to a set of 200,000 random natural 9-mer peptides.

Our results on the HIV benchmark data set confirm the importance of going beyond HLA supertypes and use full-type HLA restriction information when identifying MHC class I epitopes. In this benchmark, we found a significantly improved predictive performance, if full HLA restriction were used, in comparison to the HLA supertype information proposed in the original NetCTL publication.

In contrast to earlier published methods for MHC class I pathway prediction, NetCTLpan allows for predictions of 8- to 11-mer CTL epitopes being presented by any MHC class I molecule of known protein sequence.

NetCTLpan, the method described in this work, has shown to perform best when focusing on high specificity predictions for CTL epitope identification. In order to easily grasp the predictive performance gain, we applied the rank measure as defined by Larsen et al. [53]. The rank measure reports the average fraction of epitopes identified as a function of the percentage rank (percentage of tested peptides) for a set of proteins. This measure indicates how large a fraction of the peptides for a given protein needs to be tested in order to identify the epitope with a given likelihood. To identify new epitopes with 90likelihood by use of NetCTLpan, the rank measure reports that 3.7peptides need to be experimentally verified. For a hypothetical protein of 300 peptides, this means that on average, 11 peptides need to be tested in order to identify the epitope. The corresponding numbers for NetMHCpan and NetCTL are 13 and 17 peptides. Hence, by applying the NetCTLpan method instead of NetMHCpan, the experimental effort can be reduced by 17approximately 40that utilizing the NetCTLpan method can minimize experimental

effort needed to identify new CTL epitopes. We believe that this improved performance, combined with the methods ability to provide predictions of potential CTL epitopes of length from 8 to 11 amino acids to any MHC class I molecules of known sequence, will be useful in both rational reverse immunogenetic epitope discovery and interpretation of observed immune responses in HLA diverse patient cohorts. The NetCTLpan method and benchmark data set are available at: http://www.cbs.dtu.dk/services/NetCTLpan.

## Acknowledgments

This work is supported by a grant from the Danish Research Council for Technology and Production Sciences (project title "Disease Gene Finding, Somatic Mutations, and Vaccine Design," principal funding recipient is Søren Brunak) and by NIH (National Institute of Health) grants (contract no. HHSN266200400083C, principal funding recipient is Ole Lund; contract no. HHSN266200400025C, principal funding recipient is Søren Buus; contract no. HHSN266200400006C, principal funding recipient is Alessandro Sette).

## References

References are assembled at the end of the thesis.

# Chapter 3

# The epitope density in the alternative cancer exome

**Research Article**

## The cancer exome generated by alternative mRNA splicing dilutes predicted HLA class I epitope density

Thomas Stranzl[1], Mette V. Larsen[1], Ole Lund[1], Morten Nielsen[1], Søren Brunak[1]

[1] Department of Systems Biology DTU, Building 208, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, 2800, Denmark

## 3.1 Abstract

Several studies have shown that cancers actively regulate alternative splicing. Altered splicing mechanisms in cancer lead to cancer-specific transcripts different from the pool of transcripts occurring only in healthy tissue. At the same time, altered presentation of HLA class I epitopes is frequently observed in various types of carcinoma. Down-regulation of genes related to HLA class I antigen processing has been observed in several cancer types, leading to fewer HLA class I antigens on the cell surface. Here, we show that peptides unique to cancer splice variants comprise significantly fewer predicted HLA class I epitopes compared to peptides unique to normal transcripts. Peptides unique to carcinoma transcripts are in the case of the three most common HLA class I supertype representatives consistently found to contain fewer predicted epitopes compared to normal tissue. We observed a significant difference in amino acid composition between unique normal and carcinoma protein sequence as transcripts uniquely found in carcinoma are enriched with hydrophilic amino acids. This variation contributes to the observed significant lower likelihood of carcinoma-specific peptides to be predicted epitopes compared to peptides found uniquely in normal tissue.

## 3.2 Introduction

Cancer-specific splice variants are of significant interest as they may be involved in pathogenesis and may further potentially be used as biomarkers and generate novel targets for cancer therapy [112, 101]. The human immune system is capable of responding to some of these cancer specific antigens, as first shown by a melanoma-specific antigen, MAGE-1, able to stimulate human T cells [115, 26]. More generally, individuals with high or medium cytotoxic activity are further associated with a significantly lower risk of cancer, suggesting a role for natural immunological host defense mechanisms in cancer [37]. Alternative splicing can change the structure of mRNA by inclusion or skipping of exons, and this may alter the function, stability or binding properties of encoded proteins and thereby contribute to human diseases, such

as cancer [67]. In a study investigating alternative splicing events in ovarian and breast tissues affected by tumors it was found that about half of all splicing events in these tissues are altered in tumors, many of them due to exon skipping [117]. Similar trends have been seen in other types of cancers, e.g. in colon cancer and testicular tumor [33, 27], as well as in gastric cancer, where genes showing differential expression between cancer cell lines and corresponding normal tissues were found [72]. In addition to cancer being involved in dysregulating pathways, thus contributing to changes in alternative splicing and gene expression controlled by these proteins [16], human leukocyte antigen (HLA) class I antigen processing components and HLA expression have also been shown to be downregulated in connection with cancer [97, 90]. A study investigating alterations of HLA class I expression in 12 ovarian cancer patients reported low levels of HLA class I antigens in tumor cells from all patients. One patient-derived tumor cell line showed a complete haplotype loss, including the HLA-A2 locus [71]. These observations are interpreted as mechanisms adopted by tumors to escape immune surveillance and to avoid tumor cell recognition and destruction [62, 24]. It has been suggested that elimination of growing tumors by the immune system may lead to selection of tumor variants that are efficient in avoiding immune system recognition [123]. There thus seems to be accumulative evidence for cancer being coupled to alternative splicing as well as to an efficacy in evasion from the immune system by downregulation and altering HLA expression. Most of the studies relating cancer-specific alternative splicing to altered immune system surveillance are, however, of limited size and in most cases anecdotal. Here, we wanted to investigate, in a large scale study, if the alternative cancer exome already at the step of mRNA splicing would contain a bias compared to normal transcripts in the set of possible HLA class I epitopes.

## 3.3 Materials and Methods

### Data extraction from the ASTD database

The Alternative Splicing and Transcript Diversity database (ASTD) provides access to a collection of alternative splice events and transcripts of genes from human, mouse and rat [49]. The aim of the database is to analyze the mechanisms of alternative splicing on a genome-wide scale. It integrates a computational pipeline for detection and characterization of isoform splice patterns as well as alternative introns and exons. Our study is based on ASTD version v1.1 build 9. The database covers 14,194 human genes and lists 50,581 unique transcripts not covered by Ensembl genes. Based on related evidences from cDNA libraries, many of these transcripts are tagged with pathology information. The pathology information is given as eVOC ontologies, which is a controlled vocabulary for unifying gene expression data [43]. Two data sets were generated based on annotated pathology information. All transcripts tagged with the information of being expressed in normal tissue were assigned to subset N. This subset consisted of 30,739 transcripts derived from 11,980

| | | |
|---|---|---|
| Burkitt´s lymphoma | Glioblastoma | Myeloid leukemia |
| Ewing´s sarcoma | Glioma | Myeloma |
| T-cell leukemia | Hypertrophic cardiomyopathy | Neoplasia |
| Wilms tumor | Insulinoma | Neuroblastoma |
| Adenocarcinoma | Leiomyosarcoma | Oligodendroglioma |
| Adenoma | Leukaemia | Osteosarcoma |
| Astrocytoma | Liposarcoma | Papillary serous carcinoma |
| Carcinoid | Lymphoblastic leukemia | Phaeochromocytoma |
| Carcinoma | Lymphocytic | Polyp |
| Carcinoma in situ | Lymphoma | Retinoblastoma |
| Chondrosarcoma | Aalignant tumour | Rhabdomyosarcoma |
| Choriocarcinoma | Medulloblastoma | Sarcoma |
| Enchondroma | Melanoma | Seminoma |
| Fibrosarcoma | Meningioma | Teratocarcinoma |
| Fibrothecoma | Monocytic leukemia | Tumour |

**Table 3.1.** eVOC terms used for carcinoma subset.

| | **Normal** | **Carcinoma** |
|---|---|---|
| Number of transcripts | 30,739 | 27,967 |
| Number of genes | 11,980 | 10,730 |
| Number of uniquely associated transcripts | 16,566 | 13,794 |
| Number of uniquely associated genes | 8,741 | 7,128 |
| Average number of unique transcripts / gene | 1.90 | 1.94 |

**Table 3.2.** Number of transcripts and genes per set. Transcripts were extracted from the ASTD database. Number of transcripts and genes associated with normal and carcinoma pathology terms are given.

genes. A second subset, C, with transcripts related to carcinoma, consisted of 27,967 transcripts derived from 10,730 genes. The carcinoma subset consists of all transcripts tagged with eVOC terms related to carcinoma; that is being a subgroup of tumor in the eVOC ontology hierarchy (Table 3.1). Several eVOC terms can be associated to the same transcript. For our analysis, we were interested in transcripts uniquely associated to normal tissue or to one or more of the carcinoma eVOC terms. Two new subsets consisting of transcripts only associated to either normal or carcinoma eVOC terms were created. Out of 30,739 transcripts associated to normal, 16,566 were uniquely associated with normal tissue and not with carcinoma (unique N set). The subset of transcripts uniquely associated with carcinoma (unique C set) consists of 13,794 transcripts (see Table 3.2).

### Translation to proteins

All transcripts assigned to either normal or carcinoma pathology were translated to their respective protein sequence using Virtual Ribosome [125]. The longest ORF among all three reading frames was chosen as the translated

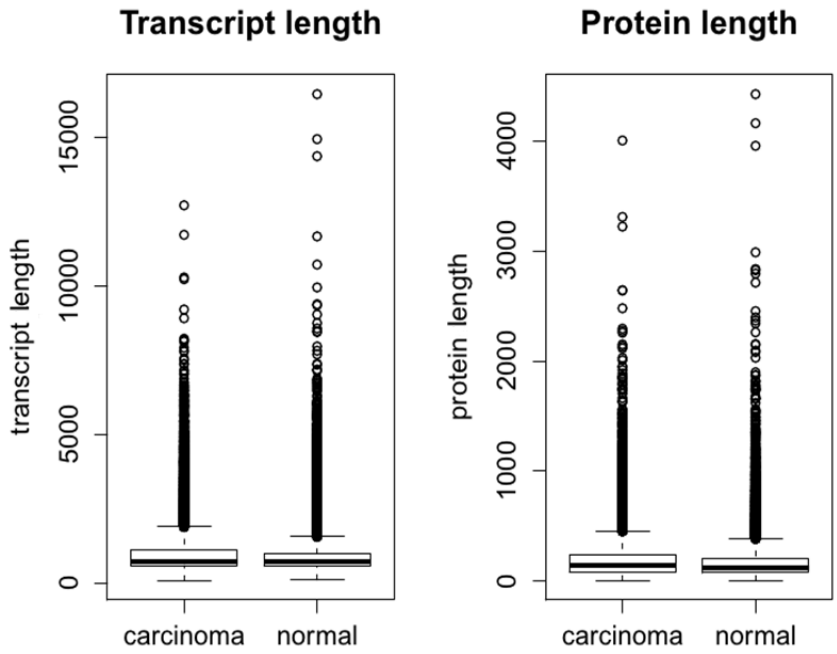


**Figure 3.1.** Lengths of transcripts and their respective proteins for transcripts assigned to normal and carcinoma groups.

protein sequence. The protein sequence and corresponding transcript were discarded if no ORF was found or if the resulting protein sequence was shorter than 9 amino acids. The threshold of 9 amino acids was chosen as we subsequently apply the epitope prediction on 9-meric peptides, although we are aware that proteins this small might not be functional. Applying this filter resulted in a normal set of 16,490 transcripts and a carcinoma set of 13,721 transcripts. The distribution of transcript and protein length is shown in Figure 3.1. The average length of transcripts from normal tissue is 1,004 nucleotides; carcinoma transcripts have an average length of 1,078 nucleotides. The average protein lengths are 185 and 212 amino acids, for normal and carcinoma transcript, respectively.

### Generation of unique 9-mers

All proteins assigned to either normal or carcinoma pathology states were divided into overlapping 9-meric peptide sequences. Peptide sequences that were found in both groups were removed, leading to the creation of two sets of unique 9-mer peptides. There are 1,856,231 unique 9-mers in the normal group (N-peptidome) and 1,684,028 unique 9-mers in the carcinoma group (C-peptidome). Note that normal and carcinoma sets do not consist of complete proteins; they only consist of unique 9-meric peptides not found in the other

set. Permutated sets of both the unique N and unique C set were created. For each set, one locally permutated and one globally permutated set of 9-meric peptides was generated. The local permutated sets were constructed by permuting each 9-mer, thus keeping the amino acid composition within each 9-mer fixed. The global permutated sets were made by randomly constructing new 9-mers out of all amino acids within each set. This preserves the overall amino acid composition within the unique N and C sets, local properties within each 9-mer are, however, destroyed.

### Prediction of possible HLA class I epitopes

The prediction method NetMHCpan-2.4 [35, 68] was used for predicting potential epitopes for the 12 HLA class I supertypes [57]. In practice, putative epitopes for a given HLA class I supertype were identified by predicting which peptides are presented by a specific HLA class I allele that represents the entire supertype (for example, HLA-A*0201 represents the A2 supertype). The NetMHCpan-2.4 method was trained on an experimentally validated data set of more than 100,000 quantitative peptide – HLA class I interactions covering more than 100 HLA molecules and has been evaluated as the best pan-specific method for HLA peptide binding in a large benchmark study [132]. A general accepted threshold for binding is a rank score of 1% [22, 81] (binding strength falling within the top 1% compared to a large set of random natural peptides), which is also the threshold, used throughout this study.

The percentages of potential epitopes per 9-mer for all 6 sets (normal 9-mers, normal globally permutated 9-mers, normal locally permutated 9-mers, carcinoma 9-mers, carcinoma globally permutated 9-mers and carcinoma locally permutated 9-mers) were calculated. P-values for difference in percentage of predicted epitopes between normal and carcinoma 9-mers for non-permutated and permutated subsets were calculated by a 2-sample test for equality of proportions and adjusted for multiple testing (Bonferroni correction).

### Amino acid scales

The amino acid abundance for normal tissue compared to carcinoma tissue was determined based on all unique 9-mers in the two data sets. The relative frequencies for all amino acids in both the normal and carcinoma sets were calculated. Observed ratio of frequencies (N/C) of amino acids among normal and carcinoma tissues was correlated with Hopp-Woods hydrophilicity [36] and Wimley-White hydrophobicity scale [128] values. The ratio was further correlated with a mean ranking scale per amino acid as published by Simpson [100]. According to Simpson, the scale is based on the mean ranking of amino acids according to the frequency of their occurrence at each sequence rank for 38 published hydrophobicity scales [113]. Other investigated scales are average volume of buried residues [84, 6], van der Waals volume [66] and total accessible surface area [61].

| Allele | Frequency |
|---|---|
| HLA-A*02:01 | 0.47 |
| HLA-A*01:01 | 0.30 |
| HLA-A*03:01 | 0.26 |
| HLA-B*07:02 | 0.24 |
| HLA-B*08:01 | 0.22 |
| HLA-A*24:02 | 0.13 |
| HLA-B*40:01 | 0.10 |
| HLA-B*15:01 | 0.07 |
| HLA-B*27:05 | 0.06 |
| HLA-A*26:01 | 0.05 |
| HLA-B*39:01 | 0.02 |
| HLA-B*58:01 | 0.02 |

**Table 3.3.** Phenotype frequencies. HLA frequencies in the European population. Data obtained from the dbMHC database [93].

Bootstrapping was applied to test if an amino acid property scale is correlated with enriched expression of residues in either unique normal or carcinoma 9-mers. For each scale, the Spearman rank correlation coefficient was calculated and the significance of the correlation was estimated using exact permutation test.

### HLA motif bias

HLA binding motifs were generated from NetMHCpan-2.4 training data. Position specific weight-matrices were calculated using sequence weighting and correction for low counts [70]. Sequence logos were visualized as described by Schneider and Stephens [95], where each letter represents its proportional frequency of the corresponding amino acid at that position. Based on amino acid frequencies and observed ratio of frequencies (N/C) of amino acids among normal and carcinoma tissues, we calculated for the HLA-A*A02:01, HLA-A*A01:01 and HLA-A*A03:01 motifs their respective overall bias towards either our unique normal or carcinoma peptide set. This was done for all 20 amino acids and for the 5 most frequent amino acid occurrences per motif. Per position, the tendency to fit preferably to either the normal or the carcinoma peptidome was calculated by summation of the respective amino acid frequencies multiplied with the related N/C values for all 20 amino acids. Likewise the calculation for the 5 most frequent amino acid occurrences per motif, where only the subset of the motif´s 5 most frequent amino acid occurrences at this position is considered. Similar to the N/C ratio, a motif´s bias to preferably fit to our normal set is given, if the average over all position for a motif is larger than 1.

## 3.4 Results

### For the three most common HLA class I supertypes, carcinoma transcripts contain fewer predicted epitopes

The aim of this study was to investigate, using a large-scale data set, if peptidomes specific for cancer and normal tissue have differential properties related to altered degree of immune system surveilance. To do this, we constructed two sets of peptides, one specific to carcinoma tissue and one uniquely expressed in normal tissue. Globally permutated versions of these sets were produced as described in Material and Methods. The global permutation destroys structural characteristics within the MHC binding 9-mers, only maintaining global compositional properties. For comparison, we constructed locally permutated normal and carcinoma sets by permutating each peptide separately thus preserving the local amino acid composition of each peptide. To investigate immune-related properties, potential epitopes covering all 12 HLA class I supertypes were predicted using NetMHCpan. For each supertype, we calculated the percentages of predicted epitopes for the six peptide data sets: normal, normal globally and normal locally permutated, carcinoma and carcinoma globally and carcinoma locally permutated. It is well known that some HLA class I supertype representatives are more common than others. It is therefore expected that for the less frequent HLA alleles the results are more likely to be more noisy. The source of our data set, the ASTD database, is to a large extent originating from EST data without HLA specific information. EST data is mostly based on Caucasian Europeans [28]; therefore we can safely assume that the more common HLA types in the European population are also more common in our dataset. The HLA allele frequencies were obtained from the dbMHC database [93]. Approximate numbers of expected phenotype per supertype in the European population are given in Table 3.3. The three most common supertype representatives in the European population are HLA-A*02:01, HLA-A*01:01 and HLA-A*03:01. For these three supertype representatives, the transcripts associated with normal tissue have a significantly higher percentage of predicted epitopes than transcripts uniquely found in carcinoma. Figure 3.2 shows the observed numbers, in percentages of predicted epitopes per 9-mers, for the different data sets for these three most common supertype representatives. All observed differences between normal and carcinoma tissues shown in Figure 3.2 are significant ($p < 0.006$, 2-sample test for equality of proportions).

### For most HLA class I supertypes, carcinoma transcripts contain fewer predicted epitopes

Further, the percentage of predicted epitopes for permutated and not-permutated sequences for all 12 supertype representatives is shown in Table 3.4. Here, we observed a similar tendency as compared to our observation for the three most common supertypes in the European population. For

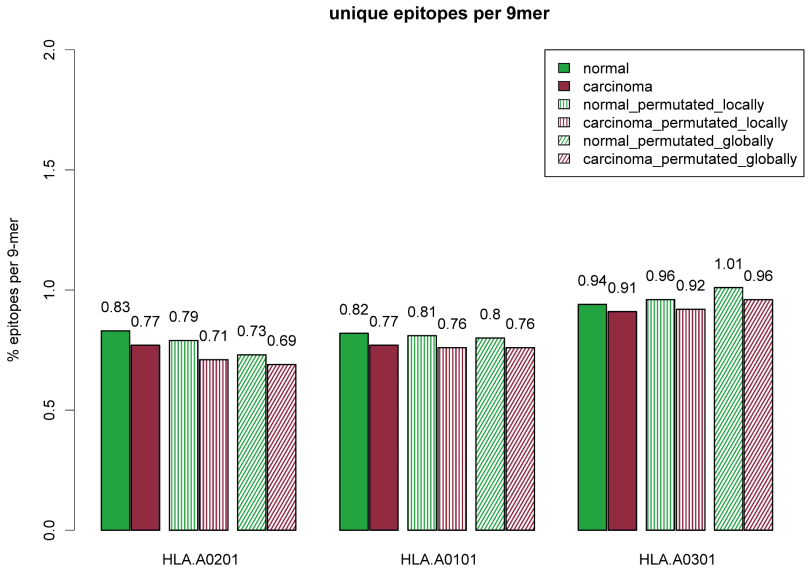


**Figure 3.2.** Percentage of epitopes per 9-mer comparison. Data is shown for the three most common HLA-I alleles in the European population. Each bar shows the percentage of predicted epitopes per 9-mer in the respective set. Each set consists of peptides that are either unique for normal or carcinoma tissue. Globally permutated or locally permutated version of the peptide sets were constructed as described in materials and methods. Based on each respective scale, more hydrophobic amino acids are colored green and more hydrophilic amino acids are colored red.

non-permutated sequences, seven out of the twelve supertype representatives (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*24:02, HLA-A*26:01, HLA-B*15:01 and HLA-B*58:01) had a significant lower fraction of predicted epitopes in sequences assigned to carcinoma pathology. A statistical significant difference, where unique carcinoma peptides contained more predicted epitopes was, on the other hand, only observed for one supertype representative, namely HLA-B*27:05. When analyzing permutated sequences, similar results were observed. Only one supertype representative (HLA-B*40:01, locally permutated) had significantly more predicted epitopes in the permutated carcinoma sequences than in the permutated normal sequences. On the other hand, permutated, normal sequences had consistently for both the local and global permutated sets more predicted epitopes for seven supertype representatives (HLA-A*01:01, HLA-A*02:01, HLA-A*03:01, HLA-A*24:02, HLA-A*26:01, HLA-B*15:01, HLA-B*58:01). For these seven supertype representatives, the difference between normal and carcinoma data sets is significant in the permutated as well the non-permutated data sets. The observation that carcinoma transcripts contain fewer predicted epitopes for most

HLA class I supertype representatives, is stable, when different thresholds for the prediction of potential epitopes are applied (data not shown).

### HLA motif and amino acid composition biases in carcinoma sequence

The relative difference in predicted epitope density between normal and carcinoma is, for our previously defined most common HLA molecules, relatively stable. Also, the difference in epitope density is largest when comparing non-permutated to globally permutated peptide sets. For HLA-A*02:01, a noticeable decrease of predicted epitopes is observed when comparing normal and carcinoma non-permutated peptides to normal and carcinoma permutated peptides. As seen from Table 3.4 and Figure 3.2, the difference in percentage of epitopes is the largest when comparing the non-permutated sequences to the globally permutated sequences (normal: 0.83 vs 0.73, carcinoma: 0.77 vs 0.69). For HLA-A*01:01, the percentage of epitopes in non-permutated versus permutated sequences appears to be relative stable (normal: 0.82 vs 0.80, carcinoma: 0.77 vs 0.76), whereas permutated HLA-A*03:01 sequences have more predicted epitopes than the corresponding non-permutated sequences (normal: 0.94 vs 1.01, carcinoma: 0.91 vs 0.96). For these three supertype representatives, the percentage of predicted epitopes in locally permutated peptides always falls between the respective percentages for non-permutated and globally permutated sequences. Locally permutated peptides preserve only local amino acid composition, and globally permutated peptides have their local structural properties destroyed and preserve only global amino acid composition. These observations indicate that both global and local structural amino acid properties are factors that define the observed differences in the epitope densities between the normal and carcinoma peptidome. An analysis of relative amino acid composition was done for all unique normal and carcinoma 9-mers. We found that hydrophilic residues are more common in unique carcinomic sequences as compared to normal sequences. The relations of N/C ratios compared to the hydrophilicity scale of amino acids by Hopp-Woods, the hydrophobicity scale by Wimley-White as well as the mean ranking of amino acids according to the frequency of their occurrence for 38 published hydrophobicity scales are shown in Figure 3.3. The Hopp-Woods and Wimley-White scales correlate strongly with the N/C ratios with a Spearman rank correlation coefficient of -0.72 and 0.78, respectively. The mean ranking amino acid scale is correlated with a correlation coefficient of -0.65. All three correlation coefficients are significant (p-value $< 0.003$, exact permutation test). No correlation was found for other amino acid properties like mass, surface area or volume (data not shown). It is striking to observe that all strong hydrophilic amino acids (KPRQ, Hopp-Woods scale) are enriched in sequences unique to carcinoma. A similar observation is made for Wimley-White scale: We identified seven amino acids significantly more common in carcinoma (APERKDQ). Six out of these (all except A) are within the seven most hydrophilic amino acids based on the Wimley-White scale. A reversed

trend is found for hydrophobic amino acids. The top significant amino acids classified by both Hopp-Woods and Wimley-Scott as hydrophobic (WFICM) are all more common in sequences uniquely assigned to transcripts from normal tissue. Based on these findings, one could suggest an explanation for the difference in epitope density between the normal and carcinoma peptidome. The binding motifs for the 3 most frequent supertype representatives are shown in Figure 3.4. Four amino acids (VMIA), which are preferred at the HLA-A*02:01 anchor positions are enriched in normal transcripts, whereas only one (L) is as common in normal as in carcinoma. This leads to the obvious conclusion that at least part of the observed differences in percentage of predicted epitopes in normal versus carcinoma transcripts are due to amino acid composition. The same tendency is found for HLA-A*01:01. The two most frequent amino acids in the motif (YT) are also more often found in normal tissue, whereas S is neutral and the next common amino acid, D, is more common in carcinoma. The most frequent amino acid for HLA-A*03:01(K) is slightly more common in carcinoma, whereas the second-next frequent (Y) is, due to a stronger preference to fit peptides from normal tissues, shifting the bias towards amino acids more common in unique normal splice variants. For all three motifs, we further calculated average weighted biases, based on N/C ratios and amino acid frequencies (see materials and methods). Covering the respective 5 most frequent amino acids per motif as well as all 20 amino acids, we observed for all three motifs an overall preference for amino acids found in our normal tissue set.

## 3.5 Discussion

Alternative splicing of mRNA transcripts is an important mechanism for generating genomic complexity and has been shown to differ between carcinoma and the corresponding normal tissues [112, 33, 27]. In addition, cancers in some cases downregulate HLA class I antigen-processing components and HLA class I expression to avoid detection by the immune system. These observations led us to investigate whether transcripts found in cancer tissue share characteristics that would reduce immune system recognition. Here, we have carried out a large-scale analysis aiming at identifying immune system related imprints that can differentiate carcinoma from normal transcripts. We identified two peptide data sets, one uniquely associated with carcinoma transcripts and one uniquely associated with normal transcripts. Using state-of-the-art immunoinformatics predictions tools, we next analyzed the two data sets for differences in terms of likelihood of being presented on prevalent HLA class I molecules, and hence potential for activating the immune system. We found that peptides, which due to alternative splicing are uniquely expressed in carcinoma tissue, contain fewer predicted epitopes restricted by the three most common HLA class I alleles than peptides expressed uniquely in normal tissue. Using globally permuted data sets we consistently, for the three most common HLA class I alleles, found that the observed loss in epitope density in the carcinoma peptidome is maintained

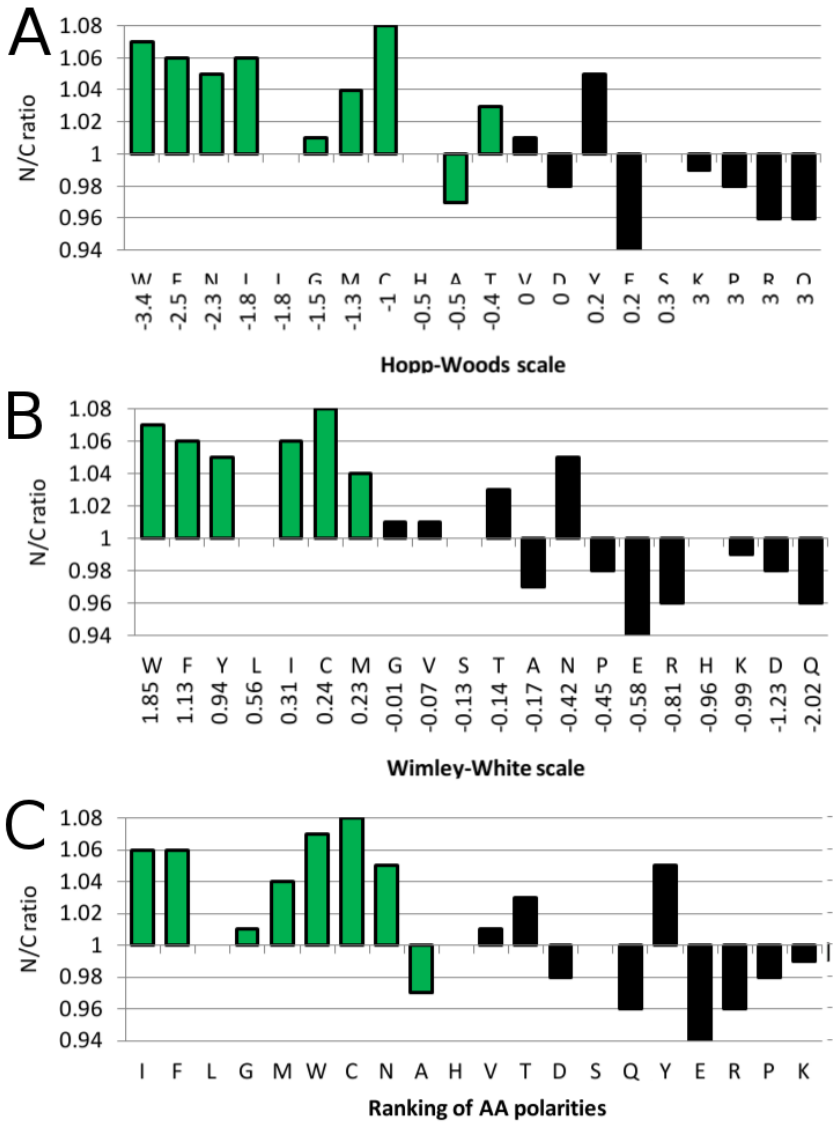


**Figure 3.3.** Hydrophilic amino acids are enriched in carcinoma. N/C ratios in relation to Hopp-Woods hydrophilicity scale (A), Wimley-White hydrophobicity scale (B) and to the mean ranking of amino acids based on 38 hydrophobicity scales (C). N/C ratio is the ratio of observed frequencies of the respective amino acids among normal and carcinoma tissues. Residues where N/C $> 1$ are more common in normal tissue; residues where N/C $<$ 1 are more common in sequence assigned to carcinoma. Green bars refer to more hydrophilic amino acids whereas black bars refer to more hydrophobic amino acids. All N/C ratios larger or smaller one are significant ($p < 0.001$, calculated using the Wilson score [127] and Bonferroni corrected).

| Allele | % epitopes | | | | % epitopes globally permutated | | | | % epitopes locally permutated | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **N** | **C** | **N/C** | **P-val** | **N** | **C** | **N/C** | **P-val** | **N** | **C** | **N/C** | **P-val** |
| HLA-A*01:01 | 0.82 | 0.77 | 1.06 | 0.000 | 0.80 | 0.76 | 1.06 | 0.000 | 0.81 | 0.76 | 1.07 | 0.000 |
| HLA-A*02:01 | 0.83 | 0.77 | 1.08 | 0.000 | 0.73 | 0.69 | 1.05 | 0.000 | 0.79 | 0.71 | 1.10 | 0.000 |
| HLA-A*03:01 | 0.94 | 0.91 | 1.04 | 0.002 | 1.01 | 0.96 | 1.05 | 0.000 | 0.96 | 0.92 | 1.04 | 0.005 |
| HLA-A*24:02 | 0.89 | 0.79 | 1.13 | 0.000 | 0.77 | 0.70 | 1.11 | 0.000 | 0.89 | 0.77 | 1.15 | 0.000 |
| HLA-A*26:01 | 0.76 | 0.71 | 1.07 | 0.000 | 0.70 | 0.66 | 1.06 | 0.000 | 0.71 | 0.68 | 1.05 | 0.001 |
| HLA-B*07:02 | 1.29 | 1.30 | 1.00 | 1.000 | 1.27 | 1.29 | 0.99 | 1.000 | 1.25 | 1.28 | 0.97 | 0.054 |
| HLA-B*08:01 | 1.02 | 1.03 | 0.99 | 1.000 | 1.00 | 0.99 | 1.01 | 1.000 | 0.97 | 0.97 | 1.00 | 1.000 |
| HLA-B*15:01 | 0.86 | 0.79 | 1.09 | 0.000 | 0.83 | 0.77 | 1.08 | 0.000 | 0.85 | 0.79 | 1.07 | 0.000 |
| HLA-B*27:05 | 0.99 | 1.02 | 0.97 | 0.021 | 0.99 | 1.00 | 0.98 | 1.000 | 1.04 | 1.04 | 0.99 | 1.000 |
| HLA-B*39:01 | 0.97 | 0.96 | 1.02 | 0.985 | 1.05 | 1.02 | 1.02 | 0.221 | 1.01 | 1.00 | 1.01 | 1.000 |
| HLA-B*40:01 | 0.87 | 0.89 | 0.98 | 1.000 | 1.03 | 1.06 | 0.98 | 0.325 | 0.95 | 0.99 | 0.96 | 0.002 |
| HLA-B*58:01 | 1.01 | 0.91 | 1.11 | 0.000 | 0.99 | 0.89 | 1.10 | 0.000 | 0.98 | 0.91 | 1.08 | 0.000 |

**Table 3.4.** Epitopes per set for all supertype representatives. Percentage of predicted epitopes is given for data extracted from ASTD database as well as for permutated sequences. N/C is the ratio between the normal and carcinoma percentages. P-values are calculated by two-tailed t test and adjusted for multiple testing by Bonferroni correction.

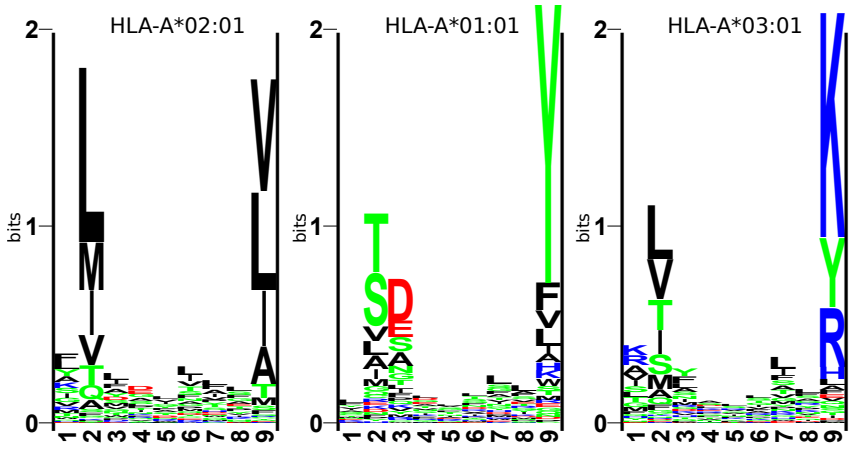


**Figure 3.4.** Human HLA motifs. The 3 most common HLA types in the European population. The height of a column of letters is equal to the information content at that position, whereas the height of each letter within a column is proportional to the frequency of the corresponding amino acid at that position [95].

also for the permuted data sets. This strongly indicates that differences in amino acid composition between peptides from alternatively spliced normal and carcinoma transcripts are the driving force of the reduced predicted epitope density. The reason for the observed change in frequency of specific amino acids in proteins unique for carcinoma as compared to normal tissue is unknown, but the phenomenon has previously been observed in studies aiming at identifying biomarkers for early stage detection of cancer: In a recent study, the levels of alanine, isoleucine, leucine and valanine were found to be increased in the pancreases of rats with pancreatic cancer as compared to samples from rats with chronic pancreatitis and healthy rats [23]. In another study, the levels of N-methylalanine and lysine were found to be significantly increased in the plasma from pancreatic cancer patients, while the level of glutamine and phenylalanine was found to be decreased [114]. These studies identified differences in amino acid composition in a single cancer type based on blood plasma and tissue samples. We, in contrast, analyze peptides unique to cancer in general. As to be expected, the findings regarding amino acid concentration reported in this study are not concurrent with those of the single cancer type studies. A possible explanation as to why we observed fewer predicted epitopes in peptides, which due to alternative splicing are uniquely expressed in carcinoma, could be that the host's immune system restricts the cancer exome. In that case, pressure from the immune system disfavors cancer cells that present new epitopes at the cell surface. An alternative explanation - which does not exclude the previous explanation - takes as starting point the observed change in amino acid frequency, especially the

increase in hydrophilic amino acids in carcinoma proteins. It has been suggested that the stabilization of a protein structure is to a large part due to the hydrophobic effect [42]. Accordingly, the increase in hydrophilic amino acids has a destabilizing effect on protein structure, which is in concordance with the protein loss-of-function that is correlated with cancer progression. This is exemplified by a study concerning inherited missense mutations of the tumor suppressor gene, BRCA-1, which may predispose to breast or ovarian cancer [25]. In this study, it was found that the mutations predominately target conserved hydrophobic amino acids that are responsible for folding and stability. Since, in particular, the most common HLA class I allele, A*02:01, prefers hydrophobic amino acids at the anchor positions, an increase in hydrophilic amino acids will inevitably lead to fewer predicted epitopes. The reduction in epitope density in unique carcinoma peptides might therefore be an intrinsic property of proteins that are destabilized by a decrease of hydrophobic amino acids as part of the progression to cancer. To our knowledge, this is the first study indicating that alternatively spliced carcinoma transcripts tend to express fewer potential epitopes than alternatively spliced transcripts found in normal tissue. The identified difference in amino acid composition towards hydrophilic amino acids in the alternative spliced cancer exome is a possible explanation for the bias in potential HLA class I epitopes. The preference for hydrophilic amino acids at the step of alternative mRNA splicing, could support the development of carcinoma by providing it with the possibility of evading the host's immune system. In this case by leading to fewer potential HLA class I epitopes presented at the cell surface.

## Grant Support

This work was supported by a grant from the Danish Research Council for Technology and Production Sciences (Project "Disease Gene Finding, Somatic Mutations, and Vaccine Design;"principal funding recipient, Søren Brunak) and was supported by the National Institutes of Health (contract HHSN26620040006C).

## References

References are assembled at the end of the thesis.

# Chapter 4

# Discovery of minor histocompatibility antigens associated with hematologic malignant diseases

## 4.1 Introduction

This study presents the analysis of 93 patients that underwent allo-HCT at the Department of Hematology, Rigshospitalet in Copenhagen. HCT is a standard treatment for a variety of hematological malignancies. Although patients are 10/10 allele-matched, small variations in proteins, so-called mHags, can still cause undesirable immune responses, such as GVHD. Donor reactivity against patient cells, as induced by mHags, is responsible for GVHD, but it may also lead to a beneficial GVT effect.

The study is initiated with a comparison to a previous study, followed by an analysis of gene-specific correlations to clinical outcome, namely the separation of genes into those in which it is beneficial or deleterious to have nsSNPs or mHags. A clear distinction is a prerequisite for identifying mHags that can be used in therapy. Ideally, one can isolate mHags that trigger the favorable GVT effect without leading to GVHD. To identify potential mHag disparities, all patients and donors were genotyped for more than one million SNPs by a BeadChip Array.

The design of the study was done in collaboration with the Department of Systems Biology (Thomas Stranzl, Mette V. Larsen, Ole Lund) at the Technical University of Denmark, the Department of Hematology (Lars Vindeløv, Bo Mortensen, Brian Kornblit, Henrik Sengeløv) at the Rigshospitalet, the Department of Biostatistics (Thomas A. Gerds) at Copenhagen University, and the Department of Health, Immunology and Microbiology (Søren Buus, Anette Stryhn) at the University of Copenhagen.

## 4.2 Patients

This study is based on 93 patients and their HLA identical related or 10/10 allele-matched unrelated donor. All patients underwent an allo-HCT with a peripheral blood graft from their respective donor at the Allo-HCT Laboratory, Department of Hematology, Rigshospitalet in Copenhagen. Donor selection was based on molecular typing for HLA-A, B, C, DRB1 and DQB1. All donors were from the same gender as the patients. Patient treatment and supportive care were done as described in Kornblit et al [48]. All patients were treated by HCT for a hematologic malignant disease. Out of 93 patients, 35 were treated for acute myeloid leukemia, 19 for chronic lymphocytic leukemia, 24 for non-Hodgkin lymphoma, 11 for myelodysplastic syndrome, three for myelofibrosis and one for mantle cell lymphoma. 56 patients where male, 37 where female. The median age of the patients at the time of transplantation was 56 years, the youngest was 31 whereas the oldest was 70. 40 of the patients had a related donor, the remaining 53 had an unrelated donor.

## 4.3 Genotyping

Illumina´s four-sample HumanOmni1-Quad BeadChip Arrays were used for determining the genomic variation of the 93 leukemia patients, as well as

their respective donor. The BeadArray delivers genome-wide coverage and covers more than one million markers. These markers, covering SNPs and CNVs, are based on published disease associate studies, all three phases of the International HapMap Project as well as markers from the 1,000 Genomes Project. The BeadArray is based on the human genome build NCBI 36.2, USCG hg18. Genotyping and data preparation was done by AROS Applied Biotechnology AS, Aarhus, Denmark. On average, there were 1,006,041 SNPs (+- 36,165 SD) successfully genotyped per sample.

## 4.4 Identification of nsSNPs differing in Graft vs Host direction

All SNP differences per patient-donor pair were extracted. Only SNPs successfully identified by genotyping in both the patient and the donor were taken into account. On average, 303,057 (+- 77,123 SD) SNPs differed per patient-donor pair. Per SNP, gene name, protein name and amino acid sequence were extracted from Ensembl build 54_36p. A SNP's effect on the transcript and further the amino acid sequence were determined by use of the Ensembl-api [86]. SNPs present on the Illumina HumanOmni1-Quad BeadChip Array are biallelic and listed with respect to their TOP/BOT strand designation; in addition, Illumina lists alleles in accordance with forward orientation of dbSNP. Strand designation is, however, not consistent with forward/backward orientation from Ensembl. Therefore, both nucleotides per SNP as well as the respective reverse complement were queried against Ensembl. A SNP is defined as protein coding if 1 of the 4 nucleotides at a given chromosome and position is consistent with the nucleotide leading to a reference protein from Ensembl. If Illumina´s forward allele is linked to a reference protein, then Illumina´s strand information is in accordance with Ensembl. If, however, the reverse complement is linked to a reference protein, then Illumina´s strand is, compared to Ensembl, reversed. Based on this, Illumina SNPs were flagged to be in accordance with Ensembl. This approach is successful if dealing with non-ambiguous SNPs. Ambiguous SNPs (A/T and C/G SNPs) would, if they are protein coding, match as given and as their reverse complement. Within our dataset, we have 12.3% ambiguous SNPs. Strand information relative to Ensembl for ambiguous SNPs was obtained by blasting the probe sequence against the human genome NCBI build 36.2. The matched strand was assigned if there was a clear match on only one strand. Overall, we identified 20,036 unique SNPs with a varying genotype in at least one patient-donor pair. A subset of 17,580 SNPs were non-ambiguous SNPs. We could assign reliable strand to additional 1,673 SNPs, whereas 783 SNPs where removed from the data set. Most of the on average 303,057 SNPs differing between patients and their donors are either in non-coding regions or do not alter the amino acid sequence of the coded protein. An mHag presumes a change in amino acid sequence, therefore we extracted the subset of non-synonymous coding SNPs (nsSNPs) with correctly identified strand of the previously identified SNP differences. On average, there are 4,922 (±

1,348 SD) nsSNP differences per patient-donor pair. A nsSNP must differ in GvH direction for being considered as a potential mHag. The removal of nsSNPs not differing in GvH direction resulted in a new set of 2,748 (mean, ± 825 SD) nsSNPs in GvH direction per patient-donor pair.

## 4.5 Identification of potential mHags

Minor histocompatibility antigens (mHags) are epitopes raising an immunological response in some organ transplants. Each of the above identified nsSNPs in the Graft vs Host direction could lead to potential mHags of interest. For each nsSNP difference in GvH direction, the binding of the peptide containing the SNP was assessed using NetMHCpan 2.4 [35]. 9-meric peptides were used as possible binders, as most peptides binding to MHC class I consists of 9 amino acids [59]. Two possible nucleotides for a nsSNP result in two protein sequences differing by one amino acid. For each possible amino acid per nsSNP, there are 9 unique 9-mers, with a varying position of the unique amino acids from one to nine within the 9-mer. Per donor-patient pair, the binding predictions were applied to the patients HLA-A, -B and -C molecules on the unique 9-mers occurring in the patient. As a patient has up to six different alleles and 9 possible 9-meric mHAgs per SNP, the number of predicted 9-meric mHAgs for a specific binding strength threshold can vary from zero per SNP to more than one per SNP. With a binding strength threshold of 1% (binding strength falling within the top 1% compared to a large set of random natural peptides) the average number of potential mHags in the GvH direction per patient-donor pair is 790 (± 267 SD).

## 4.6 Comparison to previous study

In a previous study by Larsen et al. [51], an impact on overall survival based on the number of nsSNP and mHags was shown. Patients with fewer nsSNP disparities with their donor and subsequent with fewer predicted mHags had a better survival rate than patients with many nsSNPs disparities and predicted mHags. The study was based on 126 patients who underwent allo-HCT. SNPs in this study were identified without high-throughput methods; only SNPs within genes with known minors were studied.

The study by Larsen et al. is partly based on the same patient set as the 93 patients that were genotyped with the HumanOmni1-Quad BeadChip Arrays. The two studies have 73 patient-donor pairs in common. In the study by Larsen et al. 96 SNPs were genotyped per patient-donor pair. The selection of SNPs for the study was based on proteins known to contain mHags. Eleven non-Y chromosomal proteins with known mHags were selected from the dbMinor database [106]. Although the genotyping approach using the BeadChip Array covered more than one million SNPs, only 33 of the previously genotyped SNPs were included. In the 73 patient-donor pairs that overlapped between the two studies, these 33 SNPs were identified identically by the two genotyping approaches.

| Gene | nsSNPs | occurence |
|---|---|---|
| **AKAP13** | **16** | **348** |
| **BCL2A1** | **3** | **76** |
| C19orf48 | 2 | 34 |
| CENPM | 1 | 23 |
| CTSH | 2 | 24 |
| HMHA1 | 3 | 61 |
| **HMHB1** | **1** | **16** |
| **KIAA0020** | **2** | **4** |
| MYO1F | 1 | 1 |
| **MYO1G** | **2** | **41** |
| **SP110** | **9** | **194** |
| TOR3A | 1 | 28 |
| TYMP | 1 | 10 |

**Table 4.1. Selected genes with known mHags.** The table lists 13 genes with known mHags, where we observed nsSNP-variation in the GvH direction. Genes in bold are the subset of genes in common with the study by Larsen et al. [51]. The number of nsSNPs in the GvH direction for each gene is shown. Occurrence is defined as the number of nsSNP disparities observed per gene over all 93 patient-donor pairs.

### Genes with known mHags

In order to analyze a comparable subset of SNPs and mHags, in addition to all SNPs covered by the BeadChip Array, we made a similar selection of SNPs as Larsen et al. Genes were selected based on dbMinor. As these genes have been found to contain mHags, they are of interest in the context of hematopoietic cell transplantation. Based on dbMinor, there are 14 non Y-chromosomal genes with known mHags. Two additional genes (C19orf48 and MYO1F) were extracted from [104]. 13 out of these 16 genes had nsSNPs in the GvH direction in our patient-donor set (see Table 4.1). In addition, the number of unique SNPs expressed within these genes is shown. The six genes listed in bold are also included in the analysis of Larsen, whereas the other 7 genes are only included in our study. The previous study excluded these genes (as well as C19orf48 and MYO1F) as the patient-donor pairs analyzed did not display SNP variation for these genes.

### Related and unrelated donor separation

Within our data set of 93 patient-donor pairs, 40 of the patients had a matched related donor (MRD) and 53 had a matched unrelated donor (MUD). In the study by Larsen et al., the number of nsSNPs in the GvH direction was comparable between MRDs and MUDs as shown in Figure 4.1 A. In the presented study, we analyze far more SNPs per patient and a clear

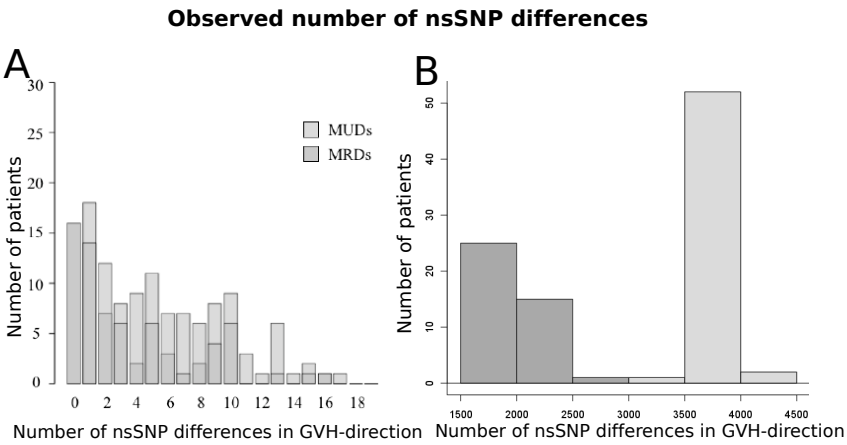


**Figure 4.1. Histogram of nsSNPs in related and unrelated patient-donor pairs.** Both plots display the distribution of nsSNPs in GvH direction. Plot A shows the distribution of analyzed nsSNP differences for MRD and MUD patient-donor pairs based on the study by Larsen et al. [51]. The distributions for the two groups in this study were roughly equally distributed. Our data (plot B) displays, based on the number of nsSNPs, a clear distinction between MRDs and MUDs.

separation between the numbers of nsSNPs in the GvH direction for MRDs versus MUDs become apparent. Figure 4.1 B shows that related and unrelated pairs have their own distribution of nsSNPs. Related donors are clearly separated from unrelated donors with regards to the number of SNPs. In the study of Larsen, patients were splitted into different cohorts based on the number of observed nsSNP and mHag differences. In the presented study, a similar separation of donor-patient pairs by the number of nsSNPs would just be a separation in related vs. unrelated donors. For the correlation of the number of predicted mHags with overall survival (OS), we have therefore chosen to analyze MRDs and MUDs separately.

### Overall survival analysis

Larsen et al. observed a non-significant higher OS for patients with fewer nsSNP differences in the GvH direction than patients with more nsSNP differences (Figure 4.2A). Applying the same analysis on predicted mHags instead of on nsSNPs boosted the difference in OS between patient-donor pairs with few versus many predicted mHags. Patients with few predicted mHags were shown to have a significantly higher OS than patients with many predicted mHags 4.2B).

In the present study, we applied a log-rank test, based on the number of observed nsSNP or mHags differences, to estimate the association between the number of observed differences and overall survival. To minimize the

effect of different follow-up lengths for different patients, the follow up time of all patients was set to 1000 days. Related and unrelated matched donor pairs were analyzed separately. Limiting our OS analysis to the 6 genes in common with the study by Larsen, we find the same tendency for MRD patients as Larsen et al. For the graphical representation, patient-donor pairs were divided by the median of nsSNP or mHag differences. As shown in Figure 4.3 we observed a comparable difference between MRD patients with few/many predicted mHags and few/many nsSNPs. Calculated p-values are not significant; in contrast to Larsen et al. our p-value is calculated on overall survival only, it is based on a smaller set of patient-donor pairs, and does not integrate other factors.

Each analysis was further done on the whole set of genes and on the subset of genes with known mHags. No difference in overall survival was observed for MUD patient-donor pairs. For the MRD pairs, we could identify tendencies, but the observed differences were not statistically significant (data not shown). Investigating nsSNP differences in the GvH direction resulted in a lower p-value (p=0.059) when analyzed for genes with known mHags only, as compared to all genes (p=0.244). The tendency of fewer nsSNPs being correlated with a higher OS is given, but only when we limit the analysis to genes with known mHags (Figure 4.4. Correlating OS with predicted mHags instead results in the same tendencies, reported p-values are, however, higher. Based on our results, we could identify some tendencies, but we could not reproduce the findings by Larsen et al. that the difference in OS is more clearly correlated to the number of predicted mHags than to the number of nsSNP differences. Our patient cohort is not independent from the cohort analyzed in the other study, but our sample size is. Due to overall fewer patients and the further split in MRDs and MUDs, it is considerably smaller. Since we could not observe an improvement when using the number of predicted mHags as a predictor for OS instead of the number of nsSNPs, we did all subsequent analyses based on both nsSNP and mHag differences.

Although we could not reproduce the results by Larsen et al., we are not stating that adding predictions of mHags is not resulting in a stronger signal than nsSNPs disparities alone. Instead, we suggest that the gene that contains the nsSNP or mHag is an important factor that must be taken into account. It is hence likely that nsSNP disparities and mHags in some genes will lead to elevated levels of GVHD and lower OS, while nsSNP disparities and mHags in other genes preferably will lead to GVT and hence higher OS [7]. This separation could for instance be based on the tissue expression of the gene in question: nSNPs and mHags in genes that are preferentially expressed in hematopoietic tissues 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, nsSNPs and mHags with a broad tissue expression would carry the risk of inducing potentially life-threatening GVHD. By chance, the nsSNP disparities in the set of patient-donor pairs analyzed by Larsen et al. are distributed in such a way that it was possible to observe an improved predictive performance on OS

when using mHags rather than nsSNPs. In another set of patient-donor pairs and in particular when including nsSNPs in other genes, this improvement is lost, since nsSNPs and mHags in some genes will lead to GVHD and lower OS, while they in other genes lead to GVT and higher OS. It is the aim of this study to try to identify these sets of genes in which it would be, respectively, beneficial and deleterious for the patients to have nsSNP disparities and mHags.

## 4.7 Gene-specific analysis

### Modeling disease course

Based on disease progression data for each patient, we modeled the disease course of patients after HCT. An association of genes with OS alone would be less sufficient, as patients could for example die, as a direct consequence of relapse, or die of because of other reasons.

We analyzed the time from the day of transplant to an event. The following events were considered: relapse or relapse related death, acute GVHD, and chronic GVHD (see Figure 4.5. Patients with grade I acute GVHD were considered as not affected by acute GVHD.

### Association analysis

We analyzed the association of the genotype (number of SNPs/number of epitopes) and the cause-specific hazards of the events acute and chronic GVHD by censoring patients who died at their time of death. To analyze the association of genotype and the cause-specific hazard of relapse we considered death due to relapse as events and censored only patients who died in remission.

We analyzed the cause-specific parameters and not the cumulative incidences because the latter also reflect the rates of the events of the competing events, and we were only interested in the biological mechanism that drives the risk of an event. See Kahl et al. [39] for a similar argument.

For each of the three time-to-event responses, the nsSNP differences and potential mHags were analyzed separately. Overall, we had 9,162 genes where one or more patient-donor pairs had at least one nsSNP differing in Graft vs. Host direction. The prediction of potential mHags resulted in a reduced set of 6,359 genes.

For the most interesting genes, all or the great majority of patients who relapsed (respectively had acute or chronic GVHD) had the same genotype: If they had negative genotype (zero SNPs or zero predicted mHags) then this gene is a candidate protective gene. If they had positive genotype (at least one SNP or one predicted mHag) then this gene is a candidate risk gene. It is important to note that for these genes, where e.g. all observed events occur in patients with concurrent genotype, a univariate or multivariate Cox regression model cannot be estimated, as there are no events in one of the genotype groups. For the same reason, GVHD could not be considered as

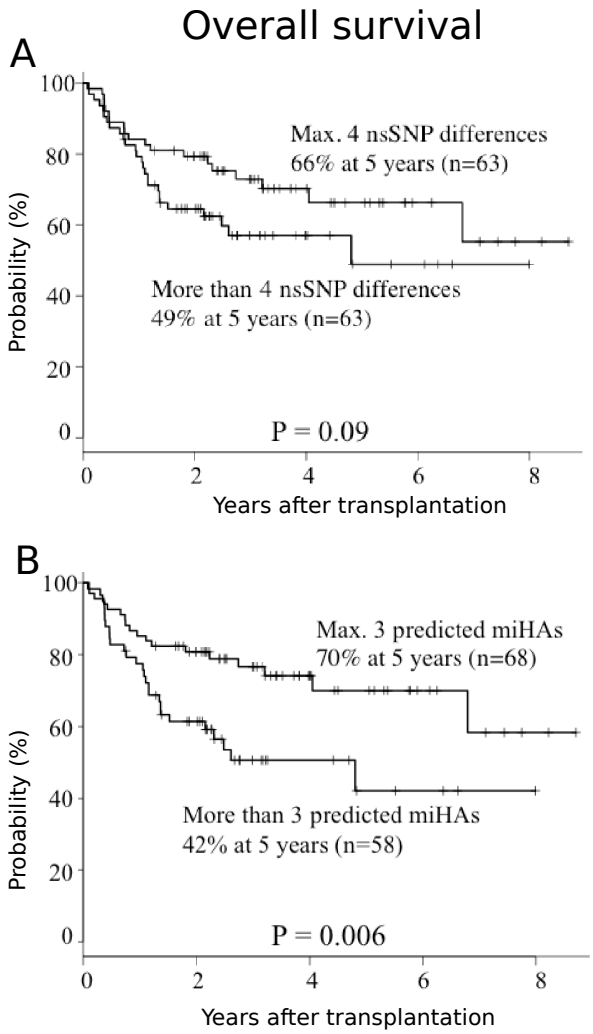


**Figure 4.2. nsSNPs and predicted mHags as possible markers for overall survival.** Difference in overall survival for patients with few/many nsSNPs/mHags as shown by Larsen. The number of predicted mHags (B) is a better marker for overall survival than the number of observed nsSNPs (A).

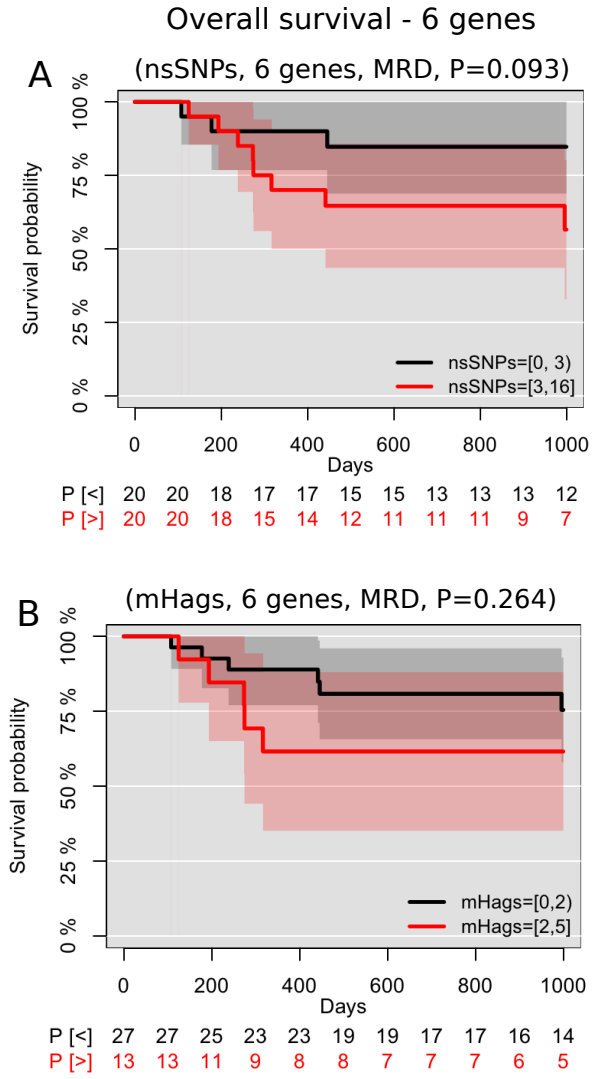


**Figure 4.3. Probability of overall survival for nssNPS and mHags, based on 6 genes.** Analysis based on genes in common with the study by Larsen et al. Selected genes are known to encode mHags. Plot A is based on nsSNP differences for matched related patients, plot B shows the analysis based on predicted mHags. P-values are based on log-rank tests with the number of observed nsSNP/mHag differences as predictors. The black line shows the survival probability for patients with fewer observed differences. The legends list the cutoffs for separating patients into groups based on number of nsSNP/mHags differences. Shaded areas are the confidence intervals.

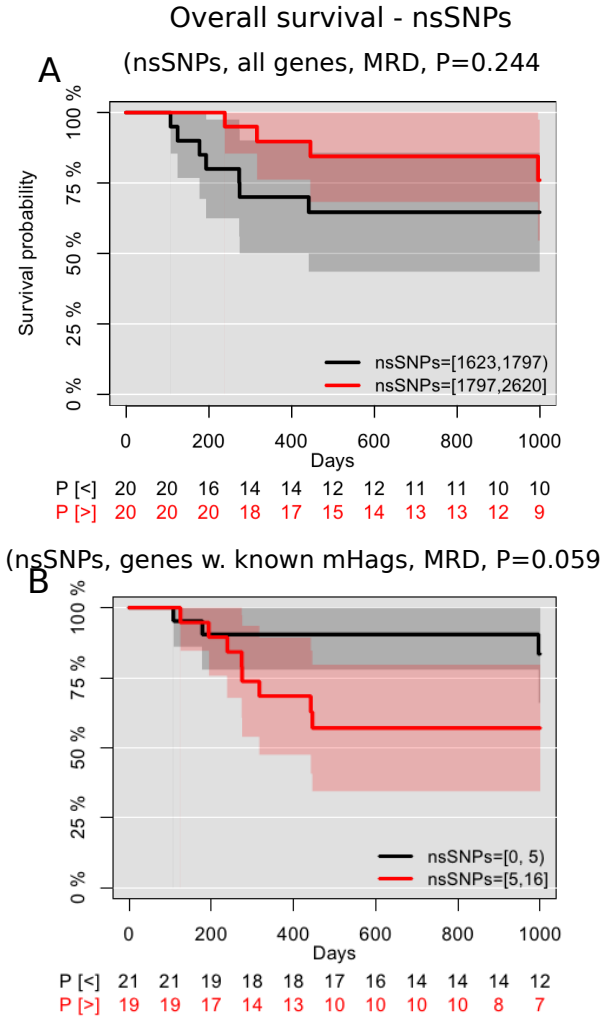


**Figure 4.4. Probability of overall survival based on observed nsSNP differences.** Plots are for matched related donors. Plot A is based on all nsSNP differences in GvH direction, as identified by our study. Plot B includes only nsSNPs from 13 genes (see Table 4.1), namely genes known to express mHags.

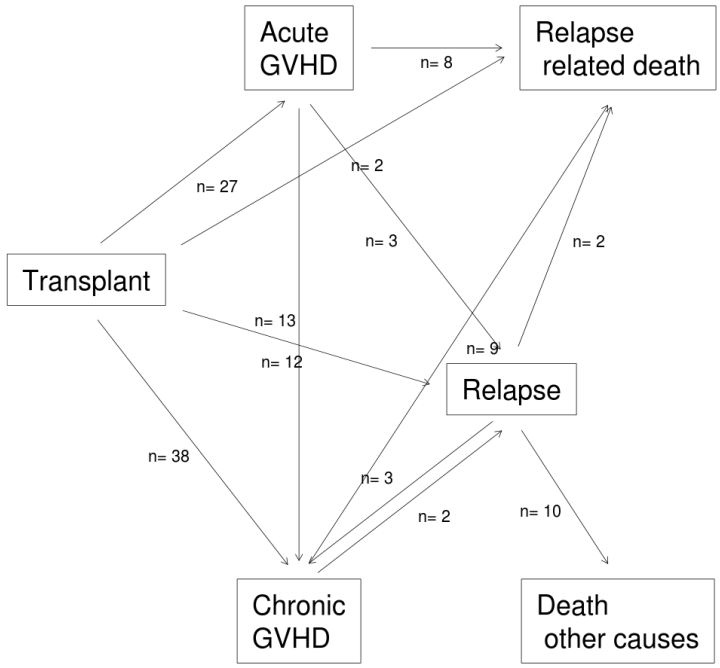


**Figure 4.5.** Multi-state model for the possible temporal courses of a patient after transplantation. The arrows describe the transitions from one state to another, and the arrow labels the number of patients with an observed transition.

a time-dependent confounder for the estimation of relapse hazards. Thus, paradoxically, in our analysis, the fact that the Cox model does not converge, makes the corresponding gene interesting. Therefore the genotype-hazard associations were analyzed with the log-rank test.

In the light of the small sample sizes and small numbers of events we decided not to trust the asymptotic approximation of the log-rank statistic and to obtain exact p-values. For calculating an exact p-value for all genes, the associated responses for each patient-donor pair were permutated 10,000 times. The observed log-rank test statistics where then compared to the distribution of the permutated log-rank statistics which should resemble the null hypothesis. Note that with 10,000 permutations the smallest possible exact p-value is 0.0001.

For each time-to-event response and separately for nsSNP and the potential mHags, all genes were ranked according to the exact log-rank p-values. For the nsSNPs analyses and with a p-value $< 0.05$ there are 417 relapse-associated genes, 467 genes are associated with acute GVHD, and 418 genes

| Event | Genotype 0 | Genotype >0 |
|---|---|---|
| Censored | 57 | 0 |
| Relapse | 15 | 2 |
| Other event | 19 | 0 |
| Probability of relapse | 0.165 | 1 |

**Table 4.2. Example for calculating the relapse risk.** The fraction of patients that had experienced relapse among all the patient-donor pairs is calculated for both patient-donor pairs with and without predicted mHags. For the given gene, with an increased number of mHags differing between patient and donor, we would have an increased risk of relapse. Patients associated to Genotype 0 have no predicted mHag variation for the analyzed gene. Patients with Genotype $> 0$ have, for the gene in question, at least one predicted mHag difference to their donor.

are associated with chronic GVHD. The mHags analyses report 213 genes for relapse, 348 for acute GVHD, and 276 genes associated with chronic GVHD. Note that p-values are not corrected for multiple testing. After correction for multiple testing (Bonferroni-Holm), there are no significant genes. The numbers of significant genes for nsSNPs and mHags are not directly comparable, as genes with potential mHags are only a subset of the genes with nsSNPs and the number of associated patient-donor pairs associated per endpoint varies.

Genes can either be positively associated (i.e. the more nsSNPs or mHags the higher the risk of the event) or negatively associated (i.e. the less nsSNPs or mHags the higher the risk of the event) to the event hazard. Based on the distribution of observed events, genes were marked as either positively or negatively associated within each list. In the case of relapse and potential mHags, this was done by calculating the fraction of patients that had experienced relapse among all the patient-donor pairs with no predicted mHags in the specific gene. Likewise, the fraction of patients with relapse among all the patient-donor pairs with predicted mHags, were calculated. A higher fraction of relapse in patient-donor pairs with more than one mHag is equivalent to an increased risk of relapse with an increased number of mHags. Similarly, a decreased relapse rate with an increasing number of mHags is given if the fraction of patients with relapse is higher for patient-donor pairs with no difference in amount of mHags. An example is shown in Table 4.7.

Table 4.7 shows the number of genes with a p-value $< 0.05$ contributing either to increased or decreased hazard for each endpoint. Overall, for each group we identified more genes associated with increased risk as compared to decreased risk. An exception is the association of relapse with having more nsSNPs. Here, we found more genes to be associated with decreased risk. Assigning all 9,162 genes with nsSNP variation as well as all 6,359

| Type | Event | Associated risk | Genes p-val < 0.05 |
|---|---|---|---|
| nsSNPs | relapse | increased | 122 |
| | | decreased | 295 |
| | aGVHD | increased | 369 |
| | | decreased | 98 |
| | cGVHD | increased | 306 |
| | | decreased | 112 |
| mHags | relapse | increased | 176 |
| | | decreased | 37 |
| | aGVHD | increased | 337 |
| | | decreased | 11 |
| | cGVHD | increased | 225 |
| | | decreased | 51 |

**Table 4.3. Classification based on risk.** For each endpoint, the number of genes with a p-value $< 0.05$, associated with either an increased or decreased risk, is given. P-values are not corrected for multiple comparisons and group sizes are not directly comparable.

genes with mHag variation in both cases resulted in a 2-2.5 times larger set of genes associated with decreased risk. The data shows that within the set of genes possibly associated with relapse, there is a higher likelihood for the nsSNPs set to be significantly associated then compared to the mHags set. Per analyzed endpoint, the top 10 ranking genes, where an observed difference in number of mHags is positively or negatively associated with the respective endpoint, are shown in Tables 4.7, 4.7 and 4.7. None of the p-values are significant after correction for multiple testing (Bonferroni-Holm). Interestingly, *AKAP13* is the top three ranked gene in the ranked gene-list associated with decreased relapse. The gene is a known source of mHags [105]. *AKAP13* has, due to its length, many possible sources for SNP mismatches. For other genes with known mHags, as listed in Table 4.1, we could not see any clustering of these within our ranked gene-lists. The top 10 ranked genes per endpoint, identified by analyzing nsSNP disparities, are listed in the Appendix.

The cumulative incidence over 5 years, for the top ranked gene based on the mHags analysis, are shown in Figure 4.6 and Figure 4.7. Per plot, two

| | | | P-vaue | | # patient-donor pairs: All pairs | | # patient-donor pairs: Associated with endpoint | |
|---|---|---|---|---|---|---|---|---|
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | ENSG00000197386 | HTT | 0.0036 | 0.0038 | 66 | 27 | 17 | 0 |
| | ENSG00000205409 | OR52E6 | 0.0041 | 0.0050 | 66 | 27 | 17 | 0 |
| | ENSG00000170776 | AKAP13 | 0.0076 | 0.0074 | 54 | 39 | 15 | 2 |
| | ENSG00000180917 | FTSJD1 | 0.0081 | 0.0072 | 68 | 25 | 17 | 0 |
| | ENSG00000138182 | KIF20B | 0.0121 | 0.0144 | 72 | 21 | 17 | 0 |
| | ENSG00000036448 | MYOM2 | 0.0166 | 0.0158 | 64 | 29 | 16 | 1 |
| | ENSG00000108375 | RNF43 | 0.0181 | 0.0165 | 58 | 35 | 15 | 2 |
| | ENSG00000183423 | LRIT3 | 0.0185 | 0.0188 | 64 | 29 | 16 | 1 |
| | ENSG00000205744 | DENND1C | 0.0194 | 0.0176 | 64 | 29 | 16 | 1 |
| | ENSG00000161996 | WDR90 | 0.0224 | 0.0206 | 64 | 29 | 16 | 1 |
| increased | ENSG00000155629 | PIK3AP1 | 0.0014 | 0.0004 | 81 | 12 | 11 | 6 |
| | ENSG00000143164 | DCAF6 | 0.0017 | 0.0000 | 87 | 6 | 13 | 4 |
| | ENSG00000149633 | KIAA1755 | 0.0017 | 0.0012 | 70 | 23 | 8 | 9 |
| | ENSG00000130377 | ACSBG2 | 0.0032 | 0.0004 | 86 | 7 | 13 | 4 |
| | ENSG00000181961 | OR4A16 | 0.0032 | 0.0031 | 69 | 24 | 8 | 9 |
| | ENSG00000117507 | FMO6P | 0.0037 | 0.0026 | 76 | 17 | 10 | 7 |
| | ENSG00000139767 | SRRM4 | 0.0049 | 0.0020 | 80 | 13 | 11 | 6 |
| | ENSG00000176510 | OR10AC1P | 0.0049 | 0.0038 | 62 | 31 | 6 | 11 |
| | ENSG00000175489 | LRRC25 | 0.0052 | 0.0007 | 86 | 7 | 13 | 4 |
| | ENSG00000153246 | PLA2R1 | 0.0053 | 0.0006 | 86 | 7 | 13 | 4 |

**Table 4.4. Top ranked genes associated with relapse.** 17 patients experienced a relapse. Analysis is based on predicted mHag differences between patient and donor. The top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased likelihood of relapse. A decreased association is given when fewer predicted mHags are associated with a higher likelihood of relapse. The ranked lists are sorted based on a permutation test (10,000 permutations). Log-rank p-values cannot be compared directly due to a varying number of patient-donor pairs associated with the number of mHag disparities. Shown p-values are not corrected for multiple comparisons. The number of patient-donor pairs, with and without difference in predicted mHags, is listed for all pairs, as well as for pairs associated with relapse.

| | | | P-vaue | | # patient-donor pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | All pairs | | Associated with endpoint | |
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | ENSG00000163092 | XIRP2 | 0.0077 | 0.0048 | 75 | 18 | 27 | 0 |
| | ENSG00000162621 | LRRC53 | 0.0147 | 0.0120 | 78 | 15 | 27 | 0 |
| | ENSG00000165512 | ZNF22 | 0.0163 | 0.0169 | 70 | 23 | 25 | 2 |
| | ENSG00000176208 | ATAD5 | 0.0224 | 0.0215 | 75 | 18 | 26 | 1 |
| | ENSG00000139767 | SRRM4 | 0.0274 | 0.0211 | 80 | 13 | 27 | 0 |
| | ENSG00000197213 | ZSCAN5B | 0.0283 | 0.0278 | 76 | 17 | 26 | 1 |
| | ENSG00000154162 | CDH12 | 0.0334 | 0.0337 | 77 | 16 | 26 | 1 |
| | ENSG00000185627 | PSMD13 | 0.0347 | 0.0354 | 77 | 16 | 26 | 1 |
| | ENSG00000141141 | DDX52 | 0.0407 | 0.0455 | 74 | 19 | 25 | 2 |
| | ENSG00000127507 | EMR2 | 0.0452 | 0.0524 | 70 | 23 | 24 | 3 |
| increased | ENSG00000140093 | SERPINA10 | 0.0001 | 0.0000 | 82 | 11 | 19 | 8 |
| | ENSG00000163635 | ATXN7 | 0.0001 | 0.0001 | 76 | 17 | 16 | 11 |
| | ENSG00000140577 | CRTC3 | 0.0002 | 0.0000 | 84 | 9 | 20 | 7 |
| | ENSG00000172572 | PDE3A | 0.0002 | 0.0000 | 88 | 5 | 22 | 5 |
| | ENSG00000114423 | CBLB | 0.0003 | 0.0000 | 89 | 4 | 23 | 4 |
| | ENSG00000114790 | ARHGEF26 | 0.0003 | 0.0000 | 90 | 3 | 24 | 3 |
| | ENSG00000101412 | E2F1 | 0.0004 | 0.0000 | 86 | 7 | 21 | 6 |
| | ENSG00000175164 | ABO | 0.0008 | 0.0006 | 65 | 28 | 13 | 14 |
| | ENSG00000213780 | GTF2H4 | 0.0008 | 0.0000 | 87 | 6 | 22 | 5 |
| | ENSG00000091542 | ALKBH5 | 0.0009 | 0.0004 | 73 | 20 | 15 | 12 |

**Table 4.5. Top ranked genes associated with acute GVHD.** 27 patients developed acute GVHD. The top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased risk of acute GVHD.

| | | | P-vaue | | # patient-donor pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | All pairs | | Associated with endpoint | |
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | ENSG00000064205 | WISP2 | 0.0008 | 0.0012 | 70 | 23 | 48 | 6 |
| | ENSG00000117400 | MPL | 0.0028 | 0.0000 | 91 | 2 | 53 | 1 |
| | ENSG00000134827 | TCN1 | 0.0044 | 0.0010 | 84 | 9 | 54 | 0 |
| | ENSG00000178795 | GDPD4 | 0.0102 | 0.0074 | 79 | 14 | 50 | 4 |
| | ENSG00000177045 | SIX5 | 0.0106 | 0.0070 | 76 | 17 | 49 | 5 |
| | ENSG00000120314 | WDR55 | 0.0108 | 0.0073 | 81 | 12 | 51 | 3 |
| | ENSG00000178163 | ZNF518B | 0.0130 | 0.0083 | 83 | 10 | 53 | 1 |
| | ENSG00000124782 | RREB1 | 0.0166 | 0.0001 | 91 | 2 | 53 | 1 |
| | ENSG00000054654 | SYNE2 | 0.0184 | 0.0166 | 66 | 27 | 41 | 13 |
| | ENSG00000158488 | CD1E | 0.0203 | 0.0157 | 81 | 12 | 50 | 4 |
| increased | ENSG00000078018 | MAP2 | 0.0004 | 0.0000 | 87 | 6 | 48 | 6 |
| | ENSG00000182612 | TSPAN10 | 0.0005 | 0.0005 | 69 | 24 | 33 | 21 |
| | ENSG00000177990 | DPY19L2 | 0.0011 | 0.0000 | 89 | 4 | 51 | 3 |
| | ENSG00000126838 | PZP | 0.0012 | 0.0013 | 63 | 30 | 32 | 22 |
| | ENSG00000176029 | C11ORF16 | 0.0012 | 0.0004 | 78 | 15 | 41 | 13 |
| | ENSG00000141441 | FAM59A | 0.0015 | 0.0000 | 91 | 2 | 52 | 2 |
| | ENSG00000189196 | RP11-129B22.2 | 0.0015 | 0.0000 | 89 | 4 | 50 | 4 |
| | ENSG00000188771 | C11ORF34 | 0.0018 | 0.0000 | 86 | 7 | 48 | 6 |
| | ENSG00000113249 | HAVCR1 | 0.0019 | 0.0000 | 91 | 2 | 52 | 2 |
| | ENSG00000184838 | PRR16 | 0.0019 | 0.0001 | 86 | 7 | 47 | 7 |

**Table 4.6. Top ranked genes associated with chronic GVHD.** 27 patients developed chronic GVHD. The top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased risk of chronic GVHD.

cumulative incidence curves are shown: one for patient-donor pairs with no mHag disparities for the given gene, and one for patient-donor pairs with one or more mHag disparities for the gene in question. For the three analyzed endpoints, the cumulative incidences, for the top ranked genes positively associated with the risks, are shown in Figure 4.6. Here, patient-donor pairs with mHags disparity have an increased risk of the event. Cumulative incidence plots for the top genes associated with decreased risk are shown in Figure 4.7. Patient-donor pairs with mHag disparity have a decreased risk of the respective event.

## 4.8 Overlap analysis

Based on our ranked gene lists, higher ranked genes are more likely correlated with the respective outcome. One obvious question is, if the same sets of genes are shown to be involved in different endpoints. To investigate this we analyzed the observed overlap of the top ranked genes from each list. The expected overlap and standard deviation between two groups was calculated by random sampling (10,000 permutations) from the respective groups. Observed overlaps as well as expected overlaps between the 300 top ranked genes of all groups are shown in Figure 4.8. This cutoff was chosen, as observed numbers are easy to interpret. The results do not change if other cutoffs, such as genes with a p-value $< 0.05$ or the top 5% of each ranked gene list, are used. Data shown in Figure 4.8 illustrates that, depending on the endpoint, some gene sets have more and some less genes in common that expected from random sets of equal size. As expected, the top 300 ranked genes for the same endpoints, but associated with either with nsSNPs or mHags differences, always share more genes as compared to random. Out of 300 genes, for all 6 endpoints there is an overlap between 68 and 89 genes. While this overlap is larger than random, it shows that analyzing predicted mHags based on nsSNPs, instead of analyzing nsSNPs directly, does result in varying gene lists. Due to the design of the study, there are no overlapping genes for each comparison, where the endpoint is only differing by the association of the genes to an increased or decreased likelihood of the respective event. Of special interest, with possible further applications for therapy, is the identification of genes uniquely associated with a GVT effect. Chronic GVHD and GVT often occur together and an actual overlap could be expected. Considering our analysis based on nsSNPs, we found a significant lower overlap between decreased relapse (GVT) and increased acute GVHD. This signal is, however, lost, if we take the respective overlap of ranked genes based on mHag differences. Based on our analysis of mHag differences for the top 300 ranked genes, there are two endpoints with a significant (mean $\pm$ 2$\times$ SD) overlap. First, decreased relapse has more than expected genes in common with decreased acute GVHD. While a possible explanation seems far-fetched, there is a special interest in genes both associated with a GVT effect and a decrease of GVHD, as both have a favorable effect for a patient. The second significant correlation is a negative correlation, where the observed overlap

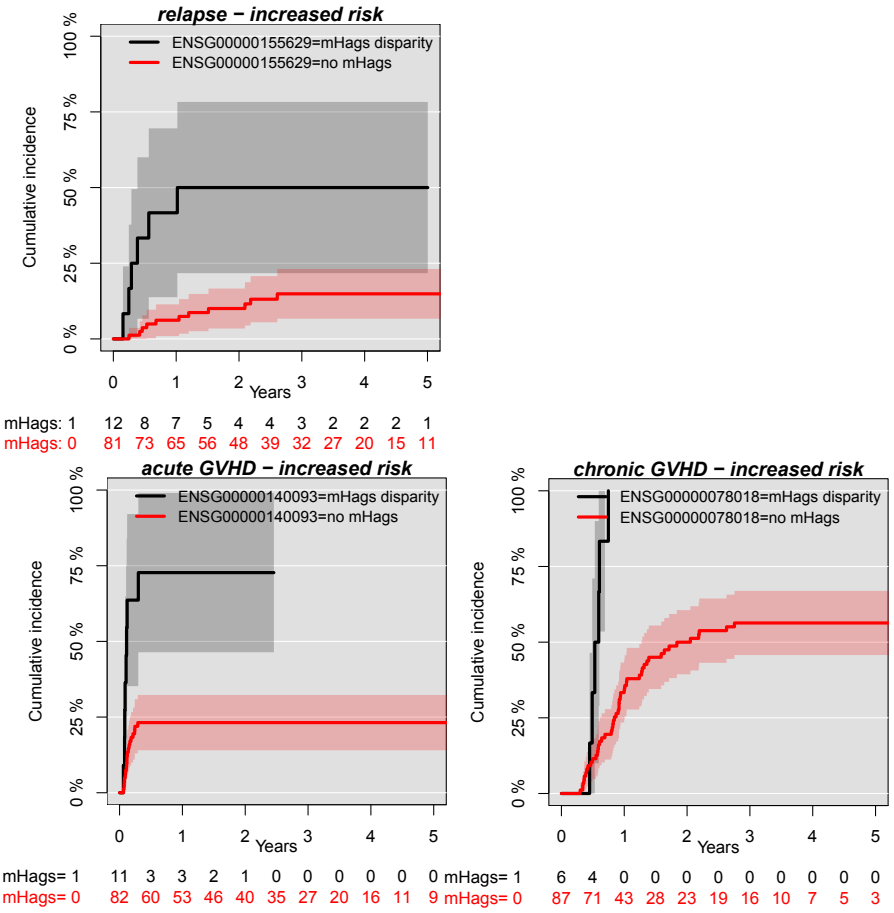


**Figure 4.6. Cumulative incidences for positively associated event hazards.** The cumulative incidence is the probability that the given event has occured before a given time. Plots are based on the top ranked genes of the respective gene list. The more mHags, the higher the risk of the given events. The plot "relapse - increased risk" is interpreted as follows: 81 patient-donor pairs had the same genotype for the given gene, while 12 patient-donor pairs had a mHag disparity of at least one mHag. The 12 patients with mHag disparity to their donor have a higher hazard of relapse as compared to patients with no mHag disparity. The shaded area is the confidence interval.

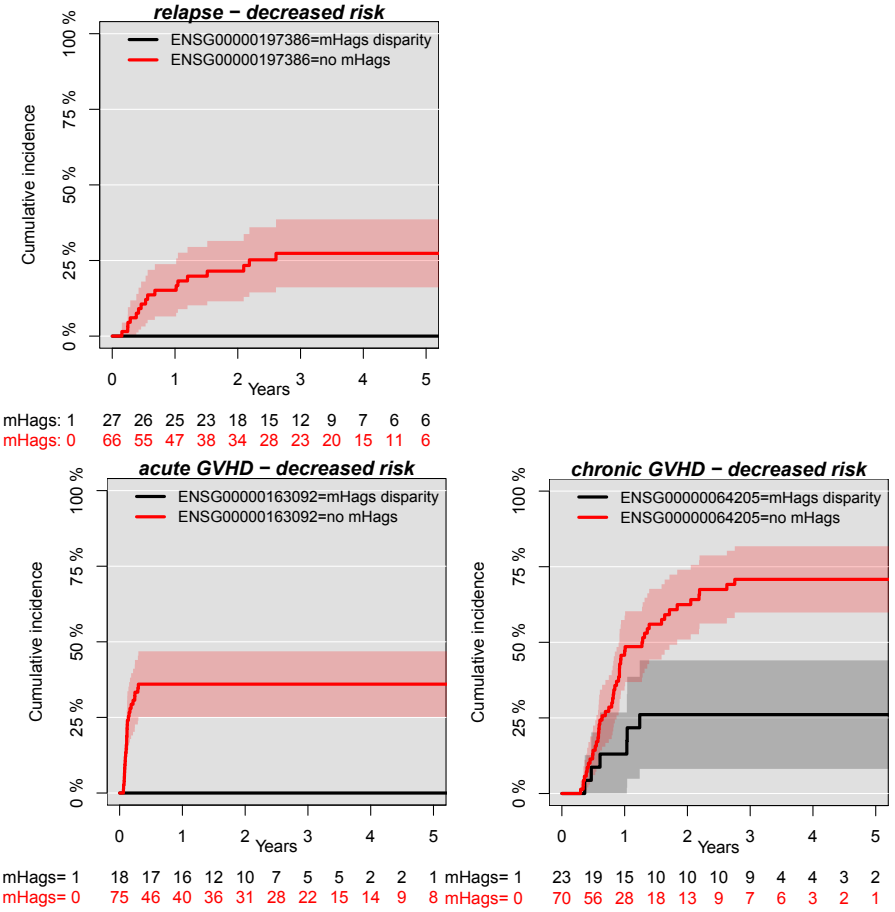


**Figure 4.7. Cumulative incidences for negatively associated event hazards.** The plots show the cumulative incidences, for the respective top ranked gene per genelist associated with mHags disparities, for the first five years. The plot "relapse - decreased risk" is interpreted as follows: 66 patients without mHag disparity to their respective donors have a higher hazard of relapse as compared to patient-donor pairs with mHag disparity. In other words, patient-donor pairs with mHag disparity have a lower likelihood (decreased risk) of relapse as pairs with the same genotype.

is smaller than the expected overlap. Here, genes associated with increased relapse are shown to have a smaller overlap than expected to genes associated with increased chronic GVHD. While both effects are unwanted, it is difficult to explain, why a gene would be associated with an increased possibility of relapse. In relation to relapse, mHags could only be associated with a decreased risk of relapse, as these mHags are playing a role in a possible GVT effect. A SNP difference may however up or down regulate the expression of gene or affect a proteins function. This change could further lead to an increased likelihood of relapse. Overall, our analysis seems to indicate, that there is no particular tendency of genes being positively or negatively correlated to either acute or chronic GVHD. On the contrary, for relapse, there are usually either more or less genes, than compared to random, associated with both acute and chronic GVHD.

## 4.9 Tissue expression analysis

Acute GVHD is not damaging all of a patient's tissues, it primarily occurs in liver, skin and the gastrointestinal tract [19]. While chronic GVHD is attacking these tissues too, it is shown to be more widespread than aGVHD. Favorable mHags contributing to a Graft vs. Leukemia effect, on the other hand, are largely limited to hematopoietic tissues. One of the challenges is the segregation of potential GVT effects from GVHD. Minor histocompatibility antigens with expression largely limited to hematopoietic tissues are shown to be able to treat hematologic malignant diseases without GVHD [96]. Based on this, we hypothesized that genes associated with the GVT effect are more commonly expressed in hematopoietic tissues. Genes associated with acute and chronic GVHD, while they might be expressed in hematopoietic tissues too, might be highly expressed in tissues such as liver and skin. In order to investigate these assumptions, we analyzed our gene lists for differences in tissue expression.

In order to test these assumptions, we compared the mRNA expressions of the top ranking genes of our analyzed endpoints. First, we analyzed differences in mRNA expression for hematopoietic tissues, and liver and skin. This analysis, based on the GNF gene expression database [109]. A variation in tissue-specific expression for the different endpoints was expected, we could however not identify any variations in mRNA expression (see Appendix). An alternative approach to investigate tissue expression is based on the Human Protein Reference Database (HPRD) [44]. The database (release 9) covers protein expression for 578 different tissue types, extracted from literature. The 300 top ranked genes associated per endpoint were compared against a pool of 300 random sampled genes. The random sampling was done 10,000 times; the genes were sampled out of all 9,162 genes associated with nsSNP differences. The number of genes associated with each tissue type was calculated per analyzed endpoint. Per tissue, an exact p-value was calculated based on how many times more genes were associated with the respective tissue, as compared to the 10,000 random permutations. Tissues significantly

| | | | mHags | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | relapse | | | | acute GVHD | | | | chronic GVHD | |
| | | | decreased | | increased | | decreased | | increased | | decreased | increased |
| mHags | relapse | decreased | | | | | | | | | | |
| | | increased | noOverlap | | | | | | | | | |
| | acute GVHD | decreased | ***38*** | *[13.81 ±3.48]* | **15** | [12.01 ±3.12] | | | | | | |
| | | increased | **9** | [13.17 ±3.38] | **5** | [11.91 ±3.16] | noOverlap | | | | | |
| | chronic GVHD | decreased | **12** | [12.03 ±3.28] | **20** | [16.64 ±3.65] | **20** | [13.46 ±3.42] | **12** | [13.94 ±3.50] | | |
| | | increased | **14** | [16.37 ±3.82] | ***4*** | *[11.74 ±3.09]* | **10** | [14.01 ±3.52] | **10** | [12.97 ±3.36] | noOverlap | |
| SNPs | relapse | decreased | ***80*** | *[11.31 ±3.21]* | **3** | [7.29 ±2.46] | **15** | [11.05 ±3.11] | **14** | [10.36 ±3.01] | **8** [8.35 ±2.71] | ***17*** *[10.57 ±3.04]* |
| | | increased | **0** | [5.45 ±2.24] | ***67*** | *[17.89 ±3.78]* | **5** | [8.90 ±2.79] | **6** | [10.12 ±2.98] | **9** [10.51 ±3.02] | **4** [8.77 ±2.77] |
| | acute GVHD | decreased | ***14*** | *[8.27 ±2.76]* | ***16*** | *[9.32 ±2.78]* | ***68*** | *[14.05 ±3.49]* | ***0*** | *[6.10 ±2.35]* | ***17*** *[10.48 ±3.02]* | **7** [9.09 ±2.80] |
| | | increased | ***2*** | *[9.56 ±2.92]* | **7** | [9.75 ±2.79] | **2** | [6.18 ±2.36] | ***89*** | *[13.63 ±3.43]* | **11** [9.33 ±2.86] | **8** [9.78 ±2.86] |
| | chronic GVHD | decreased | **9** | [9.18 ±2.88] | **14** | [9.74 ±2.84] | **8** | [8.79 ±2.75] | **13** | [9.73 ±2.94] | ***70*** *[14.24 ±3.49]* | **2** [4.70 ±2.03] |
| | | increased | **13** | [10.82 ±3.07] | **5** | [8.56 ±2.67] | **8** | [10.94 ±3.06] | **10** | [10.95 ±3.09] | **2** [5.87 ±2.32] | ***68*** *[12.31 ±3.22]* |

| | | | SNPs | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | relapse | | | | acute GVHD | | | | chronic GVHD | |
| | | | decreased | | increased | | decreased | | increased | | decreased | increased |
| mHags | relapse | decreased | | | | | | | | | | |
| | | increased | | | | | | | | | | |
| | acute GVHD | decreased | | | | | | | | | | |
| | | increased | | | | | | | | | | |
| | chronic GVHD | decreased | | | | | | | | | | |
| | | increased | | | | | | | | | | |
| SNPs | relapse | decreased | | | | | | | | | | |
| | | increased | noOverlap | | | | | | | | | |
| | acute GVHD | decreased | **12** | [9.58 ±2.96] | **13** | [11.11 ±3.09] | | | | | | |
| | | increased | ***2*** | *[10.23 ±3.02]* | **7** | [8.64 ±2.74] | noOverlap | | | | | |
| | chronic GVHD | decreased | **4** | [9.03 ±2.90] | **14** | [12.49 ±3.25] | **12** | 12 [9.84 ±2.97] | **9** | [9.63 ±2.94] | | |
| | | increased | **15** | [11.21 ±3.18] | ***2*** | *[8.92 ±2.79]* | **8** | 8 [9.01 ±2.88] | **7** | [9.09 ±2.88] | noOverlap | |

**Figure 4.8. Overlapping genes for all analyzed endpoints.** The comparisons are based on the top 300 ranked genes of the respective lists. The number of overlapping genes is given in bold. Calculated mean and standard deviation is based on a random sampling (10,000 permutations) from the corresponding groups. Observed overlaps larger or smaller than two times SD plus permutation based mean are underlined.

associated with decreased likelihood of relapse and increased risk of acute & chronic GVHD are shown in Figure 4.9. Observed differences are not significant after correction for multiple comparisons. For decreased risk of relapse (mHags analysis) and three hematopoietic tissues (hematopoietic stem cell, lymphoid stem cell, myeloid stem cell), we found the 300 top ranked genes to be overexpressed, as compared to random. While this is consistent with our expectations of GVT being associated with hematopoietic tissues, we could not find any expected correlations for other endpoints. Further, we found the top ranked genes correlated with unexpected tissues (as defined by HPRD database), such as hair and milk.

## 4.10 Conclusion

In this study we investigated a gene-specific association between the number of nsSNPs or predicted mHags and possible temporal courses of 93 patients following allo-HCT. All patients and their respective donor were genotype by a BeadChip Array. Per patient-donor pair, nsSNPs disparities in the GvH direction were identified and, based on these, mHags were predicted. The patient set is partly coherent with the set used in the study by Larsen et al. However, while we analyzed genome-wide differences, the study by Larsen et al. only analyzed 11 genes known to contain mHags.

By analyzing our data set in a similar way as to the study by Larsen, we could not reproduce their findings. This might be due to a significant smaller data set. Based on a diversification of observed nsSNP differences for related and unrelated donors, we had to further split our, from the beginning already smaller, dataset. Limiting our analysis to genes known to contain mHags resulted in a stronger signal than an analysis based on all genes covered by the array. This indicated that gene-specific distinctions for the effect of a nsSNP or mHag exist.

We defined three time-to-event responses, namely relapse, acute GVHD and chronic GVHD. For each of these temporal endpoints, an association for each gene covered by the analysis was done. Ranked gene-lists associated to each of these events, based on 10,000 permutations and log-rank tests were generated for observed nsSNP and mHag differences. Based on the distribution of observed events, genes were either positively or negatively associated to the event hazard. While the ranking of genes is correlated to their significance for each endpoint, no p-values were significant after correction for multiple testing.

An analysis of the overlap of the respective top ranked genes from each group was performed. For clinical application, genes associated with GVT (decreased relapse) and not associated with increased acute or chronic GVHD are of significant interest. Based on our analysis, such genes could be identified and they could be subsequent tested for potential mHags.

Tissue expression of mHags is believed to play a role for the GVT effect. These favorable mHags are expressed in hematopoietic tissues, as this enables them to support tumor eradication. In contrast, acute GVHD occurs

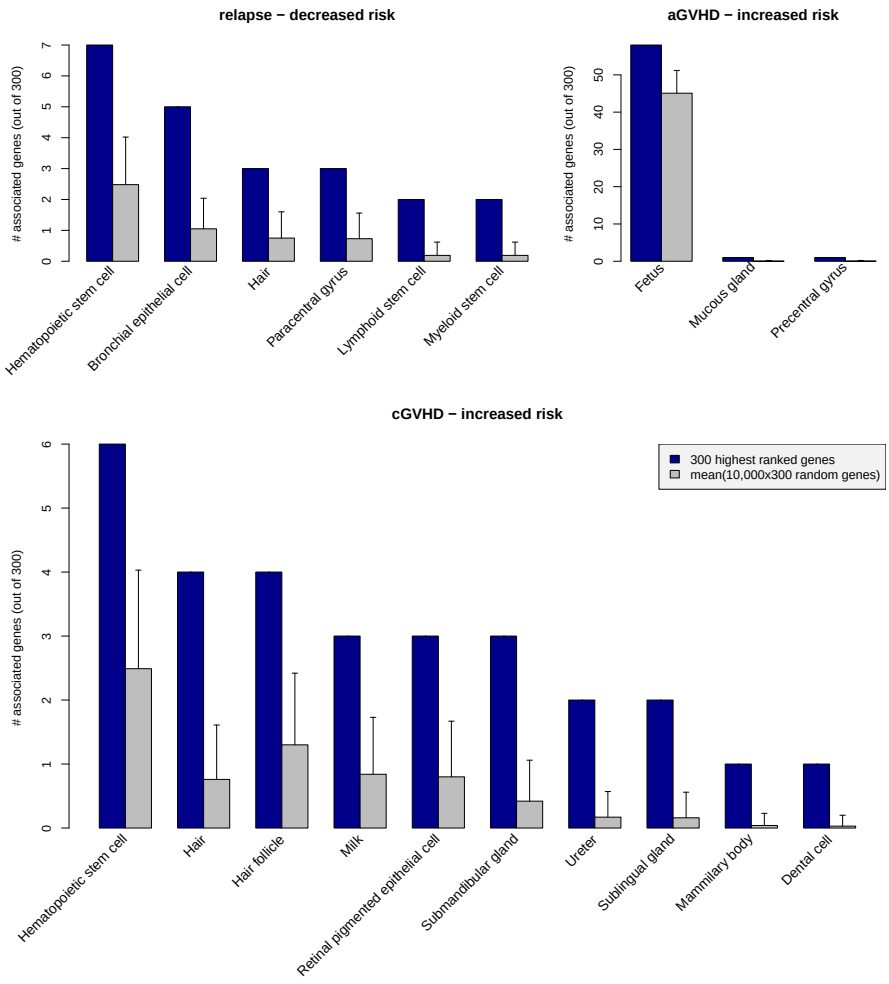


**Figure 4.9. Genes per tissue types.** Significant tissue types, where the 300 top ranked genes are overexpressed as compared to all analyzed genes, are shown. Gene lists associated with shown endpoints are based on the mHags analysis. Expression information is based on Human Protein Reference Database. Observed number of expressed genes (out of 300 top ranked) for the three shown endpoints is given by the blue bar. The expected number of genes per tissue type (grey bar) is defined by the mean of 10,000 random samples of 300 genes. Random sampling is based on all 9,162 genes with nsSNP differences. Data is not normalized. Error bars represent the SD of the random samples.

primarily in liver, skin and the gastrointestinal tract and chronic GVHD is shown to affect more tissues. For our top ranking genes, we investigated their overexpression in different tissue types. We could not find any signal based on mRNA expression data, but we could, based on HPRD database, identify a potential overexpression of genes associated with GVT in hematopoietic tissues.

Our ranked gene lists associated with event hazards are of potential importance when testing predicted mHags in the laboratory. Experimental verification of mHags is a time consuming process and a pre-selection of potential source-genes is desirable. While we could not identify significant p-values after correction for multiple testing, we believe that the ranking of genes based on observed p-values is still valuable. A larger data set would have enabled a more specific analysis of the different temporal endpoints. In our study, we had to deal with a relatively heterogeneous patient set of related and unrelated donors and different hematopoietic malignant diseases. The use of immunosuppression for treatment is another possible restriction. While grade IV acute GVHD is per definition fatal, grade II and III are treated by heavy immunosuppression. This would, however, also offset any GVT effect by the immune system. The other possible GVHD is chronic GVHD. It is assumed that mHags leading to chronic GVHD are associated with a GVT effect. Patients can have acute GVHD without subsequent chronic GVHD, or vice versa.

Each patient-donor pair has mHags with favorable as well as with undesirable effects. A mix of these beneficial and mHags results in conditions that are a mix of the above mentioned conditions. This makes it difficult to pinpoint specific genes being associated to specific temporal endpoints. Given a more homogenous data set or large enough sample size, this should, however, be possible. In this study we could identify, for each event, genes being more correlated than others genes. These findings are planned to be used for subsequent mHag identification at the Department of Hematology, Rigshospitalet in Copenhagen.

# Chapter 5

# Concluding remarks

In this thesis I have presented and discussed applications of cytotoxic T cell epitope predictions. First, I present a MHC class I pathway epitope predictor, *NetCTLpan*. Presently, we observe a growing interest in epitope based vaccine design, and at the same time high-throughput genome sequencing is becoming increasingly available [92]. Methods for experimental validation of T cell epitopes are time-consuming and allow only for testing of antigens that can be purified in large enough quantities [82]. Methods for the prediction of T cell epitopes are shown to minimize experimental efforts needed to identify new immunogenic peptides. These methods are part of the reverse immunology concept: after the selection of candidate genes, potential immunogenic peptides are predicted, with subsequent experimental validation. With the availability of high-throughput genome sequencing, the number of predicted epitopes can be huge. In this setting, *NetCTLpan* could be used to select a more promising subset of peptides for experimental validation.

While epitope prediction optimized for a low false positive rate is beneficial for the identification of new immunogenic peptides, reverse immunology approaches could also be applied, when classical approaches would be problematic due to difficulties in cultivating certain pathogens. Another application is presented in Part III of this thesis. The project utilizes *NetMHCpan* for the prediction of potential epitopes on a large scale. This predictor was used, as our goal was not the identification of new epitopes, but we were interested in observed epitope densities in different sets of proteins. We found that peptides unique to cancer splice variants comprise significantly fewer HLA class I epitopes compared to peptides unique to normal transcripts. While cancer is known to employ mechanisms to escape immune surveillance, such as dysregulating pathways and downregulating HLA expression, we presented a first report that carcinoma has a bias towards displaying fewer epitopes, which is initiated already at the step of mRNA splicing. A possible explanation could

be that the host´s immune system restricts the cancer exome. We found furthermore that hydrophilic amino acids are significantly enriched in the unique carcinoma sequence. This contributed to the lower likelihood of carcinoma-specific peptides to be predicted epitopes, as an increase in hydrophilic amino acids has a destabilizing effect on protein structure. In particular, the most common HLA class I allele, A*02:01, prefer hydrophobic amino acids at the anchor positions. Therefore, we expected fewer predicted epitopes with an increase of hydrophilic amino acids. This explanation does not exclude the possibility of the host´s immune system restricting cancer cells, by eradicating cancer cells that present new epitopes in complex with MHC. Treating cancer with epitope-based vaccines is a challenging field and despite many efforts, their use has not advanced beyond phase I or II clinical trials [119]. More knowledge of the immunology involved in cancer immunotherapy is needed; the presented work contributes to a better understanding of the mechanisms leading to cancer-specific epitopes.

The focus of the third project described in this thesis is on the role that nsSNPs and mHags play in transplantations. In the context of transplantations, mHags are epitopes unique to the patient. The donor´s transplanted immune cells classify these epitopes as foreign and mediate an immune response. This reactivity against patient cells is responsible for GVHD, a disease that, depending on the grade, requires pharmacologic immunosuppression. While GVHD is undesirable, some mHags may also lead to a beneficial GVT effect. Beneficial mHags were previously shown to be restricted to hematopoietic tissues. That is, as this enables the donor´s grafted immune cells to raise an immune reaction against diseased cells of the hematopoietic system of the host. GVHD is usually observed in liver, skin, the gastrointestinal tract, or in the connective tissue. Ideally it should be possible to identify mHags, which do not lead to GVHD, as they are restricted to hematopoietic tissue, and can aid in curing the patient for his malignancy, or at least lower relapse related mortality (RRM) and improve OS. The purpose of our study was to separate genes into those in which it is beneficial vs. deleterious to have nsSNPs or mHags. While a distinction of mHags into beneficial or undesirable effects seems reasonable, a classification of SNPs into these groups, based on an immunological response, is questionable. A SNP is only a possible target for the immune system, if it is part of a peptide that is presented in complex with MHC at the cell´s surface. While a SNP can be associated with disease progression due to other factors, such as regulation of pathways, we still expect to see a weak signal when looking at SNPs alone. That is, as a SNP in a gene that is preferentially expressed in hematopoietic tissue, may be part of a presented peptide associated with a GVT effect. On the other hand, the SNP might be indifferent. That is given, if the SNP is not part of a presented peptide. Adding predictions of which SNPs are potential mHags should strengthen the signal, since we remove indifferent SNPs.

A clear distinction of genes associated with beneficial or deleterious nsSNPs or mHags to temporal courses after HCT is valuable for subsequent experimental mHag identification. Based on our findings, we can give a tendentious association per gene, based on the number of observed nsSNP or

mHag disparities, to three temporal courses following HCT, namely relapse, acute GVHD and chronic GVHD.

All three studies taken together give new insights about the immune system and how it interplays with cancer, which brings us closer to an understanding of how epitopes could be used for cancer therapy.

# Bibliography


[1] David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Penelope E Bonnen, Paul I W de Bakker, Panos Deloukas, Stacey B Gabriel, Rhian Gwilliam, Sarah Hunt, Michael Inouye, Xiaoming Jia, Aarno Palotie, Melissa Parkin, Pamela Whittaker, Kyle Chang, Alicia Hawes, Lora R Lewis, Yanru Ren, David Wheeler, Donna Marie Muzny, Chris Barnes, Katayoon Darvishi, Matthew Hurles, Joshua M Korn, Kati Kristiansson, Charles Lee, Steven A McCarrol, James Nemesh, Alon Keinan, Stephen B Montgomery, Samuela Pollack, Alkes L Price, Nicole Soranzo, Claudia Gonzaga-Jauregui, Verneri Anttila, Wendy Brodeur, Mark J Daly, Stephen Leslie, Gil McVean, Loukas Moutsianas, Huy Nguyen, Qingrun Zhang, Mohammed J R Ghori, Ralph McGinnis, William McLaren, Fumihiko Takeuchi, Sharon R Grossman, Ilya Shlyakhter, Elizabeth B Hostetter, Pardis C Sabeti, Clement A Adebamowo, Morris W Foster, Deborah R Gordon, Julio Licinio, Maria Cristina Manca, Patricia A Marshall, Ichiro Matsuda, Duncan Ngare, Vivian Ota Wang, Deepa Reddy, Charles N Rotimi, Charmaine D Royal, Richard R Sharp, Changqing Zeng, Lisa D Brooks, and Jean E McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, September 2010. 3

[2] Y Altuvia and H Margalit. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. *Journal of molecular biology*, 295(4):879–90, January 2000. 15

[3] Mads Hald Andersen, David Schrama, Per Thor Straten, and Jürgen C Becker. Cytotoxic T cells. *The Journal of investigative dermatology*, 126(1):32–41, January 2006. 5

[4] Mads Hald Andersen, David Schrama, Per Thor Straten, and Jürgen C Becker. Cytotoxic T cells. *The Journal of investigative dermatology*, 126(1):32–41, January 2006. 7

[5] K S Anderson, J Alexander, M Wei, and P Cresswell. Intracellular transport of class I MHC molecules in antigen processing mutant cell lines. *Journal of immunology (Baltimore, Md. : 1950)*, 151(7):3407–19, October 1993. 16

[6] G Baumann, C Frömmel, and C Sander. Polarity as a criterion in protein design. *Protein engineering*, 2(5):329–34, January 1989. 38

[7] Marie Bleakley and Stanley R Riddell. Molecules and mechanisms of the graft-versus-leukaemia effect. *Nature reviews. Cancer*, 4(5):371–80, May 2004. 55

[8] Benedetto Bruno, Marcello Rotta, Francesca Patriarca, Nicola Mordini, Bernardino Allione, Fabrizio Carnevale-Schianca, Luisa Giaccone, Roberto Sorasio, Paola Omedè, Ileana Baldi, Sara Bringhen, Massimo Massaia, Massimo Aglietta, Alessandro Levis, Andrea Gallamini, Renato Fanin, Antonio Palumbo, Rainer Storb, Giovannino Ciccone, and Mario Boccadoro. A comparison of allografting with autografting for newly diagnosed myeloma. *The New England journal of medicine*, 356(11):1110–20, March 2007. 8

[9] V Brusic, P van Endert, J Zeleznikow, S Daniel, J Hammer, and N Petrovsky. A neural network model approach to the study of human TAP transporter. *In silico biology*, 1(2):109–21, January 1999. 16

[10] Esma Cecuk-Jelicić, Vesna Kerhin-Brkljacić, Zorana Grubic, and Boris Labar. [World's registry of bone marrow donors]. *Acta medica Croatica : časopis Hravatske akademije medicinskih znanosti*, 63(3):251–3, June 2009. 9

[11] R Ceppellini, P L Mattiuz, G Scudeller, and M Visetti. Experimental allotransplantation in man. I. The role of the HL-A system in different genetic combinations. *Transplantation proceedings*, 1(1):385–9, March 1969. 10

[12] IHGS Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, October 2004. 1

[13] Uniprot Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic acids research*, 37(Database issue):D169–74, January 2009. 17

[14] A Craiu, T Akopian, A Goldberg, and K L Rock. Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10850–5, September 1997. 8, 15

[15] FHC Crick. On protein synthesis. *Symp Soc Exp Biol*, (XII):139–163, 1958. 1

[16] Charles J David and James L Manley. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & development*, 24(21):2343–64, November 2010. 35

[17] Pierre Dönnes and Oliver Kohlbacher. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Science*, 14(8):2132–2140, 2005. 15

[18] I.A. Doytchinova, P. Guan, and D.R. Flower. EpiJen: a server for multistep T cell epitope prediction. *BMC bioinformatics*, 7(1):131, 2006. 15

[19] W R Drobyski, P Hari, C Keever-Taylor, R Komorowski, and W Grossman. Severe autologous GVHD after hematopoietic progenitor cell transplantation for multiple myeloma. *Bone marrow transplantation*, 43(2):169–77, January 2009. 69

[20] Aron C Eklund and Zoltan Szallasi. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome biology*, 9(2):R26, January 2008. 93

[21] T Elliott, A Willis, V Cerundolo, and A Townsend. Processing of major histocompatibility class I-restricted antigens in the endoplasmic reticulum. *The Journal of experimental medicine*, 181(4):1481–91, April 1995. 8

[22] Malene Erup Larsen, Henrik Kloverpris, Anette Stryhn, Catherine K Koofhethile, Stuart Sims, Thumbi Ndung'u, Philip Goulder, Sø ren Buus, and Morten Nielsen. HLArestrictor–a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides. *Immunogenetics*, 63(1):43–55, January 2011. 38

[23] Fang Fang, Xinhong He, Huiwen Deng, Qun Chen, Jianping Lu, Manfred Spraul, and Yihua Yu. Discrimination of metabolic profiles of pancreatic cancer from chronic pancreatitis by high-resolution magic angle spinning 1H nuclear magnetic resonance and principal components analysis. *Cancer science*, 98(11):1678–82, November 2007. 46

[24] Robert L Ferris, Theresa L Whiteside, and Soldano Ferrone. Immune escape associated with functional defects in antigen-processing machinery in head and neck cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 12(13):3890–5, July 2006. 35

[25] Meaghan A Figge and Lynda Blankenship. Missense mutations in the BRCT domain of BRCA-1 from high-risk women frequently perturb strongly hydrophobic amino acids conserved among mammals. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 13(6):1037–41, June 2004. 47

[26] Olivera J Finn. Cancer immunology. *The New England journal of medicine*, 358(25):2704–15, June 2008. 34

[27] Paul J Gardina, Tyson A Clark, Brian Shimada, Michelle K Staples, Qing Yang, James Veitch, Anthony Schweitzer, Tarif Awad, Charles Sugnet, Suzanne Dee, Christopher Davies, Alan Williams, and Yaron Turpaz. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC genomics*, 7:325, January 2006. 35, 43

[28] Bing Ge, Scott Gurd, Tiffany Gaudin, Carole Dore, Pierre Lepage, Eef Harmsen, Thomas J Hudson, and Tomi Pastinen. Survey of allelic expression using EST mining. *Genome research*, 15(11):1584–91, November 2005. 40

[29] H Goker, I C Haznedaroglu, and N J Chao. Acute graft-vs-host disease: pathobiology and management. *Experimental hematology*, 29(3):259–77, March 2001. 10

[30] 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. *The New England journal of medicine*, 334(5):281–5, March 1996. 10

[31] J Hakenberg, AK Nussbaum, H Schild, HG Rammensee, C Kuttler, HG Holzhütter, PM Kloetzel, SH Kaufmann, and HJ Mollenkopf. MAPPP: MHC class I antigenic peptide processing prediction. *Appl Bioinformatics*, 2(3):155–158, 2003. 15

[32] Lothar Hambach, Eric Spierings, and Els Goulmy. Risk assessment in haematopoietic stem cell transplantation: minor histocompatibility antigens. *Best practice & research. Clinical haematology*, 20(2):171–87, July 2007. 8

[33] Chunjiang He, Zhixiang Zuo, Hengling Chen, Liao Zhang, Fang Zhou, Hanhua Cheng, and Rongjia Zhou. Genome-wide detection of testis- and testicular cancer-specific alternative splicing. *Carcinogenesis*, 28(12):2484–90, December 2007. 35, 43

[34] R A Henderson, H Michel, K Sakaguchi, J Shabanowitz, E Appella, D F Hunt, and V H Engelhard. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science*, 255(5049):1264–6, March 1992. 16

[35] Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Sø ren Buus, and Morten Nielsen. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, January 2009. 16, 18, 26, 38, 52

[36] T P Hopp and K R Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3824–8, June 1981. 38

[37] K Imai, S Matsuyama, S Miyake, K Suga, and K Nakachi. Natural cytotoxic activity of peripheral-blood lymphocytes and cancer incidence: an 11-year follow-up study of a general population. *Lancet*, 356(9244):1795–9, November 2000. 34

[38] Agnieszka S. Juncker, Mette V. Larsen, Nils Weinhold, Morten Nielsen, Sø ren Brunak, and Ole Lund. Systematic Characterisation of Cellular Localisation and Expression Profiles of Proteins Containing MHC Ligands. *PLoS ONE*, 4(10):e7448, October 2009. 15, 93

[39] Christoph Kahl, Barry E Storer, Brenda M Sandmaier, Marco Mielcarek, Michael B Maris, Karl G Blume, Dietger Niederwieser, Thomas R Chauncey, Stephen J Forman, Edward Agura, Jose F Leis, Benedetto Bruno, Amelia Langston, Michael A Pulsipher, Peter A McSweeney, James C Wade, Elliot Epner, Finn Bo Petersen, Wolfgang A Bethge, David G Maloney, and Rainer Storb. Relapse risk in patients with malignant diseases given allogeneic hematopoietic cell transplantation after nonmyeloablative conditioning. *Blood*, 110(7):2744–8, October 2007. 56

[40] Chatchada Karanes, Gene O Nelson, Pintip Chitphakdithai, Edward Agura, Karen K Ballen, Charles D Bolan, David L Porter, Joseph P Uberti, Roberta J King, and Dennis L Confer. Twenty years of unrelated donor hematopoietic cell transplantation for adult recipients facilitated by the National Marrow Donor Program. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 14(9 Suppl):8–15, September 2008. 9

[41] S Katayama, Y Tomaru, T Kasukawa, K Waki, M Nakanishi, M Nakamura, H Nishida, C C Yap, M Suzuki, J Kawai, H Suzuki, P Carninci, Y Hayashizaki, C Wells, M Frith, T Ravasi, K C Pang, J Hallinan, J Mattick, D A Hume, L Lipovich, S Batalov, P G Engström, Y Mizuno, M A Faghihi, A Sandelin, A M Chalk, S Mottagui-Tabar, Z Liang, B Lenhard, and C Wahlestedt. Antisense transcription in the mammalian transcriptome. *Science (New York, N.Y.)*, 309(5740):1564–6, September 2005. 2

[42] W Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry*, 14:1–63, January 1959. 47

[43] Janet Kelso, Johann Visagie, Gregory Theiler, Alan Christoffels, Soraya Bardien, Damian Smedley, Darren Otgaar, Gary Greyling, C Victor Jongeneel, Mark I McCarthy, Tania Hide, and Winston Hide. eVOC: a controlled vocabulary for unifying gene expression data. *Genome research*, 13(6A):1222–30, June 2003. 35

[44] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 37(Database issue):D767–72, January 2009. 69

[45] Thomas J. Kindt, Barbara A. Osborne, and Richard A. Goldsby. *Kuby Immunology, Sixth Edition*. W. H. Freeman & Company, 2006. 6

[46] Joachim Koch, Renate Guntrum, Susanne Heintke, Christoph Kyritsis, and Robert Tampé. Functional dissection of the transmembrane domains of the transporter associated with antigen processing (TAP). *The Journal of biological chemistry*, 279(11):10142–7, March 2004. 15

[47] Hans-Jochem Kolb. Graft-versus-leukemia effects of transplantation and donor lymphocytes. *Blood*, 112(12):4371–83, December 2008. 11

[48] B Kornblit, T Masmas, H O Madsen, L P Ryder, A Svejgaard, B Jakobsen, H Sengelø v, G Olesen, C Heilmann, E Dickmeiss, S L Petersen, and L Vindelø v. Haematopoietic cell transplantation with non-myeloablative conditioning in Denmark: disease-specific outcome, complications and hospitalization requirements of the first 100 transplants. *Bone marrow transplantation*, 41(10):851–9, May 2008. 50

[49] Gautier Koscielny, Vincent Le Texier, Chellappa Gopalakrishnan, Vasudev Kumanduri, Jean-Jack Riethoven, Francesco Nardone, Eleanor Stanley, Christine Fallsehr, Oliver Hofmann, Meelis Kull, Eoghan Harrington, Stéphanie Boué, Eduardo Eyras, Mireya Plass, Fabrice Lopez, William Ritchie, Virginie Moucadel, Takeshi Ara, Heike Pospisil, Alexander Herrmann, Jens G Reich, Roderic Guigó, Peer Bork, Magnus Von Knebel Doeberitz, Jaak Vilo, Winston Hide, Rolf Apweiler, Thangavel Alphonse Thanaraj, and Daniel Gautheret. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93(3):213–20, 2009. 35

[50] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, and J Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, March 2001. 1, 4

[51] 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. Degree of predicted minor histocompatibility antigen mismatch correlates with poorer clinical outcomes in nonmyeloablative allogeneic hematopoietic cell transplantation. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 16(10):1370–81, October 2010. 52, 53, 54

[52] Mette V Larsen, Claus Lundegaard, Kasper Lamberth, Soren Buus, Ole Lund, and Morten Nielsen. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC bioinformatics*, 8:424, 2007. 15, 16, 17

[53] Mette Voldby Larsen, Claus Lundegaard, Kasper Lamberth, Sø ren Buus, Sø ren Brunak, Ole Lund, and Morten Nielsen. An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *European journal of immunology*, 35(8):2295–303, August 2005. 15, 16, 17, 28, 31

[54] Stephanie J Lee, Georgia Vogelsang, and Mary E D Flowers. Chronic graft-versus-host disease. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 9(4):215–33, May 2003. 11

[55] Frédéric Lévy, Lena Burri, Sandra Morel, Anne-Lise Peitrequin, Nicole Lévy, Angela Bachi, Ulf Hellman, Benoît J Van Den Eynde, and Catherine Servis. The final N-terminal trimming of a subaminoterminal proline-containing HLA class I-restricted antigenic peptide in the cytosol is mediated by two peptidases. *Journal of immunology (Baltimore, Md. : 1950)*, 169(8):4161–71, October 2002. 15

[56] C J Luckey, J A Marto, M Partridge, E Hall, F M White, J D Lippolis, J Shabanowitz, D F Hunt, and V H Engelhard. Differences in the expression of human class I MHC alleles and their associated peptides in the presence of proteasome inhibitors. *Journal of immunology (Baltimore, Md. : 1950)*, 167(3):1212–21, August 2001. 7

[57] Ole Lund, Morten Nielsen, Can Kesmir, Anders Gorm Petersen, Claus Lundegaard, Peder Worning, Christina Sylvester-Hvid, Kasper Lamberth, Gustav Rø der, Sune Justesen, Sø ren Buus, and Sø ren Brunak. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, 55(12):797–810, March 2004. 15, 17, 38

[58] Ole Lund, Morten Nielsen, Claus Lundegaard, Can Kesmir, and Sø ren Brunak. *Immunological bioinformatics*. MIT press, 1 ed. edition, 2005. 19

[59] Claus Lundegaard, Ole Lund, and Morten Nielsen. Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics (Oxford, England)*, 24(11):1397–8, June 2008. 18, 52

[60] Aidan MacNamara, Ulrich Kadolsky, Charles R M Bangham, and Becca Asquith. T-cell epitope prediction: rescaling can mask biological variation between MHC molecules. *PLoS computational biology*, 5(3):e1000327, March 2009. 16, 25

[61] S Miller, J Janin, A M Lesk, and C Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–56, August 1987. 38

[62] Noel F C C De Miranda, Maartje Nielsen, Dina Pereira, Marjo Van Puijenbroek, Hans F Vasen, Frederik J Hes, Tom Van Wezel, and Hans Morreau. MUTYH-associated polyposis carcinomas frequently lose HLA class I expression — a common event amongst DNA-repair-de cient colorectal cancers. *Journal of Pathology, The*, (April):69–76, 2009. 35

[63] X Y Mo, P Cascio, K Lemerise, A L Goldberg, and K Rock. Distinct proteolytic processes generate the C and N termini of MHC class I-binding peptides. *Journal of immunology (Baltimore, Md. : 1950)*, 163(11):5851–9, December 1999. 8, 15

[64] Maurizio Muscaritoli, Gabriella Grieco, Saveria Capria, Anna Paola Iori, and Filippo Rossi Fanelli. Nutritional and metabolic support in patients undergoing bone marrow transplantation. *Am J Clin Nutr*, 75(2):183–190, 2002. 9

[65] Tuna Mutis, Rob Verdijk, Ellen Schrama, Bennie Esendam, Anneke Brand, and Els Goulmy. Feasibility of Immunotherapy of Relapsed Leukemia With Ex Vivo-Generated Cytotoxic T Lymphocytes Specific for Hematopoietic System-Restricted Minor Histocompatibility Antigens. *Blood*, 93(7):2336–2341, 1999. 11

[66] T. E. Creighton N. J. Darby. *Protein structure*. 1993. 38

[67] Faustino NA and Cooper TA. Pre-mRNA splicing and human disease. *Genes Dev.*, 17(4):419–37, 2003. 35

[68] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Rø der, Bjoern Peters, Alessandro Sette, Ole Lund, and Sø ren 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, August 2007. 16, 17, 18, 38

[69] Morten Nielsen, Claus Lundegaard, Ole Lund, and Can Keşmir. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 57(1-2):33–41, April 2005. 16, 19, 29

[70] Morten Nielsen, Claus Lundegaard, Peder Worning, Christina Sylvester Hvid, Kasper Lamberth, Sø ren Buus, Sø ren Brunak, and Ole Lund. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics (Oxford, England)*, 20(9):1388–97, June 2004. 39

[71] Hå kan Norell, Mattias Carlsten, Tomas Ohlum, Karl-Johan Malmberg, Giuseppe Masucci, Kjell Schedvins, Wolfgang Altermann, Diana Handke, Derek Atkins, Barbara Seliger, and Rolf Kiessling. Frequent loss of HLA-A2 expression in metastasizing ovarian carcinomas associated with genomic haplotype loss and HLA-A2-restricted HER-2/neu-specific immunity. *Cancer research*, 66(12):6387–94, June 2006. 35

[72] Shinobu Ohnuma, Koh Miura, Akira Horii, Wataru Fujibuchi, Naoyuki Kaneko, Osamu Gotoh, Hideki Nagasaki, Takayuki Mizoi, Nobukazu Tsukamoto, Terutada Kobayashi, Makoto Kinouchi, Mitsunori Okabe, Hiroyuki Sasaki, Ken-ichi Shiiba, Kikuo Miyagawa, and Iwao Sasaki. Cancer-associated splicing variants of the CDCA1 and MSMB genes expressed in cancer cell lines and surgically resected gastric cancer tissues. *Surgery*, 145(1):57–68, January 2009. 35

[73] The International Agency for Research on Cancer. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissue (IARC WHO Classification of Tumours)*. World Health Organization, 2008. 9

[74] Edwin E. Osgood, Mathew C. Riddle, and Thomas J. Mathews. Aplastic anemia treated with daily transfusion and intravenous marrow. *Ann Intern Med*, 13(2):357–367, 1939. 8

[75] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–5, December 2008. 2

[76] P Paz, N Brouwenstijn, R Perry, and N Shastri. Discrete proteolytic intermediates in the MHC class I antigen processing pathway and MHC I-dependent peptide trimming in the ER. *Immunity*, 11(2):241–51, August 1999. 15

[77] Carina L Pérez, Mette V Larsen, Rasmus Gustafsson, Melissa M Norström, Ann Atlas, Douglas F Nixon, Morten Nielsen, Ole Lund, and Annika C Karlsson. Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *Journal of immunology (Baltimore, Md. : 1950)*, 180(7):5092–100, April 2008. 15, 26

[78] Björn Peters, Sascha Bulik, Robert Tampe, Peter M Van Endert, and Hermann-Georg Holzhütter. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *Journal of immunology (Baltimore, Md. : 1950)*, 171(4):1741–9, August 2003. 8, 16, 18

[79] 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 transplantation*, 15(6):825–8, July 1995. 11

[80] H Rammensee, J Bachmann, N P Emmerich, O a Bachor, and S Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–9, November 1999. 15, 16

[81] Xiangyu Rao, Ana Isabel C A Fontaine Costa, Debbie van Baarle, and Can Kesmir. A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant CD8+ T cell responses. *Journal of immunology (Baltimore, Md. : 1950)*, 182(3):1526–32, February 2009. 20, 38

[82] R Rappuoli. Reverse vaccinology. *Current opinion in microbiology*, 3(5):445–50, October 2000. 75

[83] Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta, Marcelo Bertalan, Kasper Nielsen, M Thomas P Gilbert, Yong Wang, Maanasa Raghavan, Paula F Campos, Hanne Munkholm Kamp, Andrew S Wilson, Andrew Gledhill, Silvana Tridico, Michael Bunce, Eline D Lorenzen, Jonas Binladen, Xiaosen Guo, Jing Zhao, Xiuqing Zhang, Hao Zhang, Zhuo Li, Minfeng Chen, Ludovic Orlando, Karsten Kristiansen, Mads Bak, Niels Tommerup, Christian Bendixen, Tracey L Pierre, Bjarne Grø nnow, Morten Meldgaard, Claus Andreasen, Sardana A Fedorova, Ludmila P Osipova, Thomas F G Higham, Christopher Bronk Ramsey, Thomas V O Hansen, Finn C Nielsen, Michael H Crawford, Sø ren Brunak, Thomas Sicheritz-Pontén, Richard Villems, Rasmus Nielsen, Anders Krogh, Jun Wang, and Eske Willerslev. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–62, February 2010. 4

[84] F M Richards. Areas, volumes, packing and protein structure. *Annual review of biophysics and bioengineering*, 6:151–76, January 1977. 38

[85] Olle Ringdén, Helen Karlsson, Richard Olsson, Brigitta Omazic, and Michael Uhlin. The allogeneic graft-versus-cancer effect. *British journal of haematology*, 147(5):614–33, December 2009. 11

[86] Daniel Rios, William M McLaren, Yuan Chen, Ewan Birney, Arne Stabenau, Paul Flicek, and Fiona Cunningham. A database and API for variation, dense genotyping and resequencing data. *BMC bioinformatics*, 11:238, January 2010. 51

[87] U Ritz and B Seliger. The transporter associated with antigen processing (TAP): structural integrity, expression, function, and its clinical relevance. *Molecular medicine (Cambridge, Mass.)*, 7(3):149–58, March 2001. 15

[88] Nathan J. Robertson, Jian-Guo Chai, Maggie Millrain, Diane Scott, Fazila Hashim, Emily Manktelow, Francois Lemonnier, Elizabeth Simpson, and Julian Dyson. Natural Regulation of Immunity to Minor Histocompatibility Antigens. *J. Immunol.*, 178(6):3558–3565, 2007. 10

[89] K L Rock and A L Goldberg. Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annual review of immunology*, 17:739–79, January 1999. 7

[90] Jose María Romero, Pilar Jiménez, Teresa Cabrera, José Manuel Cózar, Susana Pedrinaci, Miguel Tallada, Federico Garrido, and Francisco Ruiz-Cabello. Coordinated downregulation of the antigen presentation machinery and HLA class I/beta2-microglobulin complex is responsible for HLA-ABC loss in bladder cancer. *International journal of cancer. Journal international du cancer*, 113(4):605–10, February 2005. 35

[91] D C Roopenian. What are minor histocompatibility loci? A new look at an old question. *Immunology today*, 13(1):7–10, January 1992. 10

[92] Daniela Santoro Rosa, Susan Pereira Ribeiro, and Edecio Cunha-Neto. CD4+ T cell epitope discovery and rational vaccine design. *Archivum immunologiae et therapiae experimentalis*, 58(2):121–30, April 2010. 75

[93] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael Dicuccio, Scott Federhen, Michael Feolo, Ian M Fingerman, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Donna R Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D Pruitt, Gregory D Schuler, Edwin Sequeira, Stephen T Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A Tatusova, Lukas Wagner, Yanli Wang, W John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 39(Database issue):D38–51, November 2010. 39, 40

[94] Mark M Schatz, Björn Peters, Nadja Akkad, Nina Ullrich, Alejandra Nacarino Martinez, Oliver Carroll, Sascha Bulik, Hans-Georg Rammensee, Peter van Endert, Hermann-Georg Holzhütter, Stefan Tenzer, and Hansjörg Schild. Characterizing the N-terminal processing motif of MHC class I ligands. *Journal of immunology (Baltimore, Md. : 1950)*, 180(5):3210–7, March 2008. 15

[95] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–100, October 1990. 39, 46

[96] Mikkael A. Sekeres, Matt Kalaycio, and Brian J. Bolwell. *Clinical malignant hematology*. McGraw-Hill Prof Med/Tech, 2007. 69

[97] Barbara Seliger. Different regulation of MHC class I antigen processing components in human tumors. *Journal of immunotoxicology*, 5(4):361–7, October 2008. 35

[98] John Sidney, Bjoern Peters, Nicole Frahm, Christian Brander, and Alessandro Sette. HLA class I supertypes: a revised and updated classification. *BMC immunology*, 9:1, January 2008. 17

[99] E Simpson, D Scott, E James, G Lombardi, K Cwynarski, F Dazzi, M Millrain, and P J Dyson. Minor H antigens: genes and peptides. *Transplant immunology*, 10(2-3):115–23, August 2002. 10

[100] Richard J. Simpson. *Proteins and Proteomics: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 2002. 38

[101] Rolf I Skotheim and Matthias Nees. Alternative splicing in cancer: noise, functional, or systematic? *The international journal of biochemistry & cell biology*, 39(7-8):1432–49, January 2007. 34

[102] K D Smith and C T Lutz. Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. *Journal of immunology (Baltimore, Md. : 1950)*, 156(10):3755–64, May 1996. 16

[103] H L Snyder, J W Yewdell, and J R Bennink. Trimming of antigenic peptides in an early secretory compartment. *The Journal of experimental medicine*, 180(6):2389–94, December 1994. 8

[104] Charles T Spencer, Pavlo Gilchuk, Srdjan M Dragovic, and Sebastian Joyce. Minor histocompatibility antigens: presentation principles, recognition logic and the potential for a healing hand. *Current opinion in organ transplantation*, 15(4):512–25, August 2010. 53

[105] Eric Spierings, Anthony G Brickner, Jennifer A Caldwell, Suzanne Zegveld, Nia Tatsis, Els Blokland, Jos Pool, Richard A Pierce, Sahana Mollah, Jeffrey Shabanowitz, Laurence C Eisenlohr, Peter van Veelen, Ferry Ossendorp, Donald F Hunt, Els Goulmy, and Victor 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, July 2003. 62

[106] Eric Spierings, Jos Drabbels, Matthijs Hendriks, Jos Pool, Marijke Spruyt-Gerritse, Frans Claas, and Els Goulmy. A uniform genomic minor histocompatibility antigen typing methodology and database designed to facilitate clinical applications. *PloS one*, 1:e42, January 2006. 52

[107] L Stoltze, T P Dick, M Deeg, B Pömmerl, H G Rammensee, and H Schild. Generation of the vesicular stomatitis virus nucleoprotein cytotoxic T lymphocyte epitope requires proteasome-dependent and -independent proteolytic activities. *European journal of immunology*, 28(12):4029–36, December 1998. 15

[108] T Sturniolo, E Bono, J Ding, L Raddrizzani, O Tuereci, U Sahin, M Braxenthaler, F Gallazzi, M P Protti, F Sinigaglia, and J Hammer. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature biotechnology*, 17(6):555–61, June 1999. 16

[109] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–7, April 2004. 69, 93

[110] S T Tang, M Wang, K Lamberth, M Harndahl, M H Dziegiel, M H Claesson, S Buus, and O Lund. MHC-I-restricted epitopes conserved among variola and other related orthopoxviruses are recognized by T cells 30 years after vaccination. *Archives of virology*, 153(10):1833–44, January 2008. 15

[111] S Tenzer, B Peters, S Bulik, O Schoor, C Lemmel, M M Schatz, P-M Kloetzel, H-G Rammensee, H Schild, and H-G Holzhütter. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cellular and molecular life sciences : CMLS*, 62(9):1025–37, 2005. 15, 28

[112] Kasper Thorsen, Karina D Sø rensen, Anne Sofie Brems-Eskildsen, Charlotte Modin, Mette Gaustadnes, Anne-Mette K Hein, Mogens Kruhø ffer, Sø ren Laurberg, Michael Borre, Kai Wang, Sø ren Brunak, Adrian R Krainer, Niels Tø rring, Lars Dyrskjø t, Claus L Andersen, and Torben F Orntoft. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & cellular proteomics : MCP*, 7(7):1214–24, July 2008. 34, 43

[113] G Trinquier and Y H Sanejouand. Which effective property of amino acids is best preserved by the genetic code? *Protein engineering*, 11(3):153–69, March 1998. 38

[114] Shiro Urayama, Wei Zou, Kindra Brooks, and Vladimir Tolstikov. Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer. *Rapid communications in mass spectrometry : RCM*, 24(5):613–20, March 2010. 46

[115] P van der Bruggen, C Traversari, P Chomez, C Lurquin, E De Plaen, B Van den Eynde, A Knuth, and T Boon. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science (New York, N.Y.)*, 254(5038):1643–7, December 1991. 34

[116] P M van Endert, R Tampé, T H Meyer, R Tisch, J F Bach, and H O McDevitt. A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity*, 1(6):491–500, September 1994. 15

[117] Julian P Venables, Roscoe Klinck, ChuShin Koh, Julien Gervais-Bird, Anne Bramard, Lyna Inkel, Mathieu Durand, Sonia Couture, Ulrike Froehlich, Elvy Lapointe, Jean-François Lucier, Philippe Thibault, Claudine Rancourt, Karine Tremblay, Panagiotis Prinos, Benoit Chabot, and Sherif Abou Elela. Cancer-associated regulation of alternative splicing. *Nature structural & molecular biology*, 16(6):670–6, June 2009. 35

[118] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell,

R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, February 2001. 1

[119] Matteo Vergati, Chiara Intrivici, Ngar-Yee Huen, Jeffrey Schlom, and Kwong Y Tsang. Strategies for cancer vaccine development. *Journal of biomedicine & biotechnology*, 2010, January 2010. 76

[120] R B Walter, J M Pagel, T A Gooley, E W Petersdorf, M L Sorror, A E Woolfrey, J A Hansen, A I Salter, E Lansverk, F M Stewart, P V O'Donnell, and F R Appelbaum. Comparison of matched unrelated and matched related donor myeloablative hematopoietic cell transplantation for adults with acute myeloid leukemia in first remission. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, 24(7):1276–82, July 2010. 9

[121] Mingjun Wang, Kasper Lamberth, Mikkel Harndahl, Gustav Rø der, Anette Stryhn, Mette V Larsen, Morten Nielsen, Claus Lundegaard, Sheila T Tang, Morten H Dziegiel, Jø rgen Rosenkvist, Anders E Pedersen, Sø ren Buus, Mogens H Claesson, and Ole Lund. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. *Vaccine*, 25(15):2823–31, April 2007. 15

[122] Edus H. Warren, Philip D. Greenberg, and Stanley R. Riddell. Cytotoxic T-Lymphocyte-Defined Human Minor Histocompatibility Antigens With a Restricted Tissue Distribution. *Blood*, 91(6):2197–2207, 1998. 11

[123] Nicholas F S Watson, Judith M Ramage, Zahra Madjd, Ian Spendlove, Ian O Ellis, John H Scholefield, and Lindy G Durrant. Immunosurveillance is active in colorectal cancer as downregulation but not complete loss of MHC class I expression correlates with a poor prognosis. *International journal of cancer. Journal international du cancer*, 118(1):6–10, January 2006. 35

[124] P L Weiden, N Flournoy, E D Thomas, R Prentice, A Fefer, C D Buckner, and R Storb. Antileukemic effect of graft-versus-host disease in human recipients of allogeneic-marrow grafts. *The New England journal of medicine*, 300(19):1068–73, May 1979. 11

[125] Rasmus Wernersson. Virtual Ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic acids research*, 34(Web Server issue):W385–8, July 2006. 36

[126] E John Wherry, Tatiana N Golovina, Susan E Morrison, Gomathinayagam Sinnathamby, Michael J McElhaugh, David C Shockey, and Laurence C Eisenlohr. Reevaluating the generation of a "proteasome-independent" MHC class I-restricted CD8 T cell epitope. *Journal of immunology (Baltimore, Md. : 1950)*, 176(4):2249–61, February 2006. 16

[127] EB Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. 44

[128] W C Wimley and S H White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature structural biology*, 3(10):842–8, October 1996. 38

[129] J W Yewdell and J R Bennink. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. *Annual review of immunology*, 17:51–88, January 1999. 8, 15

[130] Neal S Young, Rodrigo T Calado, and Phillip Scheinberg. Current concepts in the pathophysiology and treatment of aplastic anemia. *Blood*, 108(8):2509–19, October 2006. 8

[131] Jun Yu, Songnian Hu, Jun Wang, Gane Ka-Shu Wong, Songgang Li, Bin Liu, Yajun Deng, Li Dai, Yan Zhou, Xiuqing Zhang, Mengliang Cao, Jing Liu, Jiandong Sun, Jiabin Tang, Yanjiong Chen, Xiaobing Huang, Wei Lin, Chen Ye, Wei Tong, Lijuan Cong, Jianing Geng, Yujun Han, Lin Li, Wei Li, Guangqiang Hu, Xiangang Huang, Wenjie Li, Jian Li, Zhanwei Liu, Long Li, Jianping Liu, Qiuhui Qi, Jinsong Liu, Li Li, Tao Li, Xuegang Wang, Hong Lu, Tingting Wu, Miao Zhu, Peixiang Ni, Hua Han, Wei Dong, Xiaoyu Ren, Xiaoli Feng, Peng Cui, Xianran Li, Hao Wang, Xin Xu, Wenxue Zhai, Zhao Xu, Jinsong Zhang, Sijie He, Jianguo Zhang, Jichen Xu, Kunlin Zhang, Xianwu Zheng, Jianhai Dong, Wanyong Zeng, Lin Tao, Jia Ye, Jun Tan, Xide Ren, Xuewei Chen, Jun He, Daofeng Liu, Wei Tian, Chaoguang Tian, Hongai Xia, Qiyu Bao, Gang Li, Hui Gao, Ting Cao, Juan Wang, Wenming Zhao, Ping Li, Wei Chen, Xudong Wang, Yong Zhang, Jianfei Hu, Jing Wang, Song Liu, Jian Yang, Guangyu Zhang, Yuqing Xiong, Zhijie Li, Long Mao, Chengshu Zhou, Zhen Zhu, Runsheng Chen, Bailin Hao, Weimou Zheng, Shouyi Chen, Wei Guo, Guojie Li, Siqi Liu, Ming Tao, Jian Wang, Lihuang Zhu, Longping Yuan, and Huanming Yang. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science (New York, N.Y.)*, 296(5565):79–92, April 2002. 2

[132] Hao Zhang, Claus Lundegaard, and Morten Nielsen. Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics (Oxford, England)*, 25(1):83–9, January 2009. 18, 38

# Appendix

## Tissue-specific mRNA expression levels

Expression data from the GNF gene expression database [109] covers mRNA data on 79 human tissues. The data was normalized using robust multi-array averaging (RMA) and bias corrected as described in [38, 20]. For the analysis of hematopoietic tissues, we took 16 tissues, namely Wholeblood(JJV), CD33 Myeloid, CD14 Monocytes, BDCA4 Dendritic Cells, CD56 NK Cells, CD4 T Cells, CD8 T Cells, CD19 B Cells, CD 105 Endothelial, CD34 Cells, B Lymphoblasts, Thymus, Tonsil, Lymph node, CD71 Early Erythroid and Bonemarrow. As representatives for tissues affected by acute and chronic GVHD we selected liver and skin. The mean mRNA expression level for the top ranked genes associated with different endpoints was compared. Liver and skin was analyzed separately, whereas mRNA expression of hematopoietic tissues was both analyzed as the mean of all 16 hematopoietic tissues per gene, as well as by taking only the highest expressed tissue per gene. The latter was done, as it is not expected that a gene is highly expressed in all 16 hematopoietic tissues and a potential signal would be lost if a gene is only highly expressed in one of the hematopoietic tissue types. Observed average expression and SD for the respective 300 top ranked genes of each list is shown in Figure 1. A difference in expression for the different endpoints was expected, we could however not identify any variations in mRNA expression. An analysis comparing the top 100 ranked genes is reporting similar results.

| | | | | Hematopoietic tissues | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 16 tissues | | highest expressed | | Liver | | Skin | |
| | | | # genes | mean | SD | mean | SD | mean | SD | mean | SD |
| mHags | relapse | decreased | 176 | 6.11 | 1 | 6.92 | 1.35 | 6.26 | 1.18 | 6.18 | 1.01 |
| | | increased | 158 | 6.06 | 1.12 | 6.91 | 1.49 | 6.2 | 1.2 | 6.13 | 1.14 |
| | acute GVHD | decreased | 167 | 5.99 | 1.02 | 6.77 | 1.39 | 6.12 | 1.2 | 6.08 | 1.06 |
| | | increased | 171 | 6.08 | 1.05 | 6.83 | 1.36 | 6.32 | 1.34 | 6.16 | 1.08 |
| | chronic GVHD | decreased | 173 | 6.17 | 1.07 | 6.99 | 1.47 | 6.27 | 1.29 | 6.23 | 1.09 |
| | | increased | 182 | 6.13 | 1.1 | 6.89 | 1.44 | 6.18 | 1.15 | 6.22 | 1.16 |
| SNPs | relapse | decreased | 172 | 6.13 | 1.07 | 6.93 | 1.45 | 6.28 | 1.25 | 6.2 | 1.1 |
| | | increased | 187 | 6.21 | 1.25 | 7.01 | 1.56 | 6.24 | 1.3 | 6.24 | 1.21 |
| | acute GVHD | decreased | 169 | 5.83 | 1.01 | 6.52 | 1.26 | 5.99 | 1.18 | 5.91 | 1.04 |
| | | increased | 168 | 6.2 | 1.09 | 6.88 | 1.33 | 6.34 | 1.24 | 6.31 | 1.17 |
| | chronic GVHD | decreased | 179 | 6.04 | 1.09 | 6.91 | 1.46 | 6.16 | 1.32 | 6.12 | 1.12 |
| | | increased | 191 | 6.2 | 1.07 | 7.05 | 1.41 | 6.34 | 1.16 | 6.28 | 1.07 |
| | all genes | | 5542 | 6.12 | 1.1 | 6.92 | 1.43 | 6.23 | 1.23 | 6.19 | 1.12 |

**Figure 1. mRNA expression analysis.** Tissue expression data is based on GNF database. The mean expression level for the top 300 ranked genes per analyzed endpoint is shown. Not all analyzed genes are covered by the GNF database; the number of analyzed genes in each group is given. Hematopoietic tissues are listed as a pooled mean over 16 tissues as well as the mean of the highest expressed tissue per gene. Not all genes part of the study are covered by the GNF database. The number of genes with expression value is given by # genes.

| | | | P-vaue | | # patient-donor pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | All pairs | | Associated with endpoint | |
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | ENSG00000154642 | C21orf91 | 0.0004 | 0.0014 | 50 | 43 | 15 | 2 |
| | ENSG00000188352 | KIAA1797 | 0.0004 | 0.0007 | 36 | 57 | 13 | 4 |
| | ENSG00000157119 | KBTBD5 | 0.0008 | 0.0010 | 55 | 38 | 16 | 1 |
| | ENSG00000182612 | TSPAN10 | 0.0008 | 0.0011 | 54 | 39 | 16 | 1 |
| | ENSG00000197386 | HTT | 0.0009 | 0.0014 | 43 | 50 | 14 | 3 |
| | ENSG00000205409 | OR52E6 | 0.0012 | 0.0013 | 61 | 32 | 17 | 0 |
| | ENSG00000108417 | KRT37 | 0.0015 | 0.0016 | 42 | 51 | 13 | 4 |
| | ENSG00000185313 | SCN10A | 0.0015 | 0.0016 | 50 | 43 | 15 | 2 |
| | ENSG00000170209 | ANKK1 | 0.0017 | 0.0030 | 51 | 42 | 15 | 2 |
| | ENSG00000179406 | NCRNA00174 | 0.0019 | 0.0017 | 61 | 32 | 17 | 0 |
| increased | ENSG00000140948 | ZCCHC14 | 0.0002 | 0.0001 | 73 | 20 | 8 | 9 |
| | ENSG00000163291 | PAQR3 | 0.0015 | 0.0006 | 79 | 14 | 10 | 7 |
| | ENSG00000158286 | RNF207 | 0.0016 | 0.0012 | 60 | 33 | 6 | 11 |
| | ENSG00000198711 | C1orf191 | 0.0019 | 0.0000 | 88 | 5 | 14 | 3 |
| | ENSG00000173578 | XCR1 | 0.0023 | 0.0000 | 91 | 2 | 15 | 2 |
| | ENSG00000115827 | DCAF17 | 0.0028 | 0.0000 | 90 | 3 | 14 | 3 |
| | ENSG00000188227 | ZNF793 | 0.0044 | 0.0031 | 76 | 17 | 10 | 7 |
| | ENSG00000149403 | GRIK4 | 0.0045 | 0.0000 | 91 | 2 | 15 | 2 |
| | ENSG00000165105 | RASEF | 0.0049 | 0.0040 | 65 | 28 | 7 | 10 |
| | ENSG00000198093 | ZNF649 | 0.0053 | 0.0000 | 91 | 2 | 15 | 2 |

**Table 1. Top ranked genes associated with relapse, based on nsSNP differences** 17 patients are associated with relapse. Analysis is based on nsSNP differences between patient and donor. Top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased likelihood of relapse. A decreased association is given when fewer nsSNPs are associated with a higher likelihood of relapse. The ranked lists are sorted based on a permutation test (10,000 permutations). Shown p-values are not corrected for multiple comparisons. The number of patient-donor pairs, with and without difference in nsSNPs, is listed for all pairs, as well as for pairs associated with relapse.

| | | | P-vaue | | # patient-donor pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | All pairs | | Associated with endpoint | |
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | ENSG00000128815 | WDFY4 | 0.0010 | 0.0008 | 33 | 60 | 16 | 11 |
| | ENSG00000123119 | NECAB1 | 0.0024 | 0.0018 | 72 | 21 | 27 | 0 |
| | ENSG00000129295 | LRRC6 | 0.0038 | 0.0042 | 58 | 35 | 23 | 4 |
| | ENSG00000117133 | RPF1 | 0.0043 | 0.0034 | 74 | 19 | 27 | 0 |
| | ENSG00000064419 | TNPO3 | 0.0048 | 0.0038 | 70 | 23 | 26 | 1 |
| | ENSG00000101596 | SMCHD1 | 0.0048 | 0.0043 | 40 | 53 | 18 | 9 |
| | ENSG00000105227 | PRX | 0.0051 | 0.0062 | 45 | 48 | 19 | 8 |
| | ENSG00000164458 | T | 0.0070 | 0.0058 | 44 | 49 | 19 | 8 |
| | ENSG00000149311 | ATM | 0.0077 | 0.0081 | 68 | 25 | 25 | 2 |
| | ENSG00000172869 | DMXL1 | 0.0082 | 0.0082 | 64 | 29 | 24 | 3 |
| increased | ENSG00000153560 | UBP1 | 0.0001 | 0.0000 | 73 | 20 | 14 | 13 |
| | ENSG00000198542 | ITGBL1 | 0.0001 | 0.0000 | 88 | 5 | 22 | 5 |
| | ENSG00000118479 | no HGNC | 0.0002 | 0.0002 | 65 | 28 | 12 | 15 |
| | ENSG00000175866 | BAIAP2 | 0.0004 | 0.0000 | 79 | 14 | 17 | 10 |
| | ENSG00000185187 | SIGIRR | 0.0005 | 0.0001 | 71 | 22 | 14 | 13 |
| | ENSG00000205351 | AC002347.2 | 0.0005 | 0.0003 | 73 | 20 | 15 | 12 |
| | ENSG00000127903 | ZNF835 | 0.0006 | 0.0003 | 70 | 23 | 14 | 13 |
| | ENSG00000168528 | SERINC2 | 0.0006 | 0.0000 | 83 | 10 | 20 | 7 |
| | ENSG00000184350 | MRGPRE | 0.0007 | 0.0005 | 66 | 27 | 13 | 14 |
| | ENSG00000079931 | MOXD1 | 0.0008 | 0.0000 | 85 | 8 | 21 | 6 |

**Table 2. Top ranked genes associated with acute GVHD, based on nsSNP differences** 27 patients are associated with acute GVHD. Analysis is based on nsSNP differences between patient and donor. Top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased risk of acute GVHD. A decreased association is given when fewer nsSNPs are associated with a higher likelihood of acute GVHD. The ranked lists are sorted based on a permutation test (10,000 permutations). Shown p-values are not corrected for multiple comparisons. The number of patient-donor pairs, with and without difference in nsSNPs, is listed for all pairs, as well as for pairs associated with acute GVHD.

| | | | P-vaue | | # patient-donor pairs | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | All pairs | | Associated with endpoint | |
| | | | Permutation test | Log-rank test | 0 mHags | >0 mHags | 0 mHags | >0 mHags |
| decreased | KRT78 | 0.0018 | 0.0020 | 54 | 39 | 39 | 15 | |
| | ENSG00000189241 | TSPYL1 | 0.0040 | 0.0041 | 54 | 39 | 37 | 17 |
| | ENSG00000139193 | CD27 | 0.0051 | 0.0038 | 66 | 27 | 45 | 9 |
| | ENSG00000197530 | MIB2 | 0.0055 | 0.0046 | 62 | 31 | 40 | 14 |
| | ENSG00000072840 | EVC | 0.0062 | 0.0052 | 40 | 53 | 28 | 26 |
| | ENSG00000125869 | C20orf103 | 0.0063 | 0.0065 | 71 | 22 | 47 | 7 |
| | ENSG00000142611 | PRDM16 | 0.0075 | 0.0065 | 59 | 34 | 39 | 15 |
| | ENSG00000203797 | DDO | 0.0083 | 0.0047 | 76 | 17 | 49 | 5 |
| | ENSG00000184560 | C17orf74 | 0.0088 | 0.0066 | 51 | 42 | 35 | 19 |
| | ENSG00000204277 | AC087645.1 | 0.0088 | 0.0097 | 47 | 46 | 33 | 21 |
| increased | ENSG00000153560 | UBP1 | 0.0001 | 0.0000 | 73 | 20 | 14 | 13 |
| | ENSG00000198542 | ITGBL1 | 0.0001 | 0.0000 | 88 | 5 | 22 | 5 |
| | ENSG00000118479 | no HGNC | 0.0002 | 0.0002 | 65 | 28 | 12 | 15 |
| | ENSG00000175866 | BAIAP2 | 0.0004 | 0.0000 | 79 | 14 | 17 | 10 |
| | ENSG00000185187 | SIGIRR | 0.0005 | 0.0001 | 71 | 22 | 14 | 13 |
| | ENSG00000205351 | AC002347.2 | 0.0005 | 0.0003 | 73 | 20 | 15 | 12 |
| | ENSG00000127903 | ZNF835 | 0.0006 | 0.0003 | 70 | 23 | 14 | 13 |
| | ENSG00000168528 | SERINC2 | 0.0006 | 0.0000 | 83 | 10 | 20 | 7 |
| | ENSG00000184350 | MRGPRE | 0.0007 | 0.0005 | 66 | 27 | 13 | 14 |
| | ENSG00000079931 | MOXD1 | 0.0008 | 0.0000 | 85 | 8 | 21 | 6 |

**Table 3. Top ranked genes associated with chronic GVHD, based on nsSNP differences** 27 patients are associated with chronic GVHD. Analysis is based on nsSNP differences between patient and donor. Top 10 ranked genes, based on permutation test, are shown for genes associated with increased and decreased risk of chronic GVHD. A decreased association is given when fewer nsSNPs are associated with a higher likelihood of chronic GVHD. The ranked lists are sorted based on a permutation test (10,000 permutations). Shown p-values are not corrected for multiple comparisons. The number of patient-donor pairs, with and without difference in nsSNPs, is listed for all pairs, as well as for pairs associated with chronic GVHD.