



Computational methods to improve development of T cell based cancer immunotherapies

Anja Franziska Mösch

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Dr. Fabian J. Theis

Prüfer*innen der Dissertation:

1. Prof. Dr. Dmitrij Frishman
2. Priv.-Doz. Dr. Barbara Lösch

Die Dissertation wurde am 06.12.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 23.05.2022 angenommen.

Abstract

The importance of computational methods has grown tremendously in every field of life sciences. Especially personalized therapies such as cancer immunotherapy, which rely on genomic, transcriptomic and proteomic information about an individual patient, cannot be developed without computational support, machine learning algorithms in particular. Cancer immunotherapy is based on the idea of enabling the patient's immune system to attack and destroy tumor cells. This requires a deep understanding of interactions between tumor and immune system, which are highly patient-specific and therefore need to be assessed individually by methods suitable to identify tumor antigen expression and immune profiling. For adoptive T cell therapy, where potent anti-tumor T cell receptors (TCRs) are introduced into the patient's own T cells, two concrete challenges ensue. First, potential cross-reactivity of a TCR needs to be assessed to avoid recognition and therefore destruction of healthy tissue. To tackle this task, the tool Expitope was created and further developed to version 2.0. For a given peptide sequence, which is the target sequence for a therapeutic TCR candidate, all similar peptide sequences present in healthy tissue are identified, including an estimation of cross-reaction severity by calculating a weighted tissue score. Second, the high-throughput testing of therapeutic TCR candidates can yield ambiguous results regarding the presence of more than one chain for either the α or β locus at the mRNA level, where only one chain of each locus is required for a functional TCR. To enable *in silico* screening, TCRpair was developed to predict which chain is part of the functional TCR that recognizes the tumor antigen when sequencing results show two expressed chains for one locus. Furthermore, various compu-

Abstract

tational methods used for cancer immunotherapy applications are described in the review titled “Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors”. This includes different approaches for the identification of neoepitopes, which are T cell targets derived from tumor-specific mutations, and methods to predict epitope-TCR binding.

Zusammenfassung

Computerbasierte Methoden haben in jedem Bereich der Lebenswissenschaften erheblich an Bedeutung gewonnen. Besonders personalisierte Therapien wie Krebsimmuntherapie, die sich auf genomische, transkriptomische und proteomische Informationen jedes einzelnen Patienten stützt, kann nicht ohne Computerunterstützung, allen voran maschinelles Lernen, entwickelt werden. Krebsimmuntherapie fußt auf der Idee, es dem Immunsystem einer Patientin zu ermöglichen, Tumorzellen anzugreifen und zu zerstören. Dafür ist ein tiefgehendes Verständnis der Interaktionen zwischen Tumor und Immunsystem notwendig, die für jeden Patienten spezifisch sein können und deswegen individuell bewertet werden müssen, wobei passende Methoden zur Identifikation von Tumorantigen-Expression und Analyse des Immunprofils angewendet werden. Für die adoptive T-Zell-Therapie, bei der den T-Zellen einer Patientin potente T-Zell-Rezeptoren (TCRs) hinzugefügt werden, ergeben sich daraus zwei konkrete Herausforderungen. Zum einen muss ein TCR auf mögliche Kreuzreaktivität hin untersucht werden, um die Erkennung und daraus folgende Zerstörung von gesundem Gewebe zu verhindern. Um diese Aufgabe zu lösen wurde das Programm Expitope erstellt und zu Version 2.0 weiterentwickelt. Für eine Peptidsequenz, die das Ziel eines therapeutischen TCRs ist, werden alle ähnlichen Peptide, die in gesundem Gewebe vorhanden sind, ermittelt und der Schweregrad einer Kreuzreaktivität basierend auf einem gewichteten Gewebescore berechnet. Zum anderen kann die Hochdurchsatztestung von therapeutischen TCR Kandidaten uneindeutige Resultate ergeben, wenn mehr als eine Kette des α oder β locus auf RNA-Level präsent ist, wobei nur eine Kette jedes loci für die Bildung eines funktionalen TCRs benötigt wird. Für *in*

Zusammenfassung

silico Screening wurde TCRpair entwickelt, um vorherzusagen, welche Kette Teil des funktionalen TCRs ist, der das Tumorantigen erkennt, falls zwei Ketten für einen locus in den Sequenzierergebnissen vorhanden sind. Des Weiteren werden im Review “Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors” verschiedene computerbasierte Methoden, die für Krebsimmuntherapien verwendet werden, beschrieben. Dazu gehören verschiedene Ansätze zur Identifizierung von Neoepitopen, T-Zellen-Ziele aus tumorspezifischen Mutationen, und Methoden zur Vorhersage von Bindung zwischen Epitop und TCR.

Acknowledgments

I would like to thank my supervisor Prof. Dr. Dmitrij Frishman, who not only suggested to me to pursue a PhD in Bioinformatics in collaboration with an industrial partner, but also always had great trust and confidence in my abilities and guided me through the years. I also want to thank my second examiner and Medigene supervisor Priv.-Doz. Dr. Barbara Lösch, and the chair of my thesis committee Prof. Dr. Fabian J. Theis.

Of the people with whom I worked together at Medigene Immunotherapies GmbH I want to especially thank my mentor and supervisor Dr. Silke Raffegerst, Dr. Manon Weis, the MM team and the automation team.

Thank you to Prof. Frishman's group at the School of Life Science of the Technical University Munich in Weihenstephan and the TUM Summer School members.

I owe thanks to Julia Rackerseder, who supported me as a friend and a colleague, and to many other friends who believed in me and always showed genuine interest in my work.

I would like to thank Dr. Fabian Poetke, who accompanied me during my whole academic career and a large part of my life and who always supported me without hesitation.

Finally, I want to thank my parents Prof. Dr. Christine Leib-Mösch and Siegfried Mösch, who always supported my scientific interests and my academic career. Above all, my mother is a great role model and made it natural for me that a woman can have a career in life sciences.

Publications

The following publications are part of this thesis:

- Jaravine V, **Mösch A**, Raffegerst S, Schendel DJ, Frishman D. Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. *BMC Cancer* 17 (2017)
- **Mösch A**, Raffegerst S, Weis M, Schendel DJ, Frishman D. Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors *Frontiers in Genetics* 10 (2019)
- **Mösch A**, Frishman D: TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T-cell receptor α and β chains. *Bioinformatics* 37 (21): 3038-3940 (2021)

Furthermore, during my time as member of the Graduate Center Weihenstephan I contributed to the following publication:

- Littmann M, Selig K, Cohen-Lavi L, Frank Y, Hönigschmid P, Kataka E, **Mösch A**, Qian K, Ron A, Schmid S, Sorbie A, Szlak L, Dagan-Wiener A, Ben-Tal N, Niv M Y, Razansky D, Schuller B W, Ankerst D, Hertz T, Rost B. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence* 2 (2020)

Contents

Abstract	iii
Zusammenfassung	v
Acknowledgments	vii
Publications	ix
Contents	xi
List of Figures	xiii
Acronyms	xv
1 Introduction	1
1.1 Motivation	1
1.2 T cell based immune response	2
1.2.1 Antigen presentation	2
1.2.2 T cell recognition	5
1.3 T cell receptor based cancer immunotherapy	7
1.3.1 Tumor-specific T cell antigens	7
1.3.2 Adoptive T cell therapy	9
1.3.3 Cross-reactivity	9
1.4 Computational methods for immunology	11
1.4.1 Epitope prediction	11

xi

CONTENTS

1.4.2	Long short-term memory networks for sequence based T cell receptor predictions	12
2	Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues	15
2.1	Abstract	15
2.2	Contribution	16
3	Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors	17
3.1	Abstract	17
3.2	Contribution	18
4	TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T cell receptor α and β chains	21
4.1	Abstract	21
4.2	Contribution	22
5	Conclusion	23
	Bibliography	25
A	Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues	28
B	Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors	37

List of Figures

1.1	Major histocompatibility complex (MHC) class I antigen presentation pathway for peptides recognized by CD8 ⁺ cytotoxic T cells.	3
1.2	Number of HLA alleles identified and named by the WHO Nomenclature Committee for Factors of the HLA System from 1987 to June 2021.	4
1.3	T cell receptor (TCR) binding to a peptide presented by major histocompatibility complex (MHC) class I.	6
1.4	Personalized cancer treatment with TCRs.	8
1.5	Priming of T cells with mature dendritic cells.	10
1.6	Deep learning model architecture of TCRpair using two bidirectional LSTMs	13

Acronyms

APC	Antigen presenting cell.
CAR	Chimeric antigen receptor.
CDR	Complementarity-determining region.
CNN	Convolutional neural networks.
CR	Cross-reactivity.
CTA	Cancer-testus antigen.
HLA	Human leucocyte antigen.
IFN γ	Interferon gamma.
LSTM	Long short-term memory.
MHC	Major histocompatibility complex.
pMHC	Peptide-major histocompatibility complex.
SNV	Single-nucleotide variant.
TAP	Transporter associated with antigen processing.
TCR	T cell receptor.
TIL	Tumor-infiltrating lymphocyte.

1 Introduction

1.1 Motivation

Cancer immunotherapy has made remarkable progress in the last years and has become a reliable part of cancer treatment methods. This is possible not only because our understanding about immune response to cancer, tumor microenvironment and tumor-immune system interaction grows but also because a variety of computational methods assist in tackling the vast amount of information gathered by *in vitro* experiments, especially next-generation DNA and RNA sequencing. The development of computational tools, mainly based on machine learning, expands together with the availability of data (Mösch *et al.*, 2019).

In this thesis, the focus lies on T cell based immune responses against tumors, especially the so called adoptive T cell therapy, which is described in subsection section 1.3.2. How T cells are able to detect and eliminate cancer cells is described in subsection 1.2.1 and in subsection 1.2.2.

To aid the development of tumor antigen-specific adoptive T cell therapy, the tool Expitope has been developed and was further improved as seen in the publication “Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues” (Jaravine *et al.*, 2017). TCRpair has also been created to tackle a problem associated with high-throughput scanning of T cell receptor (TCR) candidates. In about 30% of the T cell clones analyzed, two α chains can be expressed, but only one of them is

1 Introduction

part of the functional, i.e. tumor antigen recognizing TCR. Predicting which α/β chain pair is the functional one is the goal of TCRpair (Mösch and Frishman, 2021).

1.2 T cell based immune response

The human body has two ways to react to foreign particles or aberrant cells. The first one is the so called innate immune system, which is the invariable reaction to pathogenic treats and is available to an individual from birth. The second one, the adaptive immune system, is able to learn from previous encounters with pathogens and can build a long term memory that enables the immune cells to react more quickly to reoccurring threats.

T cells are a core element of the adaptive immune system. Together with antibody producing B cells, they belong to the lymphocytes, which are part of the circulating white blood cells. T cells are involved in the acute immune response by killing infected cells and cytokine signaling. A subpopulation of T cells stays in the system as so called memory T cells, which are part of the long term immune memory. Cancer immunotherapies are developed around the T cell's ability to directly eliminate cancer cells (Coulie *et al.*, 2014).

1.2.1 Antigen presentation

The process of antigen presentation happens comparably in almost every cell of the human body, although there are also specialized antigen presenting cells (APCs) like dendritic cells. APCs are very efficient in presenting peptides they acquire to T cells, which, in case of recognition, will launch an immune response to find and eliminate all other cells presenting these peptides.

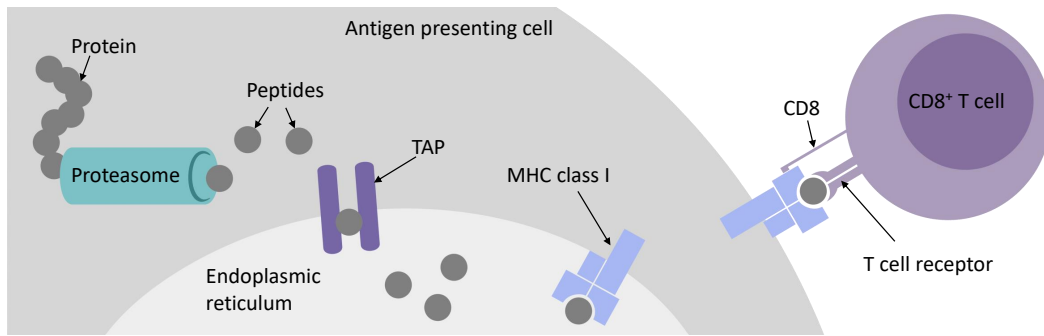


Figure 1.1: Major histocompatibility complex (MHC) class I antigen presentation pathway for peptides recognized by CD8⁺ cytotoxic T cells. Source: Mösch *et al.* (2019)

There are two pathways of antigen presentation, one mediated by the major histocompatibility complex (MHC) class I and one by MHC class II. They differ mainly in the source of the peptides and the type of T cells that can recognize these peptides. For MHC class II, peptides are acquired from external sources by APCs and are recognized by CD4⁺ T cells, so called helper T cells, which release stimulating cytokines upon binding to a peptide to activate other immune cells (Andreatta *et al.*, 2017). For MHC class I, on which the focus lies in this work since this pathway is mainly exploited for anti-tumor immune responses, CD8⁺ T cells eliminate cells upon recognizing a foreign peptide (Leone *et al.*, 2013)). These peptides come from infections, cancer-specific mutations or gene expression changes or other changes to a healthy cell's peptidome and are processed and presented on the cell surface.

A protein present in a cancer cell will eventually be digested by the cell's proteasome, which is responsible for cleaving unneeded, faulty or damaged proteins in peptides. Some of these peptides will bind to the transporter associated with antigen processing (TAP), which transports these peptides to the endoplasmic reticulum (Peters *et al.*, 2003). There, MHC class I molecules bind some of the peptides and present them on the cell surface (see Figure 1.1). These peptides are also called epitopes,

1 Introduction

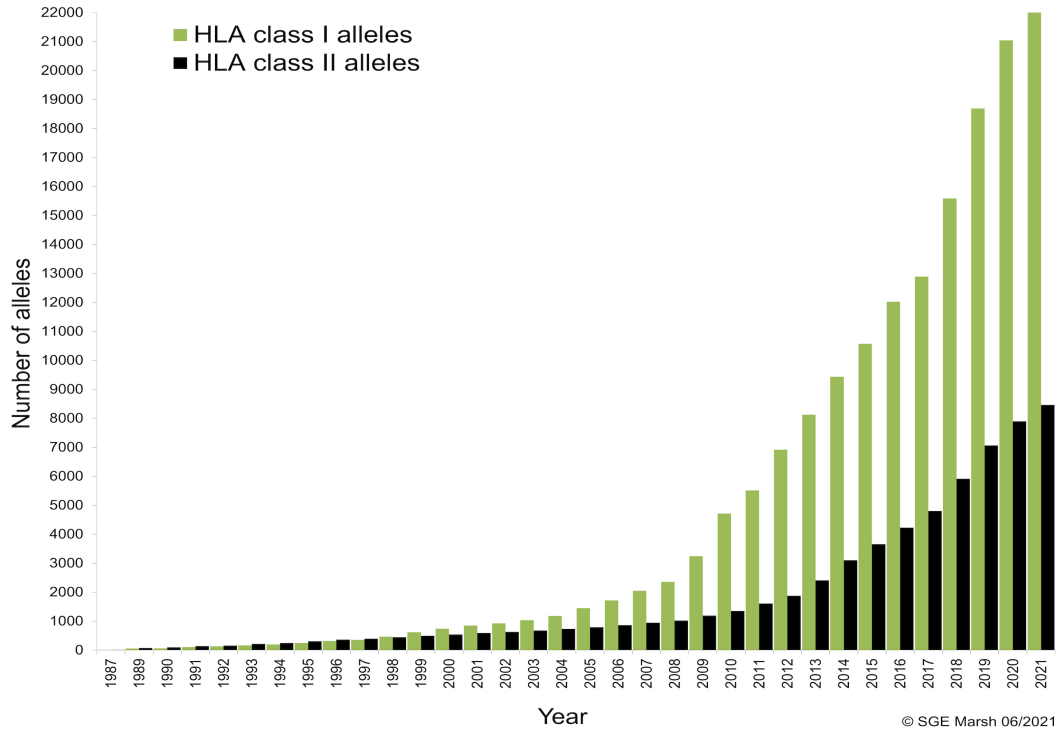


Figure 1.2: Number of HLA alleles identified and named by the WHO Nomenclature Committee for Factors of the HLA System from 1987 to June 2021. Source: <http://hla.alleles.org>, Robinson *et al.* (2015)

which is the exact part or sequence of the antigen that leads to an immune response (Coulie *et al.*, 2014).

For every aspect of the antigen presentation pathway, the peptide's properties, i.e. its amino acid composition, are responsible whether and how a peptide is processed. Therefore, computational methods can predict every step of the antigen presentation, although in recent years the focus has mainly been on MHC binding. MHC class I molecules possess a tremendous variety among the human population, because the human leukocyte antigen (HLA) loci encoding for MHC class I, especially HLA-A, HLA-B, and HLA-C, are highly polymorphic. There are

1.2 T cell based immune response

more than 30,000 HLA known alleles, which belong to five different HLA genes and 12 HLA pseudogenes (see Figure 1.2, Robinson *et al.* 2015). This variability is responsible for the large diversity of peptides that can be presented on cell surfaces, since each MHC can bind different peptides. On a person's individual level, each gene has two alleles, although there are some alleles encoding for more efficient MHCs, as there are also individuals homozygous for one or more loci (Boegel *et al.*, 2014; Gragert *et al.*, 2013). Therefore, the HLA type of a patient needs to be known in order to apply a suitable HLA-dependent immunotherapy. On population level, differing HLA allele frequencies have developed, which poses additional challenges for immunotherapeutic approaches as more data is available for predominant alleles in populations that are more often subject to data collection. One HLA allele on which plenty of data is available and one of the most reliable binding predictions is HLA-A*02:01, the most common allele in Caucasian populations. Since machine learning methods rely on diverse datasets, this can lead to bias in prediction algorithms requiring HLA-specific information.

1.2.2 T cell recognition

The T cell receptor (TCR) is a heterodimer usually consisting of an α and a β chain. Together with the respective co-receptor CD8⁺ or CD4⁺, it allows the T cell to bind to a peptide-MHC (pMHC) presented on a cell surface (see Figure 1.3). If the TCR binds, the T cell gets stimulated and, depending on whether it is a cytotoxic T cell or T helper cell, it will directly eliminate the cell presenting the antigen or release cytokines like interferon gamma (IFN γ) to signal other immune cells to react.

Each of the TCR's chains consists of three complementarity-determining regions (CDRs) that bind to different parts of the pMHC. The most important CDR responsible for the vast diversity of an individual's TCR repertoire is CDR3 (Arstila, 1999; Hughes *et al.*, 2003). This region of a

1 Introduction

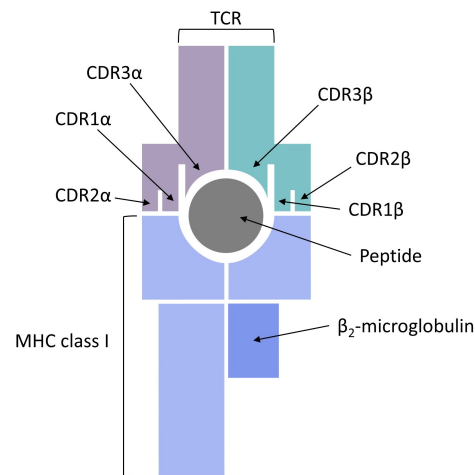


Figure 1.3: T cell receptor (TCR) binding to a peptide presented by major histocompatibility complex (MHC) class I. Source: Mösch *et al.* (2019)

TCR chain is created by the so called V(D)J recombination, referring to the V and J genes for the α chain and β chain and to the D gene for the β chain. Each human possesses several alleles of the V, D and J genes, which, in addition to the recombinant nature of the CDR3 region, allow for a high variety of CDR3 sequences. Also, TCRs are not limited to recognizing only one single pMHC (Yates, 2014). However, as this would mean that TCRs can react to all peptides presented on cell surfaces, even peptides belonging to the normal human peptidome, there is a selection process for the T cells. In the thymus, where they mature, T cells and therefore their receptors are exposed to the healthy peptidome before they are released into the blood stream. If a TCR binds to a pMHC with a peptide which is present in healthy tissue the respective T cell gets eliminated and will never be part of the TCR repertoire meant to recognize infected or aberrant cells.

T cells are activated by APCs, which present foreign peptides, and will proliferate if they are stimulated by binding to a pMHC. This procedure

1.3 T cell receptor based cancer immunotherapy

is used for cancer vaccines, which contain these peptides, as well as for identifying T cell clones suitable for adoptive T cell therapy (Hu *et al.*, 2017; Wilde *et al.*, 2009).

1.3 T cell receptor based cancer immunotherapy

1.3.1 Tumor-specific T cell antigens

The ideal T cell antigen is highly expressed in tumors but not expressed in healthy tissue. Suitable antigens can come from different sources like overexpressed tumor-associated antigens, germline-derived tumor antigens and so-called neoantigens, which are the result of tumor-specific mutations (Vigneron, 2015; Mösch *et al.*, 2019). Especially the so-called cancer-testis antigens (CTAs) are often chosen as targets for adoptive T cell therapy based on TCRs, since they are shared among patients and cancer types, which means that a TCR recognizing a CTA can be used for multiple patients over multiple indications (Almeida *et al.*, 2009; Davari *et al.*, 2021). Neoantigens, on the other hand, are patient- and tumor-specific as they are generated by individual immunogenic mutations within the tumor. This has the advantage of reduced potential cross-reactivity but makes it necessary to develop treatments individually for each patient. For this reason, cancer immunotherapy targeting neoantigens are often vaccines, which are easier to produce specifically for a patient than generating a TCR therapy from healthy donor TCR repertoires (see 1.3.2, Ott *et al.* 2017; Sahin *et al.* 2017; Brennick *et al.* 2017).

1 Introduction

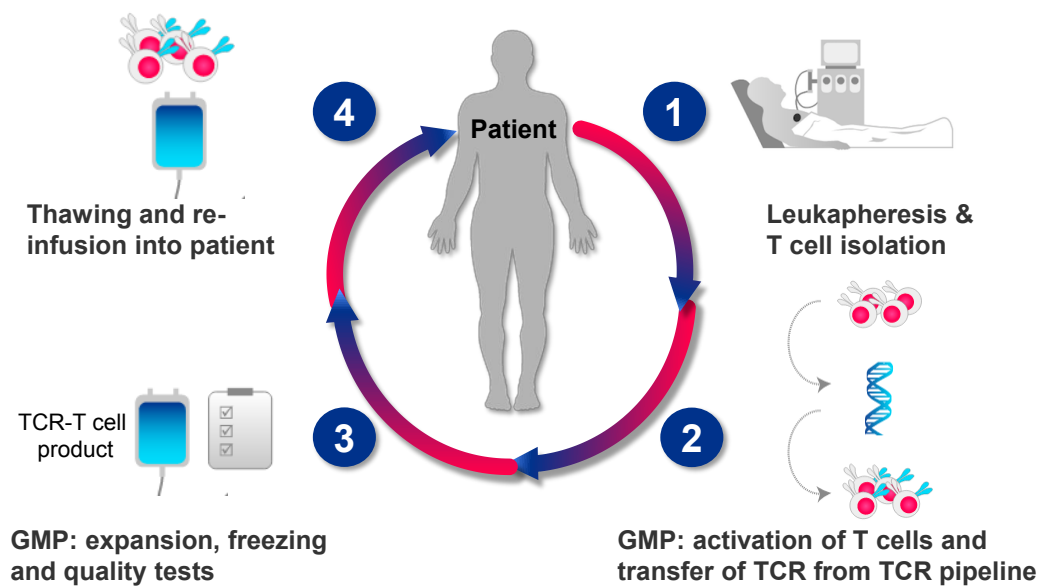


Figure 1.4: Personalized cancer treatment with TCRs. GMP means good manufacturing practice to ensure quality of the T cell product. Source: <https://www.medigene.de>, Annual Report Presentation 2018

1.3.2 Adoptive T cell therapy

Adoptive T cell therapy is based on the idea that a patient's immune response against the tumor can be boosted by modifying their T cells in a way that allows them to recognize cancer cells (Hammerl *et al.*, 2018). To achieve this, T cells are extracted from the patient's blood, cultured, receive an additional TCR and given back to the patient (see Figure 1.4). There are two different approaches regarding the type of receptor that is being added to the patient's T cells. The first one introduces a so called chimeric antigen receptor (CAR), which is an artificially designed receptor relying on antibody binding mechanisms (Sadelain *et al.*, 2013; Figueroa *et al.*, 2015). The second approach introduces an α/β T cell receptor derived from a natural T cell repertoire with optional modifications to enhance TCR pairing and surface expression as well as binding avidity. To obtain more potent TCRs, screening for candidates can be done by circumventing the thymic selection (see Figure 1.5, Wilde *et al.* 2009). In this case, T cells are challenged with peptides presented on MHC alleles which are different to the MHC alleles of the T cell donor.

1.3.3 Cross-reactivity

Unexpected cross-reactivity (CR) of TCRs introduced to a patient's T cells by adoptive T cell therapy can cause severe symptoms and even death (Linette *et al.*, 2013; Morgan *et al.*, 2013). There are two main types of CR, the first is on-target/off-tumor CR and the second is off-target CR. Both can have devastating consequences as in both cases healthy tissue is destroyed by the T cells. The difference is the antigen present on healthy cells which is recognized by the TCR. In the first case, the target antigen is not exclusively presented on tumor cells but also on healthy cells (Morgan *et al.*, 2013). In the second case, the TCR

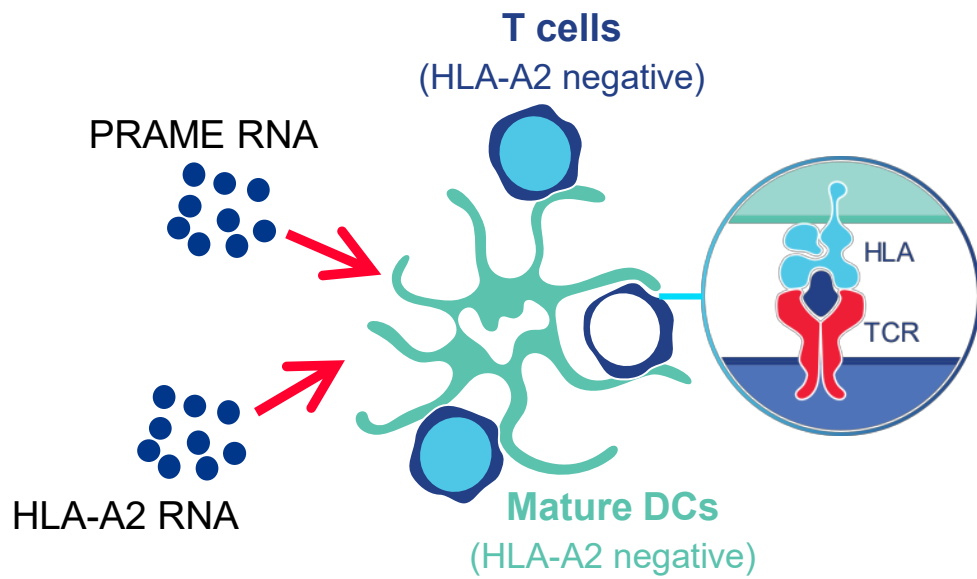


Figure 1.5: Priming of T cells with mature dendritic cells using PRAME as antigen and HLA-A2 as presenting MHC. Source: <https://www.medigene.de>, Half-Year Report Presentation 2021

1.4 Computational methods for immunology

recognizes an antigen presented on healthy cells which is similar but not identical to the target antigen (Linette *et al.*, 2013). To minimize the risk of bringing cross-reactive TCRs into the clinic, thorough screening of the candidate TCRs is necessary. To assist with this process *in silico*, the tool Expitope was developed and further improved, as more data became available (Jaravine *et al.*, 2017; Haase *et al.*, 2015). The input peptide sequence, which is the desired target epitope, is matched to a reference. Additionally, peptide sequences containing one or more mismatched amino acid positions are also identified from this reference. For all these peptide sequences, the probability of antigen processing and presentation as well as the expression of the associated protein and gene in healthy tissue is calculated. This allows quick identification of expression of the target peptide in healthy cells and provides information of potential cross-reactive peptides against which the TCR needs to be tested.

1.4 Computational methods for immunology

1.4.1 Epitope prediction

The focus of computational immunology has been on the ability to predict epitopes that elicit an immune response. Various computational methods, especially from the area of machine learning, have evolved alongside the growing availability of data. These methods also profited from the development of protein sequence-based methods and protein-peptide interaction models. One example is the usage of the diagonal of the BLOSUM matrix to encode peptide sequences, which was originally created to improve alignments of protein sequences to identify related proteins and is widely applied to epitope binding prediction meth-

ods (Henikoff and Henikoff, 1992; Andreatta and Nielsen, 2016).

The review “Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors” (see Chapter 3 and Appendix B, Mösch *et al.* 2019) describes how methods evolved which predict proteasomal cleavage, TAP transport and MHC binding and presentation. From the first sequence motif-based algorithms to highly advanced deep learning techniques like convolutional neuronal networks (CNNs), the review lists methods and presents the progress in gathering suitable training data from binding assays and MHC-eluted peptides analyzed by mass spectrometry.

Epitope prediction is also key to identifying neoepitopes, which are a highly relevant target for personalized cancer immunotherapy (see subsection 1.3.1). Pipelines are developed to determine patient-specific neoepitopes most likely to provoke an immune response if used in a vaccine or other immunotherapeutic treatments. These approaches often require only the patient’s sequencing information on DNA- and RNA-level to identify tumor-specific mutations, the patient’s HLA type, and expression levels of the mutated genes. This data will be processed and used for epitope prediction machine learning methods and neoepitope candidate rankings (Bjerregaard *et al.*, 2017; Hundal *et al.*, 2016; Bais *et al.*, 2017; Kim *et al.*, 2018).

1.4.2 Long short-term memory networks for sequence based T cell receptor predictions

Machine learning methods, especially neural networks for deep learning, are specialized for different tasks. CNNs are mainly used for image recognition and image classification tasks, whereas recurrent neural networks are more suited for temporal or sequence based predictions. A type of recurrent neural networks are long short-term memory (LSTM)

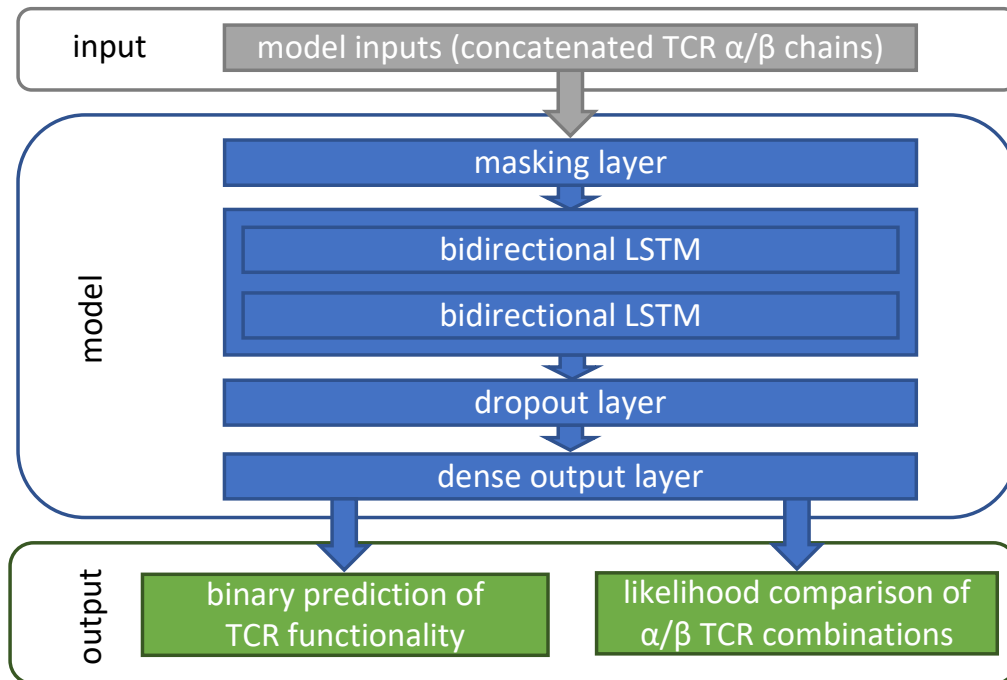


Figure 1.6: Deep learning model architecture of TCRpair using two bidirectional LSTMs. Source (slightly modified): Mösch and Frishman (2021)

1 Introduction

networks, which can remember information over a long time or a long sequence of amino acids. This makes them ideal for training on sequences with information spread across their whole length (Hanson *et al.*, 2016; Yamada and Kinoshita, 2018). LSTM layers are used for the TCRpair neural network architecture, as they ensure that relevant information from the beginning of the α chain of the TCR is kept until the end of the β chain (Mösch and Frishman, 2021). Additionally, the LSTM layers used for TCRpair are bidirectional, which means that the LSTM also runs backwards and can therefore learn information about any amino acid position with the knowledge of the sequence before and behind this position (see Figure 1.6). This helps to understand context, which is necessary for meaningful predictions using TCR sequence pairs as input.

2 Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues

2.1 Abstract

Background: Adoptive immunotherapy offers great potential for treating many types of cancer but its clinical application is hampered by cross-reactive T cell responses in healthy human tissues, representing serious safety risks for patients. We previously developed a computational tool called Expitope for assessing cross-reactivity (CR) of antigens based on tissue-specific gene expression. However, transcript abundance only indirectly indicates protein expression. The recent availability of proteome-wide human protein abundance information now facilitates a more direct approach for CR prediction. Here we present a new version 2.0 of Expitope, which computes all naturally possible epitopes of a peptide sequence and the corresponding CR indices using both protein and transcript abundance levels weighted by a proposed hierarchy of importance of various human tissues.

2 *Expitope 2.0*

Results: We tested the tool in two case studies: The first study quantitatively assessed the potential CR of the epitopes used for cancer immunotherapy. The second study evaluated HLA-A*02:01-restricted epitopes obtained from the Immune Epitope Database for different disease groups and demonstrated for the first time that there is a high variation in the background CR depending on the disease state of the host: compared to a healthy individual the CR index is on average two-fold higher for the autoimmune state, and five-fold higher for the cancer state.

Conclusions: The ability to predict potential side effects in normal tissues helps in the development and selection of safer antigens, enabling more successful immunotherapy of cancer and other diseases.

2.2 Contribution

For version 2.0 of the Expitope webserver, I developed the tissue scoring function described in section “Calculation of the tissue weighted CR-index”, which I also wrote. Additionally, I contributed to the analysis of the case studies, especially the section “IEDB epitopes”. For the full text of this publication see Appendix A.

3 Machine learning for cancer immunotherapies based on epitope recognition by T cell receptors

3.1 Abstract

In the last years, immunotherapies have shown tremendous success as treatments for multiple types of cancer. However, there are still many obstacles to overcome in order to increase response rates and identify effective therapies for every individual patient. Since there are many possibilities to boost a patient's immune response against a tumor and not all can be covered, this review is focused on T cell receptor-mediated therapies. CD8⁺ T cells can detect and destroy malignant cells by binding to peptides presented on cell surfaces by MHC (major histocompatibility complex) class I molecules. CD4⁺ T cells can also mediate powerful immune responses but their peptide recognition by MHC class II molecules is more complex, which is why the attention has been focused on CD8⁺ T cells. Therapies based on the power of T cells can, on the one hand, enhance T cell recognition by introducing TCRs that preferentially direct T cells to tumor sites (so called TCR-T therapy) or through vaccination to induce T cells *in vivo*. On the other hand, T cell activity can

be improved by immune checkpoint inhibition or other means that help create a microenvironment favorable for cytotoxic T cell activity. The manifold ways in which the immune system and cancer interact with each other require not only the use of large omics datasets from gene, to transcript, to protein, and to peptide but also make the application of machine learning methods inevitable. Currently, discovering and selecting suitable TCRs is a very costly and work intensive *in vitro* process. To facilitate this process and to additionally allow for highly personalized therapies that can simultaneously target multiple patient-specific antigens, especially neoepitopes, breakthrough computational methods for predicting antigen presentation and TCR binding are urgently required. Particularly, potential cross-reactivity is a major consideration since off-target toxicity can pose a major threat to patient safety. The current speed at which not only datasets grow and are made available to the public, but also at which new machine learning methods evolve, is assuring that computational approaches will be able to help to solve problems that immunotherapies are still facing.

3.2 Contribution

For this review, I planned the overall structure and individual sections and wrote most of the introduction, for which I also created Figure 1. For parts of the section “Prediction of T Cell Epitopes”, especially “Peptide-MHC Binding Prediction”, I contributed literature research and text. I wrote most of the section “Immunotherapy-specific Applications of Epitope Prediction”, especially “Neoepitope Identification”, for which I also did literature research and data analysis with results shown in Table 1 and Figure 3. I also participated in literature research and writing of the section “TCR Binding Prediction” and generated Figures 4 and 5. Furthermore, I contributed parts of the conclusion and outlook section

3.2 Contribution

and arranged all contributions from my co-authors for the review. All my co-authors have commented on and greatly helped with the polishing of the final manuscript. For the full text of this review see Appendix B.

4 TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T cell receptor α and β chains

4.1 Abstract

The ability of a T cell to recognize foreign peptides is defined by a single α and a single β hypervariable complementarity determining region (CDR3), which together form the T-cell receptor (TCR) heterodimer. In 30–35% of T cells, two α chains are expressed at the mRNA level but only one α chain is part of the functional TCR. This effect can also be observed for β chains, although it is less common. The identification of functional α/β chain pairs is instrumental in high-throughput characterization of therapeutic TCRs. TCRpair is the first method that predicts whether an α and β chain pair forms a functional, HLA-A*02:01 specific TCR without requiring the sequence of a recognized peptide. By taking additional amino acids flanking the CDR3 regions into account, TCRpair achieves an AUC of 0.71.

4.2 Contribution

The idea and concept of TCRpair was developed by me. I also designed, trained and evaluated the TCRpair algorithm of this publication. Furthermore, I gathered and preprocessed the data used for the training of TCRpair, designed and created Figures 1 and S1 and wrote the text with great support from my supervisor Prof. Dr. Dmitrij Frishman. The full text of this publication can be accessed online: Mösch A., Frishman D: TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T-cell receptor α and β chains. *Bioinformatics* 37 (21): 3038-3940 (2021), doi: 10.1093/bioinformatics/btab573.

5 Conclusion

Antigen presentation and recognition by cytotoxic T cells is an exquisitely complex biological process. Precise modeling of T cell antigen recognition by machine learning methods can reduce the costs and efforts required by *in vitro* methods. The tool Expitope, which identifies potential cross-reactive epitopes, is a good example for making use of such machine learning methods (Jaravine *et al.*, 2017). The quality-oriented collection and usage of data, especially on patterns of peptide processing and presentation and optimal T cell thriving conditions including T cell receptor formation, are paramount to improve the reliability of machine learning powered predictions. It is essential for the development of such prediction algorithms to be able to estimate quality, potential and limitations of different data sources. This applies to the difference between mass spectrometry derived peptide data and binding assays as well as various experimental methods to measure or identify T cell recognition (Mösch *et al.*, 2019).

As antigen processing and presentation was in the focus of machine learning applications for the last decades, the attention shifts to T cell receptors as more data becomes available, especially from single cell sequencing methods. There are, for example, needs to automate high-throughput identification of T cell receptor pairing, which we addressed with TCRpair (Mösch and Frishman, 2021), but many other applications of machine learning algorithms like TCR-peptide binding and cytotoxic qualities of T cells are possible. Different machine learning methods like

5 Conclusion

LSTMs, convolutional networks or combinations of several methods need to be explored for their suitability to predict these various features of T cell receptors, which ultimately will define the success of treatment in adoptive TCR-T cell therapy.

The developments of the past years, especially regarding the improvement of data availability and quality as well as the application of novel machine learning algorithms, demonstrate that the goal of patient specific cancer immunotherapies is achievable. Immunotherapies supported by computational methods can be expected to deliver convincing results and become an integral part of patient treatment.

Bibliography

- Almeida, L. G., Sakabe, N. J., deOliveira, A. R., Silva, M. C. C., Mundstein, A. S., Cohen, T., Chen, Y.-T., Chua, R., Gurung, S., Gnjatic, S., Jungbluth, A. A., Caballero, O. L., Bairoch, A., Kiesler, E., White, S. L., Simpson, A. J. G., Old, L. J., Camargo, A. A., and Vasconcelos, A. T. R. (2009). CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Research*, **37**(Database), D816–D819. Number: Database.
- Andreatta, M. and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, **32**(4), 511–517. Number: 4.
- Andreatta, M., Trolle, T., Yan, Z., Greenbaum, J. A., Peters, B., and Nielsen, M. (2017). An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics*.
- Arstila, T. P. (1999). A Direct Estimate of the Human T Cell Receptor Diversity. *Science*, **286**(5441), 958–961. Number: 5441.
- Bais, P., Namburi, S., Gatti, D. M., Zhang, X., and Chuang, J. H. (2017). CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics*, **33**(19), 3110–3112. Number: 19.
- Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z., and Eklund, A. C. (2017). MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunology, Immunotherapy*, **66**(9), 1123–1130. Number: 9.
- Boegel, S., Löwer, M., Bukur, T., Sahin, U., and Castle, J. C. (2014). A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *OncoImmunology*, **3**(8), e954893. Number: 8.
- Brennick, C. A., George, M. M., Corwin, W. L., Srivastava, P. K., and Ebrahimi-Nik, H. (2017). Neoepitopes as cancer immunotherapy targets: key challenges and opportunities. *Immunotherapy*, **9**(4), 361–371. Number: 4.
- Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P., and Boon, T. (2014). Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nature Reviews Cancer*, **14**(2), 135–146. Number: 2.
- Davari, K., Holland, T., Prassmayer, L., Longinotti, G., Ganley, K. P., Pechilis, L. J., Diaconu, I., Nambiar, P. R., Magee, M. S., Schendel, D. J., Sommermeyer, D., and Ellinger, C. (2021). Development of a CD8 co-receptor independent T-cell receptor specific for tumor-associated antigen MAGE-A4 for next generation T-cell-based immunotherapy. *Journal for ImmunoTherapy of Cancer*, **9**(3), e002035.

BIBLIOGRAPHY

- Figueroa, J. A., Reidy, A., Mirandola, L., Trotter, K., Suvorava, N., Figueroa, A., Konala, V., Aulakh, A., Littlefield, L., Grizzi, F., Rahman, R. L., R. Jenkins, M., Musgrove, B., Radhi, S., D’Cunha, N., D’Cunha, L. N., Hermonat, P. L., Cobos, E., and Chiriva-Internati, M. (2015). Chimeric Antigen Receptor Engineering: A Right Step in the Evolution of Adoptive Cellular Immunotherapy. *International Reviews of Immunology*, **34**(2), 154–187. Number: 2.
- Gragert, L., Madbouly, A., Freeman, J., and Maiers, M. (2013). Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology*, **74**(10), 1313–1320. Number: 10.
- Haase, K., Raffegerst, S., Schendel, D. J., and Frishman, D. (2015). Expitope: a web server for epitope expression. *Bioinformatics*, **31**(11), 1854–1856. Number: 11.
- Hammerl, D., Rieder, D., Martens, J. W., Trajanoski, Z., and Debets, R. (2018). Adoptive T Cell Therapy: New Avenues Leading to Safe Targets and Powerful Allies. *Trends in Immunology*, **39**(11), 921–936. Number: 11.
- Hanson, J., Yang, Y., Paliwal, K., and Zhou, Y. (2016). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, page btw678.
- Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, **89**(22), 10915–10919. Number: 22.
- Hu, Z., Ott, P. A., and Wu, C. J. (2017). Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nature Reviews Immunology*.
- Hughes, M., Yassai, M., Sedy, J., Wehrly, T., Huang, C.-Y., Kanagawa, O., Gorski, J., and Sleckman, B. (2003). T cell receptor CDR3 loop length repertoire is determined primarily by features of the V(D)J recombination reaction. *European Journal of Immunology*, **33**(6), 1568–1575. Number: 6.
- Hundal, J., Carreno, B. M., Petti, A. A., Linette, G. P., Griffith, O. L., Mardis, E. R., and Griffith, M. (2016). pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Medicine*, **8**(1). Number: 1.
- Jaravine, V., Mösch, A., Raffegerst, S., Schendel, D. J., and Frishman, D. (2017). Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. *BMC Cancer*, **17**(1), 892. Number: 1.
- Kim, S., Kim, H. S., Kim, E., Lee, M. G., Shin, E., Paik, S., and Kim, S. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Annals of Oncology*, **29**(4), 1030–1036. Number: 4.
- Leone, P., Shin, E.-C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). MHC Class I Antigen Processing and Presenting Machinery: Organization, Function, and Defects in Tumor Cells. *JNCI Journal of the National Cancer Institute*, **105**(16), 1172–1187. Number: 16.
- Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., Litzky, L., Bagg, A., Carreno, B. M., Cimino, P. J., Binder-Scholl, G. K., Smethurst, D. P., Gerry, A. B., Pumphrey, N. J., Bennett, A. D., Brewer, J. E., Dukes, J., Harper, J., Tayton-Martin, H. K., Jakobsen, B. K., Hassan, N. J., Kalos, M., and June, C. H. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, **122**(6), 863–871. Number: 6.

BIBLIOGRAPHY


- Morgan, R. A., Chinnasamy, N., Abate-Daga, D., Gros, A., Robbins, P. F., Zheng, Z., Dudley, M. E., Feldman, S. A., Yang, J. C., Sherry, R. M., Phan, G. Q., Hughes, M. S., Kammula, U. S., Miller, A. D., Hessman, C. J., Stewart, A. A., Restifo, N. P., Quezado, M. M., Alimchandani, M., Rosenberg, A. Z., Nath, A., Wang, T., Bielekova, B., Wuest, S. C., Akula, N., McMahon, F. J., Wilde, S., Mosetter, B., Schendel, D. J., Laurencot, C. M., and Rosenberg, S. A. (2013). Cancer Regression and Neurological Toxicity Following Anti-MAGE-A3 TCR Gene Therapy. *Journal of Immunotherapy*, **36**(2), 133–151. Number: 2.
- Mösch, A. and Frishman, D. (2021). TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T-cell receptor and chains. *Bioinformatics*, **37**(21), 3938–3940.
- Mösch, A., Raffegerst, S., Weis, M., Schendel, D. J., and Frishman, D. (2019). Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. *Frontiers in Genetics*, **10**.
- Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., Chen, C., Olive, O., Carter, T. A., Li, S., Lieb, D. J., Eisenhaure, T., Gjini, E., Stevens, J., Lane, W. J., Javeri, I., Nellaiappan, K., Salazar, A. M., Daley, H., Seaman, M., Buchbinder, E. I., Yoon, C. H., Harden, M., Lennon, N., Gabriel, S., Rodig, S. J., Barouch, D. H., Aster, J. C., Getz, G., Wucherpfennig, K., Neubergh, D., Ritz, J., Lander, E. S., Fritsch, E. F., Hacohen, N., and Wu, C. J. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, **547**(7662), 217–221. Number: 7662.
- Peters, B., Bulik, S., Tampe, R., van Endert, P. M., and Holzhutter, H.-G. (2003). Identifying MHC Class I Epitopes by Predicting the TAP Transport Efficiency of Epitope Precursors. *The Journal of Immunology*, **171**(4), 1741–1749. Number: 4.
- Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Research*, **43**(D1), D423–D431. Number: D1.
- Sadelain, M., Brentjens, R., and Rivière, I. (2013). The Basic Principles of Chimeric Antigen Receptor Design. *Cancer Discovery*, **3**(4), 388–398. Number: 4.
- Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., Bukur, V., Tadmor, A. D., Luxemburger, U., Schrörs, B., Omokoko, T., Vormehr, M., Albrecht, C., Paruzynski, A., Kuhn, A. N., Buck, J., Heesch, S., Schreeb, K. H., Müller, F., Ortseifer, I., Vogler, I., Godehardt, E., Attig, S., Rae, R., Breitkreuz, A., Tolliver, C., Suchan, M., Martic, G., Hohberger, A., Sorn, P., Diekmann, J., Ciesla, J., Waksman, O., Brück, A.-K., Witt, M., Zillgen, M., Rothermel, A., Kasemann, B., Langer, D., Bolte, S., Diken, M., Kreiter, S., Nemecek, R., Gebhardt, C., Grabbe, S., Höller, C., Utikal, J., Huber, C., Loquai, C., and Türeci, (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, **547**(7662), 222–226. Number: 7662.
- Vigneron, N. (2015). Human Tumor Antigens and Cancer Immunotherapy. *BioMed Research International*, **2015**, 1–17.
- Wilde, S., Sommermeyer, D., Frankenberger, B., Schiemann, M., Milosevic, S., Spranger, S., Pohla, H., Uckert, W., Busch, D. H., and Schendel, D. J. (2009). Dendritic cells pulsed with RNA encoding allogeneic MHC and antigen induce T cells with superior antitumor activity and higher TCR functional avidity. *Blood*, **114**(10), 2131–2139. Number: 10.
- Yamada, K. D. and Kinoshita, K. (2018). De novo profile generation based on sequence context specificity with the long short-term memory network. *BMC Bioinformatics*, **19**(1). Number: 1.
- Yates, A. J. (2014). Theories and Quantification of Thymic Selection. *Frontiers in Immunology*, **5**.

SOFTWARE

Open Access



Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues

Victor Jaravine^{1,2}, Anja Mösch^{1,2}, Silke Raffegerst², Dolores J. Schendel² and Dmitrij Frishman^{1,3*} 

Abstract

Background: Adoptive immunotherapy offers great potential for treating many types of cancer but its clinical application is hampered by cross-reactive T cell responses in healthy human tissues, representing serious safety risks for patients. We previously developed a computational tool called Expitope for assessing cross-reactivity (CR) of antigens based on tissue-specific gene expression. However, transcript abundance only indirectly indicates protein expression. The recent availability of proteome-wide human protein abundance information now facilitates a more direct approach for CR prediction. Here we present a new version 2.0 of Expitope, which computes all naturally possible epitopes of a peptide sequence and the corresponding CR indices using both protein and transcript abundance levels weighted by a proposed hierarchy of importance of various human tissues.

Results: We tested the tool in two case studies: The first study quantitatively assessed the potential CR of the epitopes used for cancer immunotherapy. The second study evaluated HLA-A*02:01-restricted epitopes obtained from the Immune Epitope Database for different disease groups and demonstrated for the first time that there is a high variation in the background CR depending on the disease state of the host: compared to a healthy individual the CR index is on average two-fold higher for the autoimmune state, and five-fold higher for the cancer state.

Conclusions: The ability to predict potential side effects in normal tissues helps in the development and selection of safer antigens, enabling more successful immunotherapy of cancer and other diseases.

Keywords: Cancer, Immunotherapy, Tumor immunology, Cross-reactivity, T cell epitope, Immunoinformatics, Tumor antigen expression

Background

The principles of how the immune system can optimally control infections and early stages of cancer underpin the development of immunotherapies. Among these approaches, adoptive transfer of antigen-specific T cells is emerging as a particularly attractive form of immunotherapy to treat patients with more advanced stages of cancer and unresolved infectious diseases. This approach utilizes transfer of tailored antigen-specific immune T cells and

provides the possibility of clinically efficient treatment of infectious diseases and human malignancies [1].

One major stumbling block precluding wider application of adoptive immunotherapy is the occurrence of adverse effects of off-target cross-reactivity (CR), which may result in significant, even lethal, toxicity. The cause of toxicity is a hyper-activated T cell response with reactivity directed against normal tissue [2]. Immune CR arises when T cells recognizing a selected target epitope are transferred back to the patient and exhibit recognition of self-epitopes in non-cancerous tissues. On the molecular level this effect is usually the consequence of a high degree of sequence similarity between the target and the

*Correspondence: d.frishman@wzw.tum.de

¹Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany

³St Petersburg State Polytechnical University, 195251 St Petersburg, Russia
Full list of author information is available at the end of the article

self-epitopes, resulting in the binding of a stable self-peptide-MHC complex to the T cell receptor (TCR) and, consequently, cross-activation of unwanted autoimmune T cell responses [3]. Depending on the sequence similarity there can be on-target/off-tumor or off-target recognition. The former is directed against the identical epitope that is also present in a non-cancerous tissue, while the latter is directed against a similar epitope also present in a healthy tissue. The ability to predict the scope and extent of on- and off-target effects can help in selection of safer antigens, and consequently enable more successful immunotherapy treatment [4].

A computational strategy for the prediction of potential peptide-HLA cancer targets and evaluation of the likelihood of off-target toxicity for the targets was developed by Dhanik et al. [5]. The strategy utilizes a sequence-based algorithm similar to the one used in our previous studies [6] and in our current work, but it is not available as a web-service.

We have developed the Expitope server as a tool to assess epitope expression in various tissues (freely accessible at <http://webclu.bio.wzw.tum.de/expitope2>). Expitope incorporates the most recent genome-wide information, including protein sequences and protein abundance data across various tissues and cell lines. It enables researchers to screen their epitopes in silico for potential CR in human tissues, before moving their therapeutic candidates into clinical trials.

Approach

CR to an immunotherapeutic epitope may arise if a protein normally expressed in healthy cells is cleaved by one of the proteasomes to produce a peptide with an amino acid sequence that is similar to the given epitope. Another prerequisite for CR is the presentation of the natural epitope by major histocompatibility complex class I molecules (MHC-I) in various tissues. We model this process by the method described by Keşmir et al. [7]. To quantitatively assess the natural occurrence of epitopes, we use experimental data on gene expression and abundance of proteins in which the epitopes are present. The methods are described in detail in our previous publication [8] on

the iCrossR tool, which has been merged into the current version 2.0 of Expitope. The iCrossR project's aim was to perform a quantitative characterization study of all MHC-I epitopes listed in the cancer immunotherapy database. A new feature of Expitope 2.0 is the calculation of the tissue-weighted cross-reactivity (CR) indices. Below we test the approach and provide information on the new data sources and a new tissue-weighted CR-index formula.

Material and Implementation

Gene and protein expression data

The previous version 1.0 of Expitope [6] assessed the expression of human antigens based on one combined gene expression database [9] and the Illumina Body Map database [10]. Interestingly, HLA-typing of samples from the Illumina Body Map and Wang et al. [9] showed that the tissues used for expression analysis are most likely derived from the same individual except for seven brain samples [11]. In order to avoid data redundancy with the new Illumina Body Map database, we now only use the brain expression data from Wang et al. [9]. The new version 2.0 of Expitope incorporates three gene expression and four protein abundance datasets (Table 1). It should be noted that in contrast to the PaxDB and Human Proteome Map datasets, which contain ppm values, the Human Protein Atlas data has been generated by immunohistochemistry, which makes the accuracy of the data dependent on the specificity of the antibodies used. The values range from 0 to 3, indicating no detectable expression (0) up to high expression (3).

IEDB datasets

We selected four groups of peptides (Table 2) from the Immune Epitope Database (IEDB) [12], containing a total of 1720 epitopes of 7-25 amino acids in length (Additional file 1: Table S1, Additional file 2: Table S2, Additional file 3: Table S3, Additional file 4: Table S4). The selection for all groups was restricted to the following tags: 'human HLA-A*02:01', 'Linear Epitopes', 'Positive Assays only', 'T cells Assays', 'MHC ligand Assays', 'No B-cell assays', 'Host: Homo Sapiens (Human)', from which the selection was

Table 1 Sources of gene expression and protein abundance data

Data source	ID	Name	Number of tissues	Type	References
PaxDB	Pax4	PaxDB v4.0	22	Protein abundance	[24]
Expression Atlas	E-Prot-3	Human Protein Atlas	44	Protein abundance	[25, 26]
Expression Atlas	E-Prot-1	Human Proteome Map	23	Protein abundance	[25, 27]
Expression Atlas	E-Mtab-513	Illumina Body Map	16	Gene expression	[10, 25]
Expression Atlas	E-Mtab-5214	GTEx	53	Gene expression	[25, 28]
Wang et al. 2008	Wang	Wang 2008	7	Gene expression	[9]
Expression Atlas	E-Mtab-3358	FANTOM5 RIKEN	56	Gene expression	[25, 29]

Table 2 Four epitope groups from the IEDB database

Group	ID in IEDB	Disease state of host	Number of entries	Peptide length range (average)
1	DOID:0050117	Infectious diseases	588	8-20 (9)
2	DTREE_00000014	Healthy (no disease)	461	8-25 (10)
3	DOID:417	Autoimmune diseases	155	8-21 (10)
4	DOID:162	Cancer	516	7-25 (11)

further restricted for each of the four groups using the tag corresponding to a disease state of the host (column 3 of Table 2).

Identification of natural epitopes

Amino acid sequences of epitopes were matched against the RefSeq database [13] of all naturally occurring human protein sequences, including annotated isoforms, downloaded from the National Center for Biotechnology Information (NCBI). The matching procedure yields a list of protein segments, which we call “natural epitopes” (NEs). Potential immunogenicity of each NE was calculated using the formula developed by Keşmir et al. [7], which combines the predicted scores for proteasomal cleavage, TAP affinity and MHC-binding predictions. The quantitative score *Q* of epitope presentation on MHC-I is defined as:

$$Q = P_{CL} / (A_{TAP} * A_{MHC}) \tag{1}$$

where *P_{CL}* is the proteasomal cleavage probability, while *A_{TAP}* and *A_{MHC}* are the IC₅₀-affinities to the transporter molecule associated with antigen processing (TAP) and to the MHC complex, respectively. Lower values for *A_{TAP}* and *A_{MHC}* correspond to higher predicted affinities, as IC₅₀-affinity is defined as a dose of peptide that displaces 50% of a competitive ligand.

Calculation of the tissue weighted CR-index

In this version, we modified the CR-index calculation formula [8] to include tissue weighting, reflecting the perceived importance of different tissue types in the human body. For each database, the tissue profile *S(t)* for a given epitope was calculated as follows:

$$S(t) = \sum_{k=0}^K \left\{ \nu(k) \cdot \log_{10} \left[\sum_{i=1}^{M(k)} a(i, t) \right] \right\} \tag{2}$$

where *k* is the allowed number of mismatches and *K* is the maximal *k*; *t* is the tissue index in a given database of *T* tissues; *i* is the running index in the list of matching NEs for each *k*, and *M(k)* is the size of the list; *ν(k)* is the normalized mismatch weight, and *a(i,t)* is the protein or transcript abundance in the tissue *t* corresponding to the *i*-th NE. The sum over *i* includes only the unique NEs that have the scores *Q(i)* (Equation 1) above a chosen threshold. The normalized mismatch weight is calculated as $\nu(k) = (1/P(k)) / \sum_k (1/P(k))$, where *P(k)* is the probability

of finding a random peptide of length *l* with *k* mismatches in our protein sequence database of the total length of *N*=6.5e7 amino acids, $P(k) = 1 - (1 - 0.05^{l-k})^{N-l+1}$. For example, for a peptide of length 9, the mismatch weights are: $\nu(k=0,1,2,3) = 0.95, 0.0475, 0.0023, 0.0002$.

The weighted CR-index is defined as a tissue-weighted average of the tissue profiles *S(t)*:

$$I_{CR} = \frac{1}{\sum_t^T w(t)} \sum_t^T w(t) S(t) \tag{3}$$

where *w(t)* represents the weight assigned to the tissue type *t* (Table 3). The *I_{CR}* index error is obtained as one standard deviation from the mean upon bootstrapping, which involves repeating index calculation 10 times using 90% of randomly subsampled data. The weight values range between 0 and 1, with the weight of 1 corresponding to the most vital organs and systems according to the Sequential Organ Failure Assessment (SOFA) score used to evaluate the condition of patients in Intensive Care Units (ICUs) [14]. The second highest weight of 0.8 is assigned to tissues that belong to vital organs where a failure does not immediately threaten a patient’s life. A weight of 0.5 is assigned to tissues where CR is not necessarily life threatening, but can nevertheless cause severe complications. The second lowest weight of 0.3 refers to tissues and organs that can be surgically removed without major complications. Finally, the weight of 0 was assigned to irrelevant tissues such as testis, where expression of an antigen does not cause an immune response, as well as to the tissues that are only present during pregnancy and other samples that do not correspond to healthy human tissue, e.g. cancer cell lines.

Consequently, large *I_{CR}* values may indicate potentially life-threatening CR of the epitope. The higher the number of hits to different NEs that are close in sequence to a therapy peptide, and their total abundance/expression levels in the tissues with high weights, the higher is the probability of CR. Higher thresholds for *Q* correspond to choosing a higher probability of the selected natural epitope to be immunogenic, while the parameter *K* controls the sequence similarity: exact match (*K*=0) for prediction of on-target/off-tumor recognition, and *K*> 0 for off-target recognition. The values of these parameters can be set by Expitope users. In this work, we chose *K*=1, i.e. up to one mismatch in amino-acid sequence, and two

Table 3 Weight values and categorization of tissue types

<i>Consequence</i>	<i>Damage immediately life threatening</i>	<i>Damage life threatening</i>	<i>Damage not immediately life threatening</i>
<i>Weight</i>	1	0.8	0.5
<i>Tissues</i>	Lung/Respiratory system Brain/Nervous system Blood/Immune system Heart Kidney Liver	Digestive system (except appendix) Soft tissue	Urinary bladder Various glands Prostate Skin Eye ^a
<i>Consequence</i>	<i>Damage not life threatening</i>	<i>Tissue not affected</i>	
<i>Weight</i>	0.3	0	
<i>Tissues</i>	Reproductive organs Mammary tissue Tonsils Appendix Gall bladder Spleen	Cancer cell lines Testis Fetal tissue	

^aThe weight for eye tissue is set to 0.5, as T cells are able to infiltrate it [30]

thresholds for Q : 0.02 corresponding to top 10% immunogenic NEs found for all epitopes in this study, and $1e-4$ corresponding to top 50% of the NEs, i.e. top-scored for proteasomal cleavage, TAP transport and MHC-I binding. However, calculation of the indices with the numbers of mismatches $K=[0,3]$ and the combined scores $Q=0.02$, $1e-4$, $1e-5$ gave very similar results (Additional file 5: Tables S5-S7; Figure S2).

While a high I_{CR} means that severe complications are expected for a target epitope, its low value hints towards minor or non-life-threatening side effects. An index greater than zero always means that there is some expression present that should be investigated in detail. The index is only an estimate, which does not take into account many patient-specific factors, and therefore should not be used as the sole measure for making decisions. As the tissue classification is not exhaustive and not all organs are completely represented by the tissue types of which they consist, a high expression value in a low rated tissue could correspond to a tissue type not covered, but also present in other more vital organs. Nonetheless, the weighted index offers a short summary of the rather extensive result tables that are produced by Expitope 2.0, and contain individual expression values for each tissue and all NEs. Therefore, the weighted index allows for quick rejection of target epitopes that are likely to cause severe side effects caused by CR.

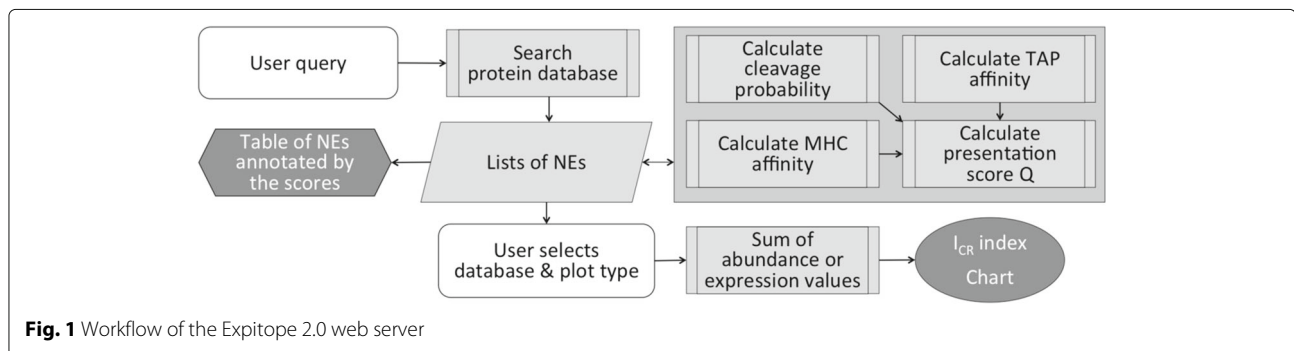
The I_{CR} indices were calculated with the default parameters (except Q and K) for each peptide and each database

using Eq. 3, and were averaged over the seven databases to give the average I_{CR} indices for each peptide. For the plots the I_{CR} indices are averaged for all peptides in each group.

Web server

Expitope 2.0 is a web application that can be easily used by the researchers inexperienced in bioinformatics, especially from the immunotherapy domain. There is no login requirement to the website and user IP addresses are not stored. Multiple clients can connect to the server, and concurrent clients are served one query at a time. The jobs are submitted to high-performance computational infrastructure. The results are displayed once they are ready; alternatively the user can return to the results later, using the session URL. It is also possible to download the results as a spreadsheet to be used with Microsoft Excel or similar software. This allows to sort and filter the results according to individual criteria, e.g. for sorting epitopes by binding affinity predicted by netMHC.

The workflow of Expitope is shown in Fig. 1. The user inputs a peptide sequence and specifies parameters for sequence matching and for the computation of MHC class I binding affinity via the html forms displayed in a web-browser (white). The server performs the search for natural epitopes (NEs) and calculates their Q scores. Computations are performed by the client process at the backend of the server (large gray rectangle). Results are returned to the user in the form of text files and graphical visualizations (dark gray). The user selects a particular



database and a plot type for visualization (white). The parameters that can be changed by users in the forms have the following default values: the TAP weight is 0.2, the cleavage threshold is 0.7, the Q score threshold is $1e-4$ and the number of mismatches is 2.

Results and discussion

Known cross-reactive epitopes

For the first version of the Expitope web server, the MAGEA3 epitope EVDPIGHLY was tested that had been associated with cross-reactivity caused by the TCR recognizing an epitope with four mismatches derived from titin, which is expressed in heart muscle tissue [6, 15]. We were able to reproduce these findings by using Expitope 2.0 with default the parameters except for allowing up to four mismatches and additionally, the newly added protein databases showed an even clearer result with values of $2.98e+03$ ppm (PaxDB) and $2.86e+03$ ppm (Human Proteome Map) and the maximum value of 3 for the Human Protein Atlas. Another case of observed cross-reactivity has been a TCR recognizing the MAGEA3/MAGEA9 epitope KVAELVHFL [16]. Expitope 2.0 with the default parameters finds this and all other epitopes from various members of the MAGE family the TCR was able to detect. This includes one epitope of MAGEA12, which was found to be expressed in brain where it led to cross-reactivity. We found expression values of 0.2 FPKM and less but no protein expression for MAGEA12, which is also not contained in the Human Protein Atlas and Human Proteome Map. This demonstrates the importance of taking even small amounts of expression into account when assessing potential cross-reactivity and also comparing the results obtained from all databases, especially for crucial tissues like heart, brain and lung.

Case studies

Cancer immunity peptides

Here we provide an overview of our previous study [8], where we analyzed short (8-15 amino acids) peptide sequences from the Cancer Immunity Peptide Database [17] as well as peptides of viral origin. The CR-index

calculation was based only on the PaxDB protein abundance database and without tissue weighting.

The peptide dataset consisted of four groups of currently known human MHC class I epitopes including: mutation antigens displayed by tumor cells (40 peptides, group A), cancer-testis (CT) antigens (67 peptides, group B), differentiation antigens (57 peptides, group C), and overexpressed proteins (94 peptides, group D). In addition, 89 epitopes originating from viral sources (group E) were investigated. When matched exactly, the group of “mutation” antigens produced no hits to the proteins normally expressed in human tissues, since the epitopes of the group have sequences that originated from mutations of normal human protein sequences. The second validation is from the CT antigens, which at small numbers of mismatches (0-1), showed few matches to proteins expressed in the majority of human tissues, with the expected exception of ovary/testis, where multiple hits were found. The hit patterns were very similar for all epitopes of this group. This is exactly as expected, since CT antigens are expressed mostly in these two tissues. In contrast to the results for groups A and B, the antigens of the groups C and D showed more hits, both for exact matches and for high numbers of mismatches. This is also as expected as the proteins containing the epitopes are expressed in a wide variety of normal tissues. Finally, the epitopes originating from the viral sources showed noticeably fewer matches to the human proteins compared to the cancer peptides.

IEDB epitopes

We sought to assess quantitatively the extent of potential “background” CR of the epitopes derived from the host individuals having different disease states - ranging from healthy to cancer. Such background CR is not caused by one single therapy but accumulates due to many factors, including an unknown history of diseases.

The I_{CR} indices of individual epitopes calculated across the seven databases used in this study are highly correlated, since for each database they are obtained by summation of the abundance (or expression) values for

the same proteins. There are high correlations between the I_{CR} values computed for the peptides using the three abundance databases as well as between the I_{CR} values derived from the four expression databases (data not shown). Similarly, the correlations between the abundance and expression indices are high (Additional file 5: Figure S1), with the Pearson's coefficients in the range 0.94-0.96. Averaging of the indices allows one to obtain a more accurate prediction of CR due to increased signal-to-noise ratio, as the databases are derived from different data sources.

Figure 2 shows the I_{CR} indices for the four epitope groups described in Table 2 (group I_{CR} indices before averaging by databases can be found in Additional file 5: Tables S5-S7). The indices for the epitopes computed from 10% top-scoring NEs ($Q=0.02$, Fig. 2 left) are on average 3-times lower, compared to those from 50% top-scoring NEs ($Q=1e-4$, Fig. 2 right), corresponding to lower numbers of matching NEs. Higher thresholds for Q correspond to a higher probability of the selected NEs to be immunogenic. It has been reported that the top-scoring 7-10% epitopes identified by the immunogenicity prediction methods have 85% probability of being immunogenic [18]. In this work we have chosen two thresholds of 10% and 50% of sequence matches. The rationale for this choice was to ensure a low amount of false positives in the immunogenicity prediction for the 10% I_{CR} index, and to compare it with the 50% value containing medium to high immunogenic peptides. Two groups - 'Infectious diseases' and 'Healthy' - have average indices close to zero on both plots, indicating low amounts of cross-reactive epitopes in the critical tissues. The groups 'Autoimmune diseases' and 'Cancer' exhibit approximately 2- to 5-fold higher average index values compared to the 'Healthy' group, in each plot respectively, corresponding to considerably higher presentation level of the cross-reactive peptides in these states.

The interpretation of these results is as follows. The epitopes in the 'Infectious diseases' group are derived from non-human organisms rather than from human hosts. Thus, compared to the epitopes from the other three groups, which are of human origin, a lower I_{CR} index is expected, implying low sequence identity to the host and thus a low probability of CR. The slightly elevated index for the 'Healthy' group is most likely due to the presence of common pathogens (such as Herpes simplex virus or Epstein-Barr virus) mimicking human sequences, an immune escape strategy known as immune camouflage [19]. A higher I_{CR} for the 'Autoimmune' group compared to the 'Healthy' group is not surprising, as autoimmunity is a response of the human body's immune system directed against human proteins overexpressed or aberrantly presented in healthy tissues. For example, multiple sclerosis, the most frequently occurring disease in this group, is due to autoimmunity to the myelin basic protein (MBP), expressed in the tissues of the central nervous system [19]. Other epitopes in this group with very high index values are derived, e.g. from the proteins actin, myosin-9, septin-2 and vimentin, which are normally expressed in various tissues. Normally, peripheral T cells are trained to recognize pathogen-derived epitopes and ignore self-antigens, however some T cells escape this selection and are able to recognize self-antigens, thus initiating an autoimmune response and becoming self-reactive. Consequently with respect to autoimmunity, the term CR is defined as the recognition by T cell TCRs of many different peptide antigens, presented by the HLA of an individual [20], which can also be referred to as cross-recognition.

The significantly higher CR index for the cancer group compared to the other three groups indicates a presence of a high background level of CR when targeting cancers. Cancer epitopes originate either from wild-type proteins overexpressed in tumors, or as a result of cancer-specific

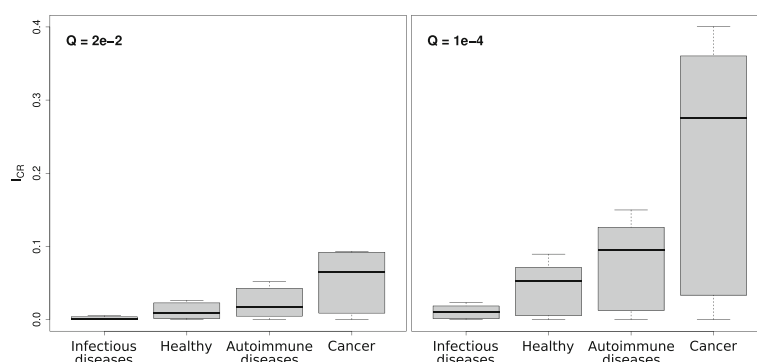


Fig. 2 The I_{CR} indices for the four IEDB peptide groups (Table 2), obtained by averaging over the seven databases listed in Table 1. $Q=2e-2$ (left), $Q=1e-4$ (right), with up to one mismatch ($K=1$). Thick black line: median; gray: the lower and the upper quartiles (25th and 75th percentiles); upper and lower whiskers: highest and lowest values

mutations in the genes, named neoepitopes. On average, neoepitopes have lower similarity to self-antigens compared to the wild-type cancer epitopes, thus potentially are less cross-reactive. Since T cells with TCRs binding to self-antigens are negatively selected in the thymus, there will generally be a lack of the T cells that can fight tumors, producing overexpressed wild-type proteins. In contrast, the cancers producing neoepitopes can be effectively controlled by the immune system provided that suitable T cells are available. Thus, different types of cancer produce epitopes of varying cross-reactivity, which explains the larger variance seen in Fig. 2 for the cancer group compared to the other groups.

High I_{CR} for the 'Autoimmune disease' and 'Cancer' groups may also be due to an activated state of the immune system, when immunoproteasomes create larger amounts of immunogenic (in comparison to standard proteasomes) epitopes, including those from the residuals of normal cells killed by the immune system [21]. In addition, disruption of the normal functioning of the ubiquitin proteasome system may result in creation of abnormally presented immunogenic epitopes, leading to many types of disorders, including malignancies, neurodegenerative diseases and systemic autoimmunity [22, 23].

Thus, multiple reasons for a high variability in presented CR epitopes appear to exist depending on the host disease state. This CR, which we tentatively call "background" CR, is independent of any immune therapy. Clearly, a collection of epitopes present in a particular individual is different from our datasets obtained from the IEDB database. Likely, it will include only a subset of the peptides, but a statistical distribution in many patients may exhibit a pattern similar to the one reported in this work. Eventually, it remains to be seen if there can be any interference between the background CR and the CR invoked by a therapy, but both types are important to assess the safety of the therapy.

Conclusion

It is a long-standing dream of many medical practitioners to use the immune system for effective treatment and permanent cure of human disease conditions. With the number of tested and approved immunotherapies growing, evidence of the side effects associated with the current therapies also increased. Consequently, therapy developers require reliable tools for predicting unwanted cross-reactions.

The Expitope web tool for predicting CR of T cell epitopes is based on experimental protein abundance and expression data obtained from a growing number of publicly available databases. We demonstrate its performance for a large number of epitopes detected in the human organism for various cancer types and at various disease states, ranging from healthy to cancer. The results of our

study of Cancer Immunity Peptides [8] showed that the currently known cancer epitopes display a very large CR variability across a range of tissues. Our predictions are in close agreement with the results of several clinical studies, with the CR indices being high in the tissues where actual side effects have been reported, and close to zero for no side-effects. Thus, Expitope enables researchers to assess potential side effects of their selected antigens for therapy and to identify specific human tissues where such side effects could be expected. Since any immunotherapy can cause side effects, we suggest using this tool at both early and late stages of a therapy development process. CR index values calculated by Expitope can serve as an estimate of the amount of potential CR for *in silico* assessment of immunotherapeutic strategies.

For the first time we demonstrate that there is a high variation in the CR of peptides presented at different disease states of the host: it is on average 2-fold higher for individuals with an autoimmune state and 5-fold higher for individuals with cancer in comparison to individuals in an apparent healthy state. Presumably, a similar background CR may exist prior to an immune therapy, which may differ by the host disease state. Since the human organism negatively pre-selects T cells binding to self-antigens, there will be a small number or no T cells fighting disease tissue cells marked by highly cross-reactive epitopes. Consequently, the similarity of presented epitopes to self-antigens is an obstacle for disease elimination both for the organism itself and for immunotherapy. Thus, therapy developers should consider the possibility of background CR interfering with a therapy.

Availability and requirements

Project name: Expitope 2.0

Project home page: <http://webclu.bio.wzw.tum.de/expitope2>

Operating system(s): Platform independent

Programming language: Java, JavaScript

Other requirements: Web browser

License: None (free to use for academic purposes)

Any restrictions to use by non-academics: None

Additional files

Additional file 1: TableS1_InfectiousDisease. Comma-separated table containing Table S1. (CSV 80 kb)

Additional file 2: TableS2_Healthy. Comma-separated table containing Table S2. (CSV 67 kb)

Additional file 3: TableS3_AutoimmuneDisease. Comma-separated table containing Table S3. (CSV 20 kb)

Additional file 4: TableS4_Cancer. Comma-separated table containing Table S4. (CSV 67 kb)

Additional file 5: Suppl-Material. Microsoft Word file containing Figures S1 and S2 and Tables S5-S7. (DOCX 191 kb)

Abbreviations

CR: Cross-reactivity; CT: Cancer-testis; HLA: Human leucocyte antigen; IEDB: Immune epitope database; MHC: Major histocompatibility complex; NE: Natural epitope; TAP: Transporter associated with antigen processing; TCR: T cell receptor

Acknowledgements

None.

Funding

None.

Availability of data and materials

The datasets generated and/or analysed during the current study are available at <http://webclu.bio.wzw.tum.de/expitope2/SupplMaterialData.tgz>.

Authors' contributions

VJ further developed the algorithm and the web implementation. AM added the tissue scoring function. VJ and AM analyzed the data and wrote the manuscript. SR, DJS and DF conducted the project and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

VJ, AM and SR are employees of Medigene Immunotherapies GmbH/Medigene AG. DJS is Managing Director of Medigene Immunotherapies GmbH and CEO/CSO of Medigene AG.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany. ²Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, 82152 Planegg, Germany. ³St Petersburg State Polytechnical University, 195251 St Petersburg, Russia.

Received: 7 June 2017 Accepted: 28 November 2017

Published online: 28 December 2017

References

- Ludewig B, (ed). *Adoptive Immunotherapy: Methods and Protocols*. Methods in molecular medicine, Vol. 109. Totowa: Humana Press; 2005.
- Weber JS, Yang JC, Atkins MB, Disis ML. Toxicities of Immunotherapy for the Practitioner. *J Clin Oncol*. 2015;33(18):2092–9.
- Kohm AP, Fuller KG, Miller SD. Mimicking the way to autoimmunity: an evolving theory of sequence and structural homology. *Trends Microbiol*. 2003;11(3):101–5.
- Maus MV, Fraietta JA, Levine BL, Kalos M, Zhao Y, June CH. Adoptive Immunotherapy for Cancer or Viruses. *Annu Rev Immunol*. 2014;32(1):189–225.
- Dhanik A, Kirshner JR, MacDonald D, Thurston G, Lin HC, Murphy AJ, Zhang W. In-silico discovery of cancer-specific peptide-HLA complexes for targeted therapy. *BMC Bioinformatics*. 2016;17:286.
- Haase K, Raffegerst S, Schendel DJ, Frishman D. Expitope: a web server for epitope expression. *Bioinformatics*. 2015;31(11):1854–6.
- Keşmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng*. 2002;15(4):287–96.
- Jaravine V, Raffegerst S, Schendel DJ, Frishman D. Assessment of cancer and virus antigens for cross-reactivity in human tissues. *Bioinformatics*. 2017;33(1):104–11.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470–6.
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. The evolutionary landscape of alternative splicing in vertebrate species. *Science*. 2012;338(6114):1587–93.
- Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci O, Diken M, Castle JC, Sahin U. HLA typing from RNA-Seq sequence reads. *Genome Med*. 2012;4(12):102.
- Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res*. 2015;43(D1):405–12.
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*. 2012;40(D1):130–5.
- Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med*. 1996;22(7):707–10.
- Linette GP, Stadtmauer EA, Maus MV, Rapoport AP, Levine BL, Emery L, Litzky L, Bagg A, Carreno BM, Cimino PJ, Binder-Scholl GK, Smethurst DP, Gerry AB, Pumphrey NJ, Bennett AD, Brewer JE, Dukes J, Harper J, Tayton-Martin HK, Jakobsen BK, Hassan NJ, Kalos M, June CH. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced t cells in myeloma and melanoma. *Blood*. 2013;122(6):863–71.
- Morgan RA, Chinnasamy N, Abate-Daga D, Gros A, Robbins PF, Zheng Z, Dudley ME, Feldman SA, Yang JC, Sherry RM, Phan GQ, Hughes MS, Kammula US, Miller AD, Hessman CJ, Stewart AA, Restifo NP, Quezado MM, Alimchandani M, Rosenberg AZ, Nath A, Wang T, Bielekova B, Wuest SC, Akula N, McMahon FJ, Wilde S, Moseetter B, Schendel DJ, Laurencot CM, Rosenberg SA. Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J Immunother*. 2013;36(2):133–51.
- Vigeneron N, Stroobant V, Van den Eynde BJ, van der Bruggen P. Database of T cell-defined human tumor antigens: the 2013 update. *Cancer Immun*. 2013;13:15.
- Larsen M, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M. An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur J Immunol*. 2005;35(8):2295–303.
- Moise L, Beseme S, Tassone R, Liu R, Kibria F, Terry F, Martin W, De Groot AS. T cell epitope redundancy: cross-conservation of the TCR face between pathogens and self and its implications for vaccines and autoimmunity. *Expert Rev Vaccines*. 2016;15(5):607–17.
- Kumar A, Delogu F. Dynamical footprint of cross-reactivity in a human autoimmune T-cell receptor. *Sci Rep*. 2017;7:42496.
- Vigeneron N. Human Tumor Antigens and Cancer Immunotherapy. *BioMed Res Int*. 2015;2015:1–17.
- Nalepa G, Rolfe M, Harper JW. Drug discovery in the ubiquitin–proteasome system. *Nat Rev Drug Discov*. 2006;5(7):596–613.
- Wang J, Maldonado MA. The ubiquitin–proteasome system and its role in inflammatory and autoimmune diseases. *Cell Mol Immunol*. 2006;3(4):255–61.
- Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*. 2015;15(18):3163–8.
- Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AM-P, Jupp S, Koskinen S, Mannion O, Huerta L, Megy K, Snow C, Williams E, Barzine M, Hastings E, Weisser H, Wright J, Jaiswal P, Huber W, Choudhary J, Parkinson HE, Brazma A. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res*. 2016;44(D1):746–52.
- Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szgyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419.

27. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabudde NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LDN, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TSK, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. *Nature*. 2014;509(7502):575–81.
28. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi L, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu Z, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen X, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struwing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–5.
29. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescato M, Alam I, Albanese D, Altschuler GM, Arakawa T, Archer JAC, Arner P, Babina M, Rennie S, et al. A promoter-level mammalian expression atlas. *Nature*. 2014;507(7493):462–70.
30. Prete M, Dammacco R, Fatone MC, Racanelli V. Autoimmune uveitis: clinical, pathogenetic, and therapeutic features. *Clin Exp Med*. 2016;16(2):125–36.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit





Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors

Anja Mösch^{1,2}, Silke Raffeggerst², Manon Weis², Dolores J. Schendel² and Dmitrij Frishman^{1*}

¹ Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany, ² Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, Planegg, Germany

OPEN ACCESS

Edited by:

Davide Chicco,
Peter Munk Cardiac Centre,
Canada

Reviewed by:

Dhruv Sethi,
Obsidian Therapeutics,
United States
Gustavo Fioravanti Vieira,
Universidade La Salle Canoas,
Brazil

*Correspondence:

Dmitrij Frishman
d.frishman@wzw.tum.de

Specialty section:

This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: 26 July 2019

Accepted: 21 October 2019

Published: 19 November 2019

Citation:

Mösch A, Raffeggerst S, Weis M,
Schendel DJ and Frishman D
(2019) Machine Learning for Cancer
Immunotherapies Based on Epitope
Recognition by T Cell Receptors.
Front. Genet. 10:1141.
doi: 10.3389/fgene.2019.01141

In the last years, immunotherapies have shown tremendous success as treatments for multiple types of cancer. However, there are still many obstacles to overcome in order to increase response rates and identify effective therapies for every individual patient. Since there are many possibilities to boost a patient's immune response against a tumor and not all can be covered, this review is focused on T cell receptor-mediated therapies. CD8⁺ T cells can detect and destroy malignant cells by binding to peptides presented on cell surfaces by MHC (major histocompatibility complex) class I molecules. CD4⁺ T cells can also mediate powerful immune responses but their peptide recognition by MHC class II molecules is more complex, which is why the attention has been focused on CD8⁺ T cells. Therapies based on the power of T cells can, on the one hand, enhance T cell recognition by introducing TCRs that preferentially direct T cells to tumor sites (so called TCR-T therapy) or through vaccination to induce T cells *in vivo*. On the other hand, T cell activity can be improved by immune checkpoint inhibition or other means that help create a microenvironment favorable for cytotoxic T cell activity. The manifold ways in which the immune system and cancer interact with each other require not only the use of large omics datasets from gene, to transcript, to protein, and to peptide but also make the application of machine learning methods inevitable. Currently, discovering and selecting suitable TCRs is a very costly and work intensive *in vitro* process. To facilitate this process and to additionally allow for highly personalized therapies that can simultaneously target multiple patient-specific antigens, especially neoepitopes, breakthrough computational methods for predicting antigen presentation and TCR binding are urgently required. Particularly, potential cross-reactivity is a major consideration since off-target toxicity can pose a major threat to patient safety. The current speed at which not only datasets grow and are made available to the public, but also at which new machine learning methods evolve, is assuring that computational approaches will be able to help to solve problems that immunotherapies are still facing.

Keywords: cancer immunotherapy, T cell receptor, neoepitope, neoantigen, cross-reactivity, MHC binding affinity prediction

INTRODUCTION

Immunotherapies have gained more and more importance over the last decades. Checkpoint inhibitors mainly targeting PD1/PDL1 and CTLA4 and personalized cancer vaccines (Gubin et al., 2014; Ott et al., 2017; Sahin et al., 2017) have been and still are heavily investigated in clinical trials. Both depend on patient individual tumor-specific mutations enabling the boost of a cancer-specific T cell-mediated immune response (Snyder et al., 2014; Rizvi et al., 2015; Łuksza et al., 2017). A more direct approach utilizes the adoptive transfer of a patient's autologous T cells, either genetically modified with a transgenic chimeric antigen receptor (CAR) or T cell receptor (TCR). For CAR-T cell as well as TCR-T cell therapy a defined target, the epitope, needs to be identified. CARs, carrying the functional antigen-binding domain of an antibody, recognize three-dimensional peptide structures on the surface of a cell (Sadelain et al., 2013). By contrast, TCRs recognize predominantly linear peptides presented by the major histocompatibility complex (MHC) called human leucocyte antigen (HLA) in humans. For MHC class I presentation and thus CD8⁺ T cell detection, these peptides come from proteins that are intracellularly processed by either the constitutive proteasome or the IFN γ induced immunoproteasome (Griffin et al., 1998; Neefjes et al., 2011). After cleavage, the peptides are transported to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) complex, where they are loaded onto MHC class I molecules. The peptide-MHCs (pMHCs) are shuttled to the cell surface where they can potentially be recognized by CD8⁺ cytotoxic T cells, either naturally carrying or engineered to bear a pMHC-specific TCR (see **Figure 1**). However, there are more than 16,000 different alleles for *HLA-A*, *-B*, and *-C* genes, which bind and present different epitopes (Robinson et al., 2015). Besides MHC class I mediated CD8⁺ cytotoxic T cell responses, MHC class II bound peptides can induce CD4⁺ T cell responses that are also reported to play an important role in tumor detection and elimination (Nielsen et al., 2010; Linnemann et al., 2014; Kreiter et al., 2015; Andreatta et al., 2017; Veatch et al., 2018).

A wide spectrum of bioinformatics tools exists for modeling all steps of the MHC class I antigen presentation pathway, including proteasomal cleavage, translocation of the peptides

to the ER by TAP, peptide binding to the MHC molecules, and TCR recognition. The overarching goal of these efforts is to enhance our understanding of how T cell epitopes are selected from a virtually unlimited number of short peptides that can be proteolytically generated from the human proteome. The origin of these T cell epitopes can be naturally occurring proteins or peptides derived from somatic mutations. For personalized cancer immunotherapy, these patient- and tumor-specific mutations are usually separately assessed for each patient by exome sequencing, mutation detection and peptide binding prediction (Robbins et al., 2013; Blankenstein et al., 2015; Schumacher and Schreiber, 2015). Predicting these so called neoepitopes or neoantigens is a prevailing challenge for computational methods for immunotherapy and essential for a high-throughput approach to narrow down mutations to be included in vaccines or to be evaluated *in vitro* for T cell recognition, since only very few mutations are truly immunogenic (Yadav et al., 2014; Strömen et al., 2016; Bjerregaard et al., 2017a).

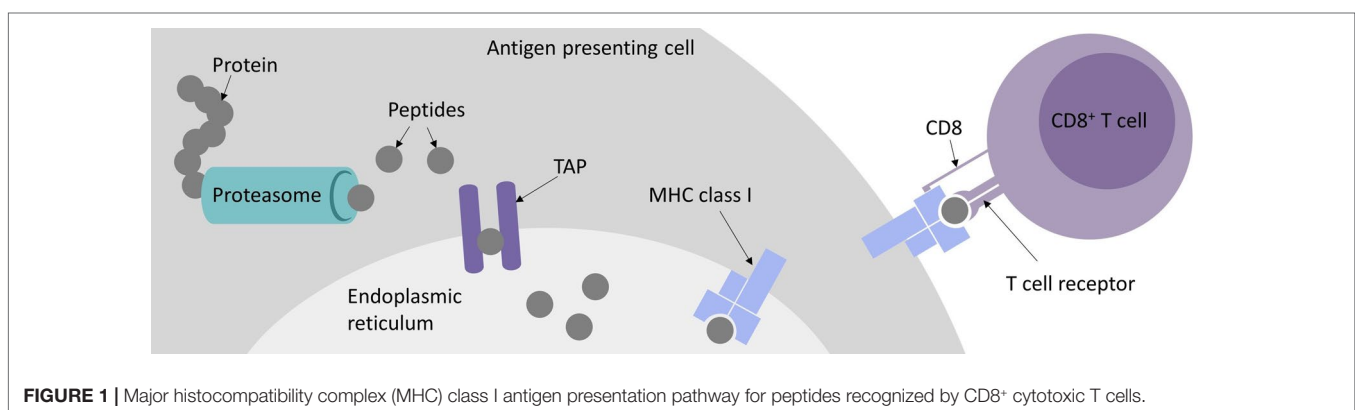
It is also of utmost importance to evaluate potential cross-reactivity of target-candidate epitopes based on various omics data such as proteomics and peptidomics (Haase et al., 2015; Jaravine et al., 2017a; 2017b). However, all existing approaches based on epitope presentation are only a surrogate for T cell recognition, for which no universal and computationally viable approach exists so far, although the first promising results have been published (Jurtz et al., 2018; Ogishi and Yotsuyanagi, 2019). By now, datasets have been generated that allow sequence-based prediction approaches using deep learning (Shugay et al., 2018; Vita et al., 2018).

In this review, we summarize the current state at the development of prediction algorithms and methods for all steps of antigen presentation, evaluate neoepitope prediction approaches, and discuss progress toward sequence-based TCR binding prediction.

PREDICTION OF T CELL EPITOPES

Proteasomal Cleavage Prediction

In order to develop an accurate prediction algorithm for proteasomal cleavages, a thorough mechanistic understanding of



the cutting process is required. The PProC algorithm by Kuttler et al. (Kuttler et al., 2000) relies on a biologically motivated model, which postulates that proteolytic sites are mostly determined by the local sequence context, generally not further away in the sequence than six amino acid residues. The two residues immediately adjacent to the cut make the greatest contribution to the affinity to the active subunits of the proteasome, while the influence of the other surrounding residues is lower. The recognition model is additive in that the total affinity, which ultimately determines the probability of the cut, is considered to be the sum of all individual contributions. Bioinformatics analyses revealed that the amino acids in the six positions preceding the cut and four positions downstream contain sufficient information to reproduce a training dataset of experimentally determined cleavage motifs of 20S proteasomes by a network-based technique. Keşmir et al. (Keşmir et al., 2002) demonstrated that good results in detecting proteasomal cleavage motifs can be achieved by combining experimental data on degradation by the constitutive proteasome with the sequences of peptides bound by the MHC class I molecules, which may be generated either by the constitutive or by the immunoproteasomes. A neural network trained on such a composite dataset, called NetChop, and an updated version NetChop 3.0 (Nielsen et al., 2005), achieved a reasonable accuracy and also yielded useful insights into cleavage-promoting and inhibiting residues as well as into N-terminal extension of peptides after proteasomal cleavage. A recurrent difficulty in predicting proteasomal cleavage is the lack of experimentally verified noncleavage sites. However, such negative data can be artificially generated by considering internal positions of confirmed MHC ligands or randomly generated sites.

TAP Binding Prediction

An early study of Daniel et al. (1998), in which the TAP binding affinity for a large number of peptides of length nine was measured by a peptide binding assay, revealed that positions one to three and nine of the 9-mers make the largest contribution to the selectivity of TAP to peptides. An artificial neural network trained on these data was able to predict the IC_{50} values with high accuracy. The study also found that HLA class I molecules differed significantly with respect to TAP affinities of their ligands. The predictive scope was later extended to peptides of arbitrary length using a stabilized matrix approach and a scoring scheme that only considers the first three N-terminal residues and the last C-terminal residue (Peters et al., 2003). Since it has been established that the selectivity of peptide transport by TAP is entirely determined by the peptide-binding step (Gubler et al., 1998), affinity predictions can be equated with translocation likelihood predictions. A number of further machine learning methods for predicting peptide binding to TAP were trained on 9-mer data, which is the typical length of the peptides that will subsequently bind to the MHC complex (Bhasin, 2004; Zhang et al., 2006; Diez-Rivero et al., 2010; Lam et al., 2010).

Peptide-MHC Binding Prediction

Sequencing of peptides eluted from MHC class I molecules (Falk et al., 1991) as well as mass-spectrometric (MS) (Hunt

et al., 1992) and crystallographic (Madden, 1995) evidence revealed common properties of the epitopes, in particular the typical length range of 8–12 residues. Additionally, it showed the existence of MHC allele-specific anchor residues, usually in positions two and nine of the core nonameric segments, as well as auxiliary anchors, where amino acid preferences are less strict (Rammensee et al., 1993).

Starting from the early nineties, efforts were made to collect available information on MHC class I ligands (Brusic et al., 1994; Rammensee et al., 1995; Rammensee et al., 1999) and to predict them using simple motif- and profile-based techniques (Rothbard and Taylor, 1988; Parker et al., 1994; Reche et al., 2002), based on the notion that peptides highly similar in sequence to experimentally characterized ligands will have a higher binding potential than more distantly related peptides and that individual amino acid side chains make independent contributions to the overall binding energy. Machine learning techniques, such as neural networks and hidden Markov models (Bisset and Fierz, 1993; Mamitsuka, 1998; Nielsen et al., 2003) outperform matrix-based methods in predicting peptide binding affinity (Peters et al., 2006; Lin et al., 2008). They are able to deal with peptides of variable length (Lundegaard et al., 2008) and to take into account nonadditive effects, which may arise, e.g., when two amino acids compete for the same site in the peptide-binding groove of the MHC heterodimer. The latest version of the widely used NetMHC algorithm 4.0 (Andreata and Nielsen, 2016) was trained on many thousands of quantitative affinity measurements for peptides of length 8–11 and the total of 118 MHC class I alleles from human, other primates, and mouse. Neural networks trained on all peptides (allmer networks) significantly outperformed the networks trained on peptides of each individual length separately. The study also suggested specific binding modes for 10- and 11-mers, which are predicted to bulge out of the MHC groove in contrast to 8- and 9-mers, which are strictly linear epitopes. MHCflurry, which relies on affinity measurement and peptide elution MS data, also uses neural networks trained individually for each HLA allele (O'Donnell et al., 2018b). Additionally, it allows users to train networks locally on data of their choice. This can be important especially for cancer immunotherapy applications, since peptide-binding affinity predictions are traditionally focused on viral epitopes.

There is also a growing group of pan-specific methods, including PickPocket (Zhang et al., 2009), NetMHCpan 4.0 (Jurtz et al., 2017), PSSMHCpan (Liu et al., 2017), and ACME (Hu et al., 2019), which take as input both the peptide and the HLA sequence and are able to predict the binding of any peptide to any allele. Most predictions are focused on MHC class I, but there are also methods available for MHC class II, such as NetMHCII 2.3 and NetMHCIpan 3.2 (Jensen et al., 2018), ProPred (Singh and Raghava, 2001), SMM-align (Nielsen et al., 2007), and NNAlign (Nielsen and Andreata, 2017), of which the latter also allows to train and use own models, as Garde et al. did for MHC class II prediction using both affinity measurement and MS data (Garde et al., 2019). Many of the aforementioned prediction methods for both MHC class I and II and consensus methods, such as NetMHCcons (Karosiene et al., 2012) and the consensus method by Moutaftsi et al. (Moutaftsi et al., 2006), are integrated into

the IEDB epitope analysis resource and can be accessed online (Wang et al., 2010; Fleri et al., 2017; Vita et al., 2018; Dhanda et al., 2019). In addition, combinatory pipelines and frameworks have been published, namely, EpiJen (Doytchinova et al., 2006), NetCTL (Larsen et al., 2007), NetCTLpan (Stranzl et al., 2010), and FRED2 (Schubert et al., 2016), modeling the complete antigen presentation pathway by including proteasomal cleavage and TAP transport predictions.

Epitope presentation, however, is only one step toward T cell recognition. NetMHCstab (Jørgensen et al., 2014) and NetMHCstabpan (Rasmussen et al., 2016) are methods to predict the stability of pMHC complexes, presuming that epitope presentation lasting longer increases the likelihood of T cell recognition and thus immunogenicity. Calis et al. proposed a scoring model to predict true immunogenicity of T cell epitopes (Calis et al., 2013). Despite these efforts, however, true immunogenicity remains far more difficult to predict than mere MHC-binding affinity.

Beyond sequence-based approaches, significant methodological progress has been made in modeling peptide binding to MHC class I molecules on structure level. The diversity of the cognate peptide repertoire and the experimental binding profiles for a particular MHC protein can be accurately captured using both general purpose modeling packages, such as Rosetta (Yanover and Bradley, 2011), and faster specialized methods, such as GradDock (Kyeong et al., 2018), DockTope (Menegatti Rigo et al., 2015), and LYRA (Klausen et al., 2015), of which the latter two are also integrated in the IEDB. Docking experiments are becoming increasingly successful in reproducing crystallographically known peptide-MHC binding geometry (Bordner and Abagyan, 2006; Antunes et al., 2018).

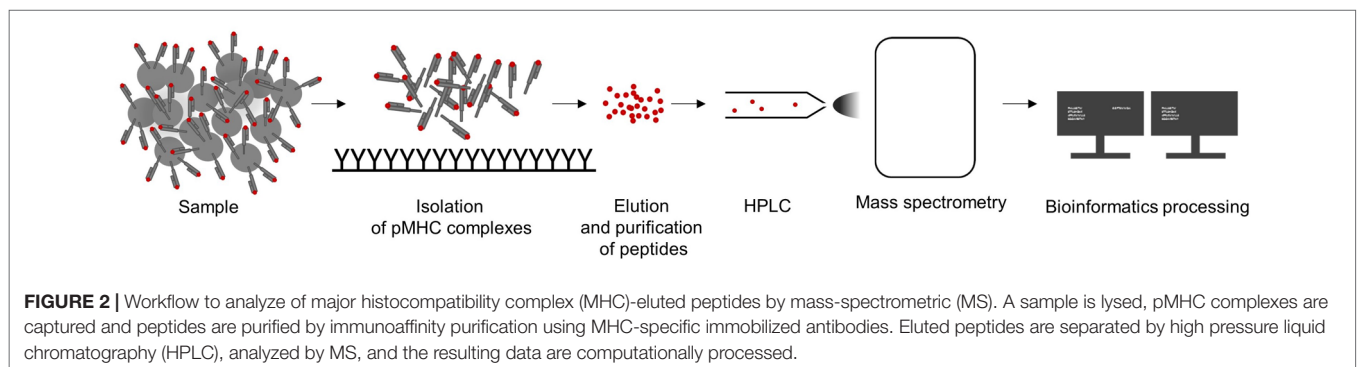
Immunopeptidomics Data

The recent availability of large-scale immunopeptidomics data allowed to explicitly model peptide length distributions and the interdependence between individual sequence positions, leading to more accurate predictions of naturally presented MHC class I ligands (Gfeller et al., 2018). MS profiling provides novel insights into the antigen processing rules, including the discovery of binding motifs, improved description of proteasomal cleavage signatures, cellular localization and sequence features of peptide source proteins, and better understanding of the role of gene

expression, protein abundance and degradation (Bassani-Sternberg et al., 2015; Bassani-Sternberg et al., 2017; Abelin et al., 2017). In particular, Abelin et al. (2017) reported that neural networks trained on MS-derived peptides bound to 16 different HLA alleles outperformed affinity-trained predictors.

For immunogenicity, T cell epitope verification by TCRs or TCR-like antibodies would constitute an ideal dataset to train prediction algorithms (Dolan, 2019), but both approaches are highly dependent on specificity and affinity of TCRs and antibodies used and do not reach the high-throughput efficiency of immunopeptidomics. HLA-peptidomics, which is the MS analysis of MHC-eluted peptides, is the most sophisticated method for high-throughput qualitative and quantitative detection of MHC ligands and thereby of potential T cell epitopes (Hunt et al., 1992; Caron et al., 2011; Hassan et al., 2014; Álvaro-Benito et al., 2018; Freudenmann et al., 2018).

The isolation of pMHC complexes from cell surfaces (Sugawara et al., 1987; Storkus et al., 1993; Bassani-Sternberg et al., 2015; Marino et al., 2019) or out of serum (Ritz et al., 2016, 2017) is the first critical step for a high-quality MS HLA-peptidome analysis. After elution from pMHC complexes, peptides are purified, separated by high pressure liquid chromatography (HPLC), and directly injected and analyzed in a mass spectrometer followed by computational processing of MS spectra data (see **Figure 2**). Successful peptide detection is determined by various factors, such as HLA enrichment, which is dependent on HLA-antibody quality, efficient elution, and physicochemical characteristics of a peptide defined by its amino acid composition. Relevant peptide properties can be mass, hydrophilicity, and hydrophobicity, its ability to be ionized, as well as cysteine content (Gfeller and Bassani-Sternberg, 2018). Therefore, not all peptides are equally likely to be detected by MS but it is difficult to assess how many peptides are missed. Peptide sequences are often determined by tandem MS: a precursor mass spectrum called MS1 spectrum of the eluted peptides is generated and only peptides with high intensities are isolated for fragmentation and analyzed, resulting in a MS2 or MS/MS spectrum. Observed mass spectra are then compared with theoretical mass spectra in general reference databases. Proteogenomic computational pipelines using customized reference datasets also allow the identification of peptides originating from noncanonical and allegedly noncoding reading frames (Laumont and Perreault, 2017; Laumont et al., 2018), unconventional, genomic coding-sequences (Erhard et al.,



2018) as well as neoepitopes from somatic alterations (Yadav et al., 2014; Carreno et al., 2015) or intron retentions (Smart et al., 2018). In addition, the generation of customized spectral library databases of high confidence peptides can be used for data-independent acquisition approaches (Ritz et al., 2017), resulting in increased reproducibility and sensitivity.

Peptides are often assigned to the HLA molecule from which they were originally eluted by predicting the binding affinity (Freudenmann et al., 2018; Bilich et al., 2019). For common HLA alleles, usually a sufficient number of peptides are identified as binders, resulting in datasets large enough to train prediction algorithms. However, for less frequent HLA alleles, the pool of identified and correctly assigned peptides is more limited, which leads to variability in performance of prediction techniques depending on the rarity of each HLA allele (O'Donnell et al., 2018b). If MS datasets annotated by binding affinity predictions are used to train machine learning algorithms, a self-amplifying bias is introduced. MS profiling of mono-allelic cells (Giam et al., 2015; Abelin et al., 2017) as well as deconvolution approaches (Bassani-Sternberg and Gfeller, 2016) can circumvent this problem and improve the quality of available training data and prediction performance.

IMMUNOTHERAPY-SPECIFIC APPLICATIONS OF EPITOPE PREDICTION

Neoepitope Identification

Cancer-specific mutations have been demonstrated to be viable targets for tumor-infiltrating lymphocytes (TILs) enabled by checkpoint inhibitors that block CTLA4 or PD1/PDL1 or by vaccine-induced immune responses (van Rooij et al., 2013; Carreno et al., 2015; Cohen et al., 2015; Gros et al., 2016; McGranahan et al., 2016; Ott et al., 2017; Zacharakis et al., 2018; Hilf et al., 2019). These mutations alter amino acid sequences of proteins and are recognized as so called neoepitopes or neoantigens, with both terms used ambiguously and oftentimes synonymously in the literature. Here, we use the term neoepitopes for epitopes predicted to be presented by a certain MHC and the term neoantigens for confirmed immunogenic mutations. By definition, neoantigens are tumor-specific, which makes them ideal immunotherapy targets, but they are also to a large degree patient-specific. Despite many efforts, only very few shared neoantigens such as KRAS^{G12D/V} or BRAF^{V600E}, could be identified, making an off-the-shelf therapy approach hardly feasible (Warren and Holt, 2010; Angelova et al., 2015; Tran et al., 2015; Thorsson et al., 2018). Furthermore, a high individual tumor mutation burden and the ambition to provide personalized medicine for more patients do not allow for testing the immunogenicity of every mutation *in vitro*. Therefore, the current standard procedure for individual patients relies on exome sequencing followed by mutation calling and machine learning-based neoepitope prediction, which represents the main application of pMHC-binding prediction algorithms in the field of cancer immunotherapy. Here, we reviewed more than 70 publications using binding prediction algorithms to identify neoepitopes of which 49, that provided quantifiable data, are shown in **Table 1**.

Not all studies stated all steps of their neoepitope selection process, including which algorithm parameters were used, how many neoepitopes were found when applying a threshold or how many and what types of mutation were used for predicting neoepitopes, which makes quantitative evaluation and reproducibility difficult. This is aggravated by the large variance in ratio of predicted neoepitopes per mutation, which is caused by thresholds of varying strictness, the number of features used for filtering, and the approach to counting neoepitopes or neoantigens, i.e., if a mutation was counted only once even if presented by more than one HLA allele or contained in multiple epitopes predicted to be immunogenic. Furthermore, some studies could only experimentally validate a subset of predicted neoepitopes and experimental validation was determined by biological assays of varying sensitivity from MHC-ligand confirmation to ELISPOT assays using patient-specific TILs.

Not surprisingly, most publications investigated cancer types known for high mutation loads, such as non-small cell lung carcinoma and melanoma, but glioblastoma and chronic lymphocytic leukemia were also shown to harbor neoantigens identified by neoepitope prediction (Rajasagi et al., 2014; Hilf et al., 2019; Keskin et al., 2019). Regarding mutation types, the focus clearly lies on single nucleotide variants (SNVs) considering their abundance in tumors above all other types of mutation, their comparatively easy detection by mutation calling software and easier computational generation of mutated and wild-type peptide sequences (Bailey et al., 2018; Ellrott et al., 2018). However, larger indels, frameshifts, and other more complex mutation types can be the source of more neoepitopes that are also less similar to self and thus highly interesting immunotherapeutic targets. More recent studies from Kahles et al., Koster et al., and Schischlik et al. investigated these types of mutation, benefitting from improvements on sequencing and mutation calling techniques (Kahles et al., 2018; Koster and Plasterk, 2019; Schischlik et al., 2019). Nevertheless, identification of cancer-specific mutation remains a critical step in every neoepitope identification pipeline and the number of mutations obtained varies greatly dependent on the software and thresholds employed (Tran et al., 2015; Karasaki et al., 2017).

The focus of most publications lies on MHC class I presented neoepitopes that can be detected by CD8⁺ T cells. MHC class I prediction algorithms are more commonly used but there is clear evidence that MHC class II mediated CD4⁺ T cell responses play a major role in neoantigen immune responses and thus should also be considered for neoepitope detection. (Linnemann et al., 2014; Kreiter et al., 2015; Tran et al., 2015; Hugo et al., 2016; Ott et al., 2017; Reuben et al., 2017; Sahin et al., 2017; Sonntag et al., 2018; Vrecko et al., 2018).

All studies, except Koster et al., who investigated 10-mers only, looked at peptides with a length of 8–10 or 8–11 amino acids or just at 9-mers alone, which are the majority of peptides presented by MHC class I (Trolle et al., 2016). Most studies also relied on matching HLA types for the samples used, often determined by one of the following HLA typing algorithms: ATHLATES, HLAMiner, OptiType, PHLAT, POLYSOLVER, and seq2HLA (Boegel et al., 2012; Warren et al., 2012; Liu et al., 2013; Szolek et al., 2014; Shukla et al., 2015; Bai et al., 2018). In contrast,

TABLE 1 | Publications describing the application of machine learning approaches to neoepitope prediction.

Publication	Indication	Sample type and number	number of HLAs used	Estimated ratio of predicted neoepitopes from mutations	Estimated ratio of experimentally confirmed neoantigens	Number of features	Algorithms
(Segal et al., 2008)	BRCA/CRC	11 patients	1	0.17	N/A	1	NetMHC, SYFPEITHI, BIMAS, RANKPEP
(Castle et al., 2012)	MEL	1 murine cell line	N/S	0.05	0.32 ^T	2	NetMHC
(Khalili et al., 2012)	various	312 genes (COSMIC)	57	1.40	N/A	2	NetMHC 3.2
(Robbins et al., 2013)	MEL	3 patients	2	0.18	0.03 ^T	3	NetMHCpan 2.4
(van Rooij et al., 2013)	MEL	1 patient	4	0.42	<0.01 ^T	3	NetChop, NetMHC 3.2
(Boegel et al., 2014)	various	167 cancer cell lines	6	0.44	N/A	1	IEDB 2.9
(Duan et al., 2014)	SARC	2 murine tumors	3	0.75	0.56 ^T	2	NetMHC 3.0
(Snyder et al., 2014)	MEL	64 patients	6	0.42	<0.01 ^T	3	NetMHC 3.4, RANKPEP, IEDB immunogenicity, CTLPred
(Yadav et al., 2014)	CRC/PRAD	2 murine cell lines	2	0.03	0.02 ^T	3	NetMHC 3.4
(Angelova et al., 2015)	CRC	552 TCGA patients	6	0.41	N/A	2	NetMHCpan
(Carreno et al., 2015)	MEL	7 samples/3 patients	1	0.04	0.43 ^B	3	NetMHC 3.4
(Cohen et al., 2015)	MEL	8 patients	2	0.02	0.02 ^T	2	IEDB
(Rizvi et al., 2015)	NSCLC	34 patients	6	0.62	<0.01 ^T	2	NetMHC 3.4
(Rooney et al., 2015)	various	4250 TCGA patients	6	0.14	N/A	2	NetMHCpan 2.4
(Tran et al., 2015)	GIC	10 patients	12	0.03	0.21 ^T	2	NetMHCpan 2.8, NetMHCIIpan 3.0
(Van Allen et al., 2015)	MEL	110 patients	6	1.56	N/A	2	NetMHCpan 2.4
(van Gool et al., 2015)	UCEC	245 TCGA patients	1	0.06	N/A	3	NetMHCpan 2.8
(Bassani-Sternberg and Gfeller, 2016)	MEL	1 patient	6	1.43	<0.01 ^B	1	NetMHCpan 2.8
(Goh et al., 2016)	MCC	49 patients	4	0.09	N/A	1	NetMHC 3.4
(Gros et al., 2016)	MEL	3 patients	6	0.03	0.55 ^T	2	IEDB
(Hugo et al., 2016)	MEL	38 patients	12	0.06	N/A	3	NetMHCpan 2.8, NetMHCIIpan 3.0
(Kalaora et al., 2016)	MEL	1 patient	6	5.30	<0.01 ^B	1	NetMHCpan 2.8
(Karasaki et al., 2016)	NSCLC	15 patients	6	0.62	N/A	1	NetMHCpan 2.8
(Löffler et al., 2016)	CHOL	1 patient	6	3.68	0 ^B	2	NetMHC 3.4, NetMHCpan 2.8, SYFPEITHI
(Strønen et al., 2016)	MEL	3 patients	1	0.05	0.19 ^T	4	NetChop, NetMHC 3.2, NetMHCpan 2.0
(Anagnostou et al., 2017)	NSCLC	10 patients	6	0.76	<0.01 ^T	4	SYFPEITHI, NetMHCpan, NetCTLpan
(Chang et al., 2017)	PED	540 patients	6	0.42	N/A	2	NetMHCcons 1.1
(Karasaki et al., 2017)	NSCLC	4 patients	6	0.20	N/A	2	NetMHCpan 2.8
(Kato et al., 2017)	BRCA	5 patients	6	0.47	N/A	2	NetMHC 3.4, NetMHCpan 2.8
(Miller et al., 2017)	MM	664 patients	6	0.16	N/A	3	NetMHC 4.0
(Ott et al., 2017)	MEL	6 patients	6	0.01	0.60 ^T	3	NetMHCpan 2.4
(Sahin et al., 2017)	MEL	13 patients	10	0.02	0.60 ^T	2	IEDB 2.5 (MHC class I & II)
(Zhang et al., 2017)	BRCA	3 patients	6	0.01	0.16 ^T	3	NetMHC 3.2
(Kalaora et al., 2018)	MEL	15 patients/cell lines	6	9.57	0.15 ^T	2	NetMHCpan 3.0
(Kinhead et al., 2018)	PAAD	1 murine cell line	2	0.27	0.16 ^T	2	NetMHC 3.2/3.4, NetMHCpan 2.8
(Martin et al., 2018)	OV	1 patient	6	1.57	0.09 ^T	2	NetMHCpan 2.4
(O'Donnell et al., 2018a)	OV	92 patients	6	0.02	N/A	2	NetMHCpan 2.8
(Sonntag et al., 2018)	PDAC	1 patient	10	2.00	0.75 ^T	3	NetMHC, NetMHCIIpan 3.1, SYFPEITHI

(Continued)

TABLE 1 | Continued

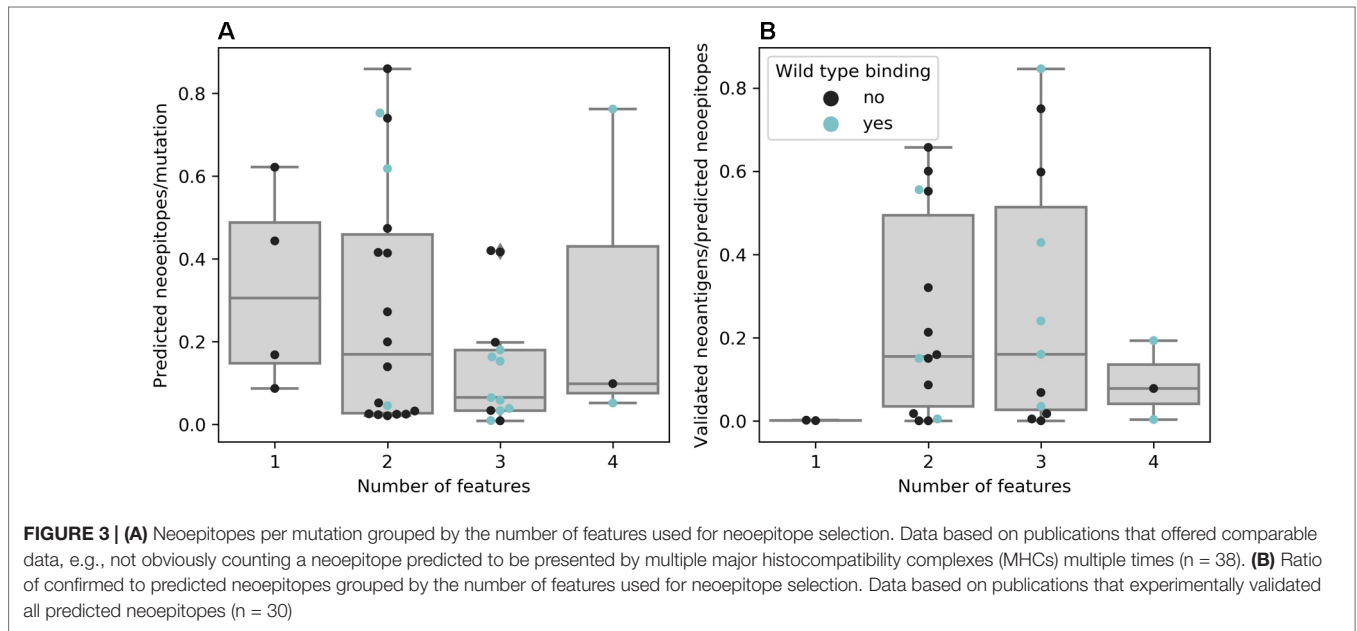
Publication	Indication	Sample type and number	number of HLAs used	Estimated ratio of predicted neopeptides from mutations	Estimated ratio of experimentally confirmed neoantigens	Number of features	Algorithms
(Thorsson et al., 2018)	various	8546 TCGA patients	6	0.74	N/A	2	NetMHCpan 3.0, pVAC-Seq 4.0.8
(Vreko et al., 2018)	HCC	1 patient	3	0.05	0.15 [†]	2	SYFPEITHI, IEDB (MHC class II)
(Wu et al., 2018)	various	7748 TCGA samples	100	1.18	N/A	1	NetMHCpan 4.0
(Bulik-Sullivan et al., 2019)	NSCLC	7 patients	6	0.10	0.08 [†]	>4	EDGE
(Hilf et al., 2019)	GBM	10 patients	1	0.03	0.85 [†]	3	IEDB 2.5
(Keskin et al., 2019)	GBM	8 patients	6	0.20	0.07 [†]	3	NetMHCpan 2.4
(Koster and Plasterk, 2019)	various	10186 TCGA patients	1	0.02	N/A	2	NetMHC 4.0
(Liu et al., 2019)	OV	20 patients	12	0.15	0.24 [†]	3	NetMHCpan 3.0, NetMHCIIpan 3.1
(Löffler et al., 2019)	HCC	16 patients	6	1.79	0 [‡]	2	NetMHC 4.0, NetMHCpan 3.0, SYFPEITHI
(Rosenthal et al., 2019)	NSCLC	164 samples/64 patients	6	0.86	N/A	2	NetMHC 4.0, NetMHCpan 2.8
(Schischlik et al., 2019)	PNMN	113 patients	6	2.53	0.66 [‡]	2	NetMHCpan

N/S means not specified. Cancer type abbreviations: adenocarcinoma (AC), breast cancer (BRCA), cholangiocarcinoma (CHOL), colorectal cancer (CRC), glioblastoma (GBM), gastrointestinal cancer (GIC), hepatocellular carcinoma (HCC), merkel cell carcinoma (MCC), melanoma (MEL), multiple myeloma (MM), non-small cell lung cancer (NSCLC), ovarian cancer (OV), pancreatic ductal adenocarcinoma (PDAC), pediatric cancers (PED), Ph-negative myeloproliferative neoplasms (PNMN), prostate adenocarcinoma (PRAD), sarcoma (SARC) and uterine corpus endometrial cancer (UCEC). [†] indicates experimentally confirmed T cell responses (e.g., IFN γ ELISPOT), [‡] indicates experimentally confirmed major histocompatibility complex (MHC) binding (e.g., mass spectrometric [MS] of eluted peptides), and N/A indicates that no experimental validation was done. Features are mutated peptide binding prediction, wild-type peptide binding prediction, gene expression, sequence-based features like sequence similarity scores, and immunogenicity predictions. If available, version information of algorithms is included.

Wu et al. made predictions based on the 100 most frequent HLA alleles in their dataset and Wood et al. based on the general 145 most frequent alleles (Wood et al., 2018; Wu et al., 2018). Whether or not such approaches yield substantial information gain is a debatable issue since most immunogenic mutations are highly individual and restricted by a patient's individual HLA type (Marty et al., 2017; McGranahan et al., 2017; Rosenthal et al., 2019). HLA-A*02:01 has been extensively studied since it is the most common allele in Caucasian populations and therefore was exclusively used by Segal et al. for their analysis (Segal et al., 2008). Since predictions for A*02:01 still belong to the best performing group and can be more easily validated compared to other alleles due to established *in vitro* protocols and reagents, Carreno et al., Spranger et al., Strønen et al., van Gool et al., and Hilf et al. also only used A*02:01 for their predictions and the studies that carried out experimental validation accomplished high confirmation of predicted neopeptides (Carreno et al., 2015; van Gool et al., 2015; Spranger et al., 2016; Strønen et al., 2016; Hilf et al., 2019). Similarly, Koster et al. only used A*02:01 for an unfiltered TCGA dataset although they did not perform experimental validation. Similar to Wood et al., they did not use HLA typing information for TCGA samples, which has been generated but can only be obtained by applying for access to restricted data (Shukla et al., 2015; Charoentong et al., 2017; Marty et al., 2017).

For most studies, algorithms from the NetMHC family were chosen as they are widely known and represent the

state-of-the-art prediction methods for binding of a peptide to a given MHC molecule. Van Allen et al. showed that out of 17 validated neoantigens, 14 passed the 500 nM standard threshold, indicating high sensitivity (van Buuren et al., 2014). However, only a handful of the predicted binders will also be recognized by T cells, which requires additional filtering or prediction improvement (Anonymous, 2017). Indeed, using more filtering criteria leads to fewer predicted neopeptides per mutation, as seen in **Figure 3A**, although the false negative rate remains unknown. Only a few publications rely on predicting the binding affinity of mutated peptides alone and most use at least one additional threshold criterion, of which gene expression as a premise for antigen recognition is the most common. As RNA-Seq data was not available for Anagnostou et al., Le et al. and Reuben et al., they used TCGA expression data as a proxy to further filter the mutations to test for immunogenicity. Binding of the wild-type peptide was also considered by some studies, but not always used for filtering. Duan et al. proposed a “differential agretopicity index” (DAI), which is the difference between the predicted mutated and wild-type binding affinity, to use as a filtering criterion for neopeptide prediction. Although it yielded promising results based on their mouse data, it seemed less reliable in further investigations by Bjerregaard et al. and Koşaloğlu-Yalçın et al. using human data (Duan et al., 2014; Bjerregaard et al., 2017b; Koşaloğlu-Yalçın et al., 2018). In another study by Ghorani et al., DAI was more predictive for



immune infiltration in melanoma and lung cancer compared to neoantigen or mutation load, suggesting that while some neoepitope responses might be enhanced by a reduced cross-reactivity potential, there are also many validated neoantigens whose wild-type counterparts are predicted to bind comparably strong (Ghorani et al., 2018; Koşaloğlu-Yalçın et al., 2018).

There is evidence that taking more than one feature into account promises greater success for experimentally validating predicted neoepitopes (see **Figure 3B**). However, the results of experimental validation are dependent on the sensitivity of the technique used and the reactivity of neoantigen-specific TILs can additionally be hampered by other factors, such as tumor immune suppression or T cell exhaustion (Anonymous, 2017; Bulik-Sullivan et al., 2019).

Some studies chose a quantitative approach, mostly linking neoepitope load and survival (Brown et al., 2014; Rizvi et al., 2015; Miller et al., 2017; Ghorani et al., 2018). It has to be mentioned that neoepitope load and mutational burden are usually highly

correlated (Pearson $r = 0.89$ based on 38 publications with less than 1 neoepitope per mutation from **Table 1**) and although it can be assumed that an increased survival is linked to the immunogenicity of mutations, quantifying predicted neoepitopes does not necessarily transport more information than mutation burden alone (Nathanson et al., 2017). There are, however, also studies that correlated survival with neoepitopes but not mutational burden or found contradictory results depending on patient cohorts (Snyder et al., 2014; Ghorani et al., 2018).

Among well-described approaches for neoepitope identification based on affinity binding prediction algorithms, there are also pipelines available that automate all analytic steps and rank potential neoepitopes based on peptide affinity prediction and other features (see **Table 2**). They differ greatly as to their properties and outputs, thus offering choices depending on research questions and dataset sizes. Their availability demonstrates how important neoepitope prediction has become as an application for binding affinity prediction algorithms.

TABLE 2 | Neoepitope prediction pipelines based on mutation data input. Additional features are cancer driver status of the mutated gene used by MuPeXI; differential agretopicity index (DAI), sequence-based immunogenicity score, and more used by Neopepsee; DAI, cleavage, and stability prediction used by pVACtools.

	MuPeXI	CloudNeo	Neopepsee	pVACTools
Algorithms	NetMHCpan	NetMHCpan	NetCTLpan, IEDB Bayes classifier	8 MHC class I predictors 4 MHC class II predictors
Input	VCF gene expression TSV	VCF BAM	VCF RNA-Seq FASTQ	VCF BAM (RNA and DNA)
HLA typing	user input	integrated	user input or integrated	user input or integrated
Mutation types	SNVs indels frameshifts	SNVs	SNVs	SNVs indels fusions (additional input)
Wild type peptide	yes	yes	yes	yes
Gene expression	yes (optional)	no	yes	yes
Additional features	yes	no	yes	yes
Availability	local, webservice	cloud	local	local
Reference	(Bjerregaard et al., 2017a)	(Bais et al., 2017)	(Kim et al., 2018)	(Hundal et al., 2019)

Since a variety of different neoepitope identification approaches exist and it is not clear which features are predictive for immunogenicity, Koşaloğlu-Yalçın et al. and Kim et al. integrated and compared features additional to the standard MHC binding affinity by either comparing areas under the curve of receiver operating characteristics or evaluating feature importance derived from trained classifiers (Kim et al., 2018; Koşaloğlu-Yalçın et al., 2018). Both studies found that binding affinity prediction performs best or is the most informative feature. This is not surprising for viral epitopes constituting a major part of data on which most prediction algorithms are trained nor for neoantigens from literature mainly selected by predicted binding affinity, which introduces a bias toward this feature. It still remains unclear how many potential neoantigens are not detected because their binding affinity is predicted to lie beyond thresholds. An approach avoiding this bias has been proposed by Bulik-Sullivan et al. (Bulik-Sullivan et al., 2019). Like the most recent generation of neural network binding prediction algorithms, they developed a deep learning neural network trained on MS data, but apart from improved peptide sequence modeling, they also included features unrelated to the pMHC interaction, namely, quantified gene expression, flanking sequence, and protein family. Although their model is currently limited to HLA alleles of the training data, the approach demonstrated an increased performance of neoepitope discovery over peptide binding prediction and can also be expanded to MHC class II presented antigens.

Cross-Reactivity Assessment

A major challenge for immunotherapies introducing TCRs into patient recipient T cells is the choice of safe target antigens. If an engineered TCR-T cell cross-reacts with self-antigens in healthy tissue, the side-effects can be devastating. Possible TCR toxicity scenarios can be generally divided into on-target and off-target toxicities. On-target toxicities include all aspects of a specific target antigen or epitope expression that lead to an unintentional TCR-mediated destruction of healthy tissues. An example of on-target toxicity is melanocyte destruction, hearing loss, and retina infiltration mediated by MART1-targeting TCR-T cells relating to the same epitope in all cases (Johnson et al., 2009).

Off-target toxicities, in contrast, can appear by unexpected recognition of alternative epitopes that contain amino acid exchanges (mismatches) compared to the known epitope sequence. In rare cases, these mismatched peptides are presented identically on corresponding MHC molecules and are recognized equally well by deployed TCRs.

Targeting epitope sequences of proteins originating from highly homologous family members can cause unforeseen tissue damage as exemplified by the study performed by Morgan et al. (Morgan et al., 2013). Using autologous anti-MAGEA3 TCR-T cells, adoptive transfer led to severe neurotoxicity in several patients. The MAGEA3-specific TCR used in this clinical trial also recognized a MAGEA12, which was retrospectively found to be expressed in the brain. In the Linette et al. study, clinicians adoptively transferred MAGEA3-TCR-modified lymphocytes that also recognized an alternative epitope derived from the

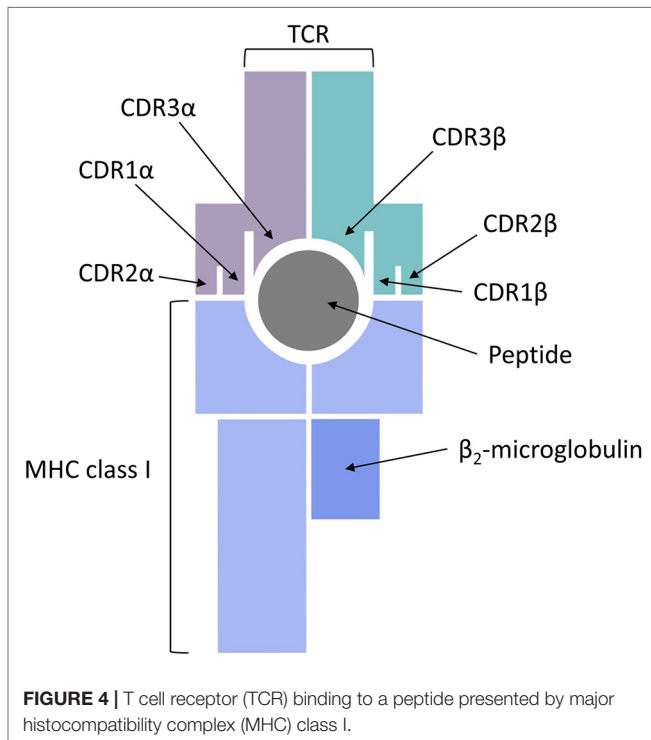
protein titin, causing fatal heart failure in two patients (Linette et al., 2013). Each of these examples underline the importance and need of comprehensive preclinical target and TCR analysis to prevent potential adverse events at later stages of clinical development.

With Expitope, we presented the first web server for assessing epitope sharing when evaluating new potential target candidates (Haase et al., 2015). Based on predictions for proteasomal cleavage, TAP transport, and MHC class I binding affinity, Expitope lists peptides with a given number of mismatches including the original target peptide. For these peptides, which are linked to genes by transcripts, the expression values in various healthy tissues, representing all vital human organs, are extracted from RNA-Seq data. However, transcript abundance only indirectly indicates protein expression. Meanwhile, proteome-wide human protein abundance data has become available and now facilitates a more direct approach for the prediction of potential cross-reactivity. The development of a new version 2.0 of Expitope, which computes all possible, naturally occurring epitopes of a peptide sequence and the corresponding cross-reactivity indices using both protein and transcript abundance levels weighted by a proposed hierarchy of importance of various human tissues, should help addressing this issue (Jaravine et al., 2017a). Cross-reactivity potential can also be assessed by calculating structural similarities between pMHC complexes obtained by molecular docking (Antunes et al., 2010) and by clustering pMHC complexes based on their electrostatic properties and the accessible surface area (Mendes et al., 2015). A comprehensive review by Baker et al. (2012) is covering these aspects in great detail.

TCR BINDING PREDICTION

The final piece of the epitope recognition puzzle is the interaction of the pMHC complex with the TCR, which represents a very difficult problem for modeling studies and sequence-based predictions. One reason for that is the complex and noncontiguous nature of the interaction interface, with the CDR1 and CDR2 regions of the TCR α and β chains making contacts with the MHC class I molecule and the CDR3 regions directly interacting with the bound peptide (see **Figure 4**). Another major hurdle in predicting TCR recognition is the scarcity of experimentally confirmed TCR complementarity determining regions and the sequences of their respective binding partners on the pMHC complex. For example, one of the first feasibility studies of CDR3 sequence patterns was only based on two immunogenic HIV peptides (De Neuter et al., 2018). An additional complication is posed by the fact that repertoire sequencing combined with immune assays determines antigen-specific clonotypes, but does not yield negative controls, i.e., validated pairs of CDRs and pMHC complexes that do not bind each other.

CDR3 β chains appear to always be in contact with the antigen bound to the MHC class I molecule, whereas the direct contact of CDR3 α chains to the peptide is not always required (Glanville et al., 2017). The involvement of short linear stretches of CDR3 β sequence in peptide-TCR interactions creates the opportunity to cluster TCRs in groups of common specificity

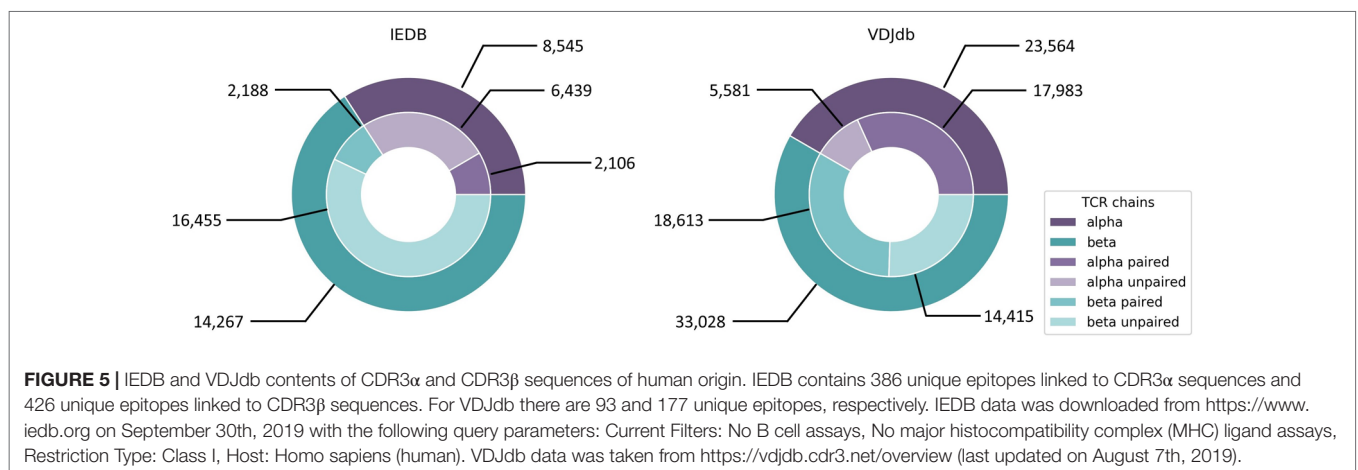


(Dash et al., 2017; Glanville et al., 2017) and also serves as the basis for developing specialized algorithms for sequence-based prediction of pMHC/TCR binding. Two recent publications addressed this problem from two completely different perspectives. Jurtz et al. presented a proof of concept study, in which they predicted TCR interactions with their cognate HLA-A*02:01-presented peptide targets (Jurtz et al., 2018). A machine learning approach, called NetTCR, was trained on 8,920 TCRβ CDR3 sequences and 91 cognate peptide targets obtained from IEDB and from the immune assay data published by Klinger et al. (2015). A dataset of negative interactions was

assembled by randomly matching TCR and peptide pairs. The NetTCR project in its current form is limited to a small number of peptides and it does not consider CDR1/CDR2 interactions with the MHC molecules or CDR3α sequences, but it is an important step forward because it demonstrates that TCR recognition of pMHCs is specific enough to be captured by sequence-level prediction tools.

Ogishi and Yotsuyanagi exploited the existence of immunodominant epitopes, which are targeted by the adaptive immune system in different individuals and would therefore be expected to exhibit some prominent features that make them especially prone to be recognized by T cells (Ogishi and Yotsuyanagi, 2019). The idea behind their repertoire-wide TCR-epitope contact potential profiling is that intermolecular contacts between relevant portions of the epitope and the TCR CDR3β region that closely resemble the contact structure of the interactions involving immunodominant peptides would be more likely to be immunogenic. To quantitatively assess the interaction affinity, they used physicochemical properties of amino acids and an energetic potential, calculated as the sum of all pairwise contact potentials for individual amino acids. The latter were obtained from several previously published amino acid contact potential scales, available from the AAINDEX database (Kawashima et al., 2007). These features were converted to immunogenicity scores using machine learning. It should be noted that the knowledge-based potentials, derived from crystal structures of proteins and protein complexes, reflect either intramolecular interactions driving protein folding and stability or contacts at protein interfaces and may only be a coarse approximation of peptide-TCR interactions. Yet, Ogishi and Yotsuyanagi demonstrated that the most informative contact-based and property-based features strongly correlate with experimentally measured TCR-peptide affinities.

Both approaches by Jurtz et al. and Ogishi and Yotsuyanagi are solely based on CDR3β chains and do not incorporate CDR3α sequence information. This is due to the fact that most datasets and databases such as IEDB and VDJdb did, until recently, consist mainly of CDR3β sequences (Figure 5)



derived from bulk sequencing (Shugay et al., 2018; Vita et al., 2018), since identifying functional TCR pairing in repertoire data is technically challenging (Holec et al., 2018). Single cell sequencing eliminates this problem and a large dataset has just been added to VDJdb, which is, however, dominated by only few epitopes and HLA alleles. Another problem regarding TCR-epitope data is the lack of true negative datasets and the inclusion of cross-reactivity information, since many TCRs are able to recognize more than one epitope, which has been elaborated in section “Cross-reactivity assessment.” For this reason, pMHC/TCR binding prediction would also add valuable information to the detection of potential cross-reactivity for clinical candidate TCRs.

Further light on the details of pMHC/TCR interactions can be shed by molecular dynamics simulations. This entails understanding the role of hydrogen bonds, hydrophobic contacts, and interactions with the solvent in determining the specificity and cross-reactivity of each individual complex and proposing specific models of TCR engagement with the CDR1, CDR2, and CDR3 regions (Cuendet et al., 2011). Moreover, molecular modeling can help to compare the surface morphology between the complexed wild-type and mutated peptides and their relationship with immunogenicity (Park et al., 2013) and can also help to predict affinity-enhancing TCR mutations (Malecek et al., 2014). In cases where three-dimensional structures are not yet available, accurate models of pMHC/TCR complexes can be obtained by homology modeling (Zoete et al., 2013; Lanzarotti et al., 2019). Finally, a number of both rigid and flexible pMHC/TCR docking protocols have been proposed, which, in many cases, are able to produce accurate complex models starting from unbound structures (Pierce and Weng, 2013).

CONCLUSION AND OUTLOOK

Machine learning has become an indispensable tool for immunotherapeutic applications over the last decades. The established core method is peptide binding affinity prediction and thus target identification for TCR-T therapy or personalized neoantigen vaccination. The constant evolution of available training data as well as machine learning techniques, building on growing computational power, has improved the quality of binding affinity predictions. Focus has been on CD8⁺ cytotoxic T cells, but the substantial role of CD4⁺ T cells is increasingly gaining attention and efforts are made to also improve predictions for MHC class II presented epitopes, which poses a more challenging task compared to MHC class I binding due to the larger variety in peptide length and the open binding groove (Brown et al., 1993).

Additional challenges which can be tackled by machine learning remain. Immunogenicity is still an elusive aim for prediction tools, especially when it comes to personalized therapies relying on neoepitope identification. This is owed to the fact that patient immune systems and tumors undergo a process of mutual influence and therefore are highly

individual and heterogeneous. The identification of features derived from the immune system that affect T cell recognition of individual epitopes within a tumor could be the key toward more reliable personalized immunotherapy predictions, thereby opening the process to a broader number of patients. Although neoantigens are currently in the focus of cancer immunotherapy, the detection of shared tumor antigens beyond coding DNA regions remains necessary since not all tumors harbor enough immunogenic mutations and the creation of potent TCRs for individual patients is currently impossible. Another challenge, which can be tackled with the help of ongoing data acquisition, is TCR binding prediction. Being able to reliably predict which TCR will recognize which epitope is extremely valuable not only for target epitope identification for TCR-T therapies, but also especially for TCR safety assessment, since it can speed up the process of selecting TCRs for the clinic by reducing *in vitro* screening of TCR candidates.

As the TCR-T adoptive immunotherapy community grows and data on the impact of sequence variations in both TCR alpha and beta chains on peptide fine specificity, sensitivity of peptide-MHC recognition and TCR cross-reactivity for partially mismatched epitopes emerge, artificial intelligence in the form of machine learning will be critical to advance understanding of pMHC/TCR interactions for many types of antigen and many different HLA allotypes. In particular, these issues will become additionally relevant as this form of immunotherapy is developed for patient populations worldwide. High-throughput TCR discovery platforms, yielding TCR sequence information from natural repertoires of T cells or through TCR mutational analyses, coupled with functional assessment of peptide variants as a means to assess cross-reactivity, offer many opportunities to continually improve understanding of pMHC/TCR interactions that will not only advance the cause of basic science but also help to meet medical needs for patients with cancer, infectious diseases or autoimmunity, where it is envisioned that TCR-Ts have the potential to provide improved therapies worldwide.

In particular, the push to couple TCR sequence data with neoantigen recognition for single patients through analysis of individual tumor samples in order to develop more potent cancer vaccines or TCR-T immunotherapies has already fostered strong collaborations and commercial endeavors to advance the interplay of machine learning and TCR recognition. While it currently seems daunting to imagine how the enormous and fast flow of information now emerging from many sources can be accessed and assembled to rapidly support the broader needs for personalized patient-individualized TCR-based immunotherapies, this review summarizes the challenges as well as the substantial progress that has already been achieved in defining some of the most relevant parameters in the complex cell biology of antigen processing and presentation and pMHC interactions with TCRs that lead to successful immune recognition. Important gaps have also been defined, alerting the community to the types of control data that may already exist in many laboratories, or could be collected, that would help in

the refinement of prediction tools to achieve better results in the future. Increased interest and collaborative efforts of machine learning and HLA and TCR specialists will certainly foster further developments to support the rapidly expanding field of T cell-based immunotherapy of high medical relevance.

With the support of bioinformatic tools and improved prediction algorithms, immunotherapy holds the potential to become more precise, more personalized, and more effective

than current cancer treatments—and potentially with fewer side effects.

AUTHOR CONTRIBUTIONS

AM, SR, MW, DS, and DF all contributed to the writing and all approved the content of this review article.

REFERENCES

- Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46, 315–326. doi: 10.1016/j.immuni.2017.02.007
- Alvaro-Benito, M., Morrison, E., Abualrous, E. T., Kuroпка, B., and Freund, C. (2018). Quantification of HLA-DM-dependent major histocompatibility complex of class II immunopeptidomes by the peptide landscape antigenic epitope alignment utility. *Front. Immunol.* 9, 872. doi: 10.3389/fimmu.2018.00872
- Anagnostou, V., Smith, K. N., Forde, P. M., Niknafs, N., Bhattacharya, R., White, J., et al. (2017). Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discovery* 7, 264–276. doi: 10.1158/2159-8290.CD-16-0828
- Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511–517. doi: 10.1093/bioinformatics/btv639
- Andreatta, M., Jurtz, V. I., Kaefer, T., Sette, A., Peters, B., and Nielsen, M. (2017). Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* 152, 255–264. doi: 10.1111/imm.12763
- Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., et al. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* 16, 64. doi: 10.1186/s13059-015-0620-6
- Anonymous (2017) The problem with neoantigen prediction. *Nat. Biotechnol.* 35, 97–97. doi: 10.1038/nbt.3800
- Antunes, D. A., Devaurs, D., Moll, M., Lizée, G., and Kaviraki, L. E. (2018). General Prediction of peptide-mhc binding modes using incremental docking: a proof of concept. *Sci. Rep.* 8, 4327. doi: 10.1038/s41598-018-22173-4
- Antunes, D. A., Vieira, G. F., Rigo, M. M., Cibulski, S. P., Sinigaglia, M., and Chies, J. A. B. (2010). Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of mhc:peptide complexes to cross-reactivity assessment. *PLoS One* 5, e10353. doi: 10.1371/journal.pone.0010353
- Bai, Y., Wang, D., and Fury, W. (2018). “PHLAT: Inference of high-resolution HLA types from RNA and whole exome sequencing,” in *HLA Typing*. Ed. Boegel, S. (New York, NY: Springer New York), 193–201. doi: 10.1007/978-1-4939-8546-3_13
- Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e18. doi: 10.1016/j.cell.2018.02.060
- Bais, P., Namburi, S., Gatti, D. M., Zhang, X., and Chuang, J. H. (2017). CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics* 33, 3110–3112. doi: 10.1093/bioinformatics/btx375
- Baker, B. M., Scott, D. R., Blevins, S. J., and Hawse, W. F. (2012). Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism. *Immunol Rev.* 250, 10–31. doi: 10.1111/j.1600-065X.2012.01165.x
- Bassani-Sternberg, M., and Gfeller, D. (2016). Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide–HLA interactions. *J. Immunol.* 197, 2492–2499. doi: 10.4049/jimmunol.1600808
- Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., et al. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allosterically regulating HLA specificity. *PLoS Comput. Biol.* 13, e1005725. doi: 10.1371/journal.pcbi.1005725
- Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* 14, 658–673. doi: 10.1074/mcp.M114.042812
- Bhasin, M. (2004). Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* 13, 596–607. doi: 10.1110/ps.03373104
- Bilich, T., Nelde, A., Bichmann, L., Roerden, M., Salih, H. R., Kowalewski, D. J., et al. (2019). The HLA ligandome landscape of chronic myeloid leukemia delineates novel T-cell epitopes for immunotherapy. *Blood* 133, 550–565. doi: 10.1182/blood-2018-07-866830
- Bisset, L. R., and Fierz, W. (1993). Using a neural network to identify potential HLA-DR1 binding sites within proteins. *J. Mol. Recognit.* 6, 41–48. doi: 10.1002/jmr.300060105
- Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z., and Eklund, A. C. (2017a). MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* 66 (9), 1123–1130. doi: 10.1007/s00262-017-2001-3
- Bjerregaard, A.-M., Nielsen, M., Jurtz, V., Barra, C. M., Hadrup, S. R., Szallasi, Z., et al. (2017b). An analysis of natural t cell responses to predicted tumor neoepitopes. *Front. Immunol.* 8, 1566. doi: 10.3389/fimmu.2017.01566
- Blankenstein, T., Leisegang, M., Uckert, W., and Schreiber, H. (2015). Targeting cancer-specific mutations by T cell receptor gene therapy. *Curr. Opin. Immunol.* 33, 112–119. doi: 10.1016/j.coi.2015.02.005
- Boegel, S., Löwer, M., Bukur, T., Sahin, U., and Castle, J. C. (2014). A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *OncoImmunology* 3, e954893. doi: 10.4161/21624011.2014.954893
- Boegel, S., Löwer, M., Schäfer, M., Bukur, T., de Graaf, J., Boisguérin, V., et al. (2012). HLA typing from RNA-Seq sequencing reads. *Genome Med.* 4, 102. doi: 10.1186/gm403
- Bordner, A. J., and Abagyan, R. (2006). Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins: Struct. Funct. Bioinf.* 63, 512–526. doi: 10.1002/prot.20831
- Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., et al. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364, 33–39. doi: 10.1038/364033a0
- Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., et al. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* 24, 743–750. doi: 10.1101/gr.165985.113
- Brusic, V., Rudy, G., and Harrison, L. C. (1994). MHCPEP: a database of MHC-binding peptides. *Nucleic Acids Res.* 22, 3663–3665. doi: 10.1093/nar/22.17.3663
- Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., et al. (2019). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37, 55–63. doi: 10.1038/nbt.4313
- Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., et al. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* 9, e1003266. doi: 10.1371/journal.pcbi.1003266
- Caron, E., Vincent, K., Fortier, M.-H., Laverdure, J.-P., Bramouille, A., Hardy, M.-P., et al. (2011). The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Syst. Biol.* 7, 533–533. doi: 10.1038/msb.2011.68
- Carreno, B. M., Magrini, V., Becker-Hapak, M., Kaabinejad, S., Hundal, J., Petti, A. A., et al. (2015). A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* 348, 803–808. doi: 10.1126/science.1253828
- Castle, J. C., Kreiter, S., Diekmann, J., Lower, M., van de Roemer, N., de Graaf, J., et al. (2012). Exploiting the mutanome for tumor vaccination. *Cancer Res.* 72, 1081–1091. doi: 10.1158/0008-5472.CAN-11-3722

- Chang, T.-C., Carter, R. A., Li, Y., Li, Y., Wang, H., Edmonson, M. N., et al. (2017). The neoepitope landscape in pediatric cancers. *Genome Med.* 9, 78. doi: 10.1186/s13073-017-0468-3
- Charoentong, P., Finotello, E., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019
- Cohen, C. J., Gartner, J. J., Horovitz-Fried, M., Shamalov, K., Trebska-McGowan, K., Bliskovsky, V. V., et al. (2015). Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *J. Clin. Invest.* 125, 3981–3991. doi: 10.1172/JCI82416
- Cuendet, M. A., Zoete, V., and Michielin, O. (2011). How T cell receptors interact with peptide-MHCs: A multiple steered molecular dynamics study. *Proteins: Struct Funct. Bioinf.* 79, 3007–3024. doi: 10.1002/prot.23104
- Daniel, S., Brusici, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., et al. (1998). Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J. Immunol.* 161, 617–624.
- Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547, 89–93. doi: 10.1038/nature22383
- De Neuter, N., Bittremieux, W., Beirnaert, C., Cuypers, B., Mrzic, A., Moris, P., et al. (2018). On the feasibility of mining CD8⁺ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics.* 70 (3), 159–168. doi: 10.1007/s00251-017-1023-5
- Dhanda, S. K., Mahajan, S., Paul, S., Yan, Z., Kim, H., Jespersen, M. C., et al. (2019). IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res.* 47, W502–W506. doi: 10.1093/nar/gkz452
- Diez-Rivero, C. M., Chenlo, B., Zuluaga, P., and Reche, P. A. (2010). Quantitative modeling of peptide binding to TAP using support vector machine. *Proteins: Struct Funct. Bioinf.* 78, 63–72. doi: 10.1002/prot.22535
- Dolan, B. P. (2019). “Quantitating MHC Class I ligand production and presentation using TCR-like antibodies,” in *Antigen Processing*. Ed. van Endert, P. (New York, NY: Springer New York), 149–157. doi: 10.1007/978-1-4939-9450-2_12
- Doytchinova, I. A., Guan, P., and Flower, D. R. (2006). EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinf.* 7, 131. doi: 10.1186/1471-2105-7-131
- Duan, F., Duitama, J., Al Seesi, S., Ayres, C. M., Corcelli, S. A., Pawashe, A. P., et al. (2014). Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* 211, 2231–2248. doi: 10.1084/jem.20141308
- Elliott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandath, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 6, 271–281.e7. doi: 10.1016/j.cels.2018.03.002
- Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., et al. (2018). Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* 15, 363–366. doi: 10.1038/nmeth.4631
- Falk, K., Röttschke, O., Stevanović, S., Jung, G., and Rammensee, H.-G. (1991). Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351, 290–296. doi: 10.1038/351290a0
- Fleri, W., Paul, S., Dhanda, S. K., Mahajan, S., Xu, X., Peters, B., et al. (2017). The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.* 8, 278. doi: 10.3389/fimmu.2017.00278
- Freudenmann, L. K., Marcu, A., and Stevanović, S. (2018). Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* 154, 331–345. doi: 10.1111/imm.12936
- Garde, C., Ramarathinam, S. H., Jappe, E. C., Nielsen, M., Kringelum, J. V., Trolle, T., et al. (2019). Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics.* 71 (7), 445–454. doi: 10.1007/s00251-019-01122-z
- Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting antigen presentation—what could we learn from a million peptides? *Front. Immunol.* 9, 1716. doi: 10.3389/fimmu.2018.01716
- Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R. T., Racle, J., et al. (2018). The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* 201, 3705–3716. doi: 10.4049/jimmunol.1800914
- Ghorani, E., Rosenthal, R., McGranahan, N., Reading, J. L., Lynch, M., Peggs, K. S., et al. (2018). Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* 29, 271–279. doi: 10.1093/annonc/mdx687
- Giam, K., Ayala-Perez, R., Illing, P. T., Schittenhelm, R. B., Croft, N. P., Purcell, A. W., et al. (2015). A comprehensive analysis of peptides presented by HLA-A1: A comprehensive analysis of peptides. *Tissue Antigens* 85, 492–496. doi: 10.1111/tan.12565
- Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547, 94–98. doi: 10.1038/nature22976
- Goh, G., Walradt, T., Markarov, V., Blom, A., Riaz, N., Doumani, R., et al. (2016). Mutational landscape of MCPyV-positive and MCPyV-negative Merkel cell carcinomas with implications for immunotherapy. *Oncotarget* 7, 3403–3415. doi: 10.18632/oncotarget.6494
- Griffin, T. A., Nandi, D., Cruz, M., Fehling, H. J., Kaer, L. V., Monaco, J. J., et al. (1998). Immunoproteasome assembly: cooperative incorporation of interferon γ (IFN- γ)-inducible subunits. *J. Exp. Med.* 187, 97–104. doi: 10.1084/jem.187.1.97
- Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., et al. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat. Med.* 22, 433–438. doi: 10.1038/nm.4051
- Gubin, M. M., Zhang, X., Schuster, H., Caron, E., Ward, J. P., Noghuchi, T., et al. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* 515, 577–581. doi: 10.1038/nature13988
- Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S., and van Endert, P. M. (1998). Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol. Immunol.* 35, 427–433. doi: 10.1016/s0161-5890(98)00059-5
- Haase, K., Raffegerst, S., Schendel, D. J., and Frishman, D. (2015). Expitope: a web server for epitope expression. *Bioinformatics* 31, 1854–1856. doi: 10.1093/bioinformatics/btv068
- Hassan, C., Kester, M. G. D., Oudgenoeg, G., de Ru, A. H., Janssen, G. M. C., Drijfhout, J. W., et al. (2014). Accurate quantitation of MHC-bound peptides by application of isotopically labeled peptide MHC complexes. *J. Proteomics* 109, 240–244. doi: 10.1016/j.jpro.2014.07.009
- Hilf, N., Kutttruff-Coqui, S., Frenzel, K., Bukur, V., Stevanović, S., Gouttefangeas, C., et al. (2019). Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* 565, 240–245. doi: 10.1038/s41586-018-0810-y
- Holec, P. V., Berleant, J., Bathe, M., and Birnbaum, M. E. (2018). A Bayesian framework for high-throughput T cell receptor pairing. *Bioinformatics.* 35 (8), 1318–1325. doi: 10.1093/bioinformatics/bty801
- Hu, Y., Wang, Z., Hu, H., Wan, F., Chen, L., Yuanpeng, X., et al. (2019). ACME: Pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics.* doi: 10.1093/bioinformatics/btz427
- Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., et al. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 165, 35–44. doi: 10.1016/j.cell.2016.02.065
- Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Wollam, A. T., Xia, H., et al. (2019). pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *bioRxiv.* doi: 10.1101/501817
- Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., et al. (1992). Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255, 1261–1263. doi: 10.1126/science.1546328
- Jaravine, V., Mösch, A., Raffegerst, S., Schendel, D. J., and Frishman, D. (2017a). Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. *BMC Cancer* 17, 892. doi: 10.1186/s12885-017-3854-8
- Jaravine, V., Raffegerst, S., Schendel, D. J., and Frishman, D. (2017b). Assessment of cancer and virus antigens for cross-reactivity in human tissues. *Bioinformatics* 33, 104–111. doi: 10.1093/bioinformatics/btw567
- Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., et al. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154, 394–406. doi: 10.1111/imm.12889
- Johnson, L. A., Morgan, R. A., Dudley, M. E., Cassard, L., Yang, J. C., Hughes, M. S., et al. (2009). Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood* 114, 535–546. doi: 10.1182/blood-2009-03-211714

- Jørgensen, K. W., Rasmussen, M., Buus, S., and Nielsen, M. (2014). NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* 141, 18–26. doi: 10.1111/imm.12160
- Jurtz, V. I., Jessen, L. E., Bentzen, A. K., Jespersen, M. C., Mahajan, S., Vita, R., et al. (2018). NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. doi:10.1101/433706
- Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199, 3360–3368. doi: 10.4049/jimmunol.1700893
- Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., et al. (2018). Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* 34, 211–224.e6. doi: 10.1016/j.ccell.2018.07.001
- Kalaora, S., Barnea, E., Merhavi-Shoham, E., Qutob, N., Teer, J. K., Shimony, N., et al. (2016). Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* 7, 5110–5117. doi: 10.18632/oncotarget.6960
- Kalaora, S., Wolf, Y., Feferman, T., Barnea, E., Greenstein, E., Reshef, D., et al. (2018). Combined analysis of antigen presentation and t-cell recognition reveals restricted immune responses in melanoma. *Cancer Discovery* 8, 1366–1375. doi: 10.1158/2159-8290.CD-17-1418
- Karasaki, T., Nagayama, K., Kawashima, M., Hiya, N., Murayama, T., Kuwano, H., et al. (2016). Identification of individual cancer-specific somatic mutations for neoantigen-based immunotherapy of lung cancer. *J. Thoracic Oncol.* 11, 324–333. doi: 10.1016/j.jtho.2015.11.006
- Karasaki, T., Nagayama, K., Kuwano, H., Nitadori, J., Sato, M., Anraku, M., et al. (2017). Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer Sci.* 108, 170–177. doi: 10.1111/cas.13131
- Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64, 177–186. doi: 10.1007/s00251-011-0579-8
- Kato, T., Park, J.-H., Kiyotani, K., Ikeda, Y., Miyoshi, Y., and Nakamura, Y. (2017). Integrated analysis of somatic mutations and immune microenvironment of multiple regions in breast cancers. *Oncotarget* 8, 62029–62038. doi: 10.18632/oncotarget.18790
- Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998
- Keskin, D. B., Anandappa, A. J., Sun, J., Tirosh, I., Mathewson, N. D., Li, S., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* 565, 234–239. doi: 10.1038/s41586-018-0792-9
- Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., and Brunak, S. (2002). Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 15, 287–296. doi: 10.1093/protein/15.4.287
- Khalili, J. S., Hanson, R. W., and Szallasi, Z. (2012). In silico prediction of tumor antigens derived from functional missense mutations of the cancer gene census. *OncolImmunology* 1, 1281–1289. doi: 10.4161/onci.21511
- Kim, S., Kim, H. S., Kim, E., Lee, M. G., Shin, E., Paik, S., et al. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann. Oncol.* 29 (4), 1030–1036. doi: 10.1093/annonc/mdy022
- Kinhead, H. L., Hopkins, A., Lutz, E., Wu, A. A., Yarchoan, M., Cruz, K., et al. (2018). Combining STING-based neoantigen-targeted vaccine with checkpoint modulators enhances antitumor immunity in murine pancreatic cancer. *JCI Insight* 3, e122857. doi: 10.1172/jci.insight.122857
- Klausen, M. S., Anderson, M. V., Jespersen, M. C., Nielsen, M., and Marcatili, P. (2015). LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.* 43, W349–W355. doi: 10.1093/nar/gkv535
- Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., et al. (2015). Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One* 10, e0141561. doi: 10.1371/journal.pone.0141561
- Koşaloğlu-Yalçın, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premalal, A., Sidney, J., Vaughan, K., et al. (2018). Predicting T cell recognition of MHC class I restricted neopeptides. *OncolImmunology* 7, e1492508. doi: 10.1080/2162402X.2018.1492508
- Koster, J., and Plasterk, R. H. A. (2019). A library of Neo Open Reading Frame peptides (NOPs) as a sustainable resource of common neoantigens in up to 50% of cancer patients. *Sci. Rep.* 9, 6577. doi: 10.1038/s41598-019-42729-2
- Kreiter, S., Vormehr, M., van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692–696. doi: 10.1038/nature14426
- Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H.-G., Schild, H., and Haderl, K.-P. (2000). An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.* 298, 417–429. doi: 10.1006/jmbi.2000.3683
- Kyeong, H.-H., Choi, Y., and Kim, H.-S. (2018). GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics* 34, 469–476. doi: 10.1093/bioinformatics/btx589
- Lam, T., Mamitsuka, H., Ren, E., and Tong, J. (2010). TAP Hunter: a SVM-based system for predicting TAP ligands using local description of amino acid sequence. *Immunome Res.* 6, S6. doi: 10.1186/1745-7580-6-S1-S6
- Lanzarotti, E., Marcatili, P., and Nielsen, M. (2019). T-cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. *Front. Immunol.* 10, 2080. doi: 10.3389/fimmu.2019.02080
- Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinf.* 8, 424. doi: 10.1186/1471-2105-8-424
- Laumont, C. M., and Perreault, C. (2017). Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell. Mol. Life Sci.* 75 (4), 607–621. doi: 10.1007/s00018-017-2628-4
- Laumont, C. M., Vincent, K., Hesnard, L., Audemar, É., Bonneil, É., Laverdure, J.-P., et al. (2018). Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Trans. Med.* 10, eaau5516. doi: 10.1126/scitranslmed.aau5516
- Lin, H., Ray, S., Tongchusak, S., Reinherz, E. L., and Brusica, V. (2008). Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research. *BMC Immunol.* 9, 8. doi: 10.1186/1471-2172-9-8
- Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* 122, 863–871. doi: 10.1182/blood-2013-03-490565
- Linnemann, C., van Buuren, M. M., Bies, L., Verdegaal, E. M. E., Schotte, R., Calis, J. J. A., et al. (2014). High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4⁺ T cells in human melanoma. *Nat. Med.* 21, 81–85. doi: 10.1038/nm.3773
- Liu, C., Yang, X., Duffly, B., Mohanakumar, T., Mitra, R. D., Zody, M. C., et al. (2013). ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* 41, e142–e142. doi: 10.1093/nar/gkt481
- Liu, G., Li, D., Li, Z., Qiu, S., Li, W., Chao, C., et al. (2017). PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *GigaScience* 6, 1–11. doi: 10.1093/gigascience/gix017
- Liu, S., Matsuzaki, J., Wei, L., Tsuji, T., Battaglia, S., Hu, Q., et al. (2019). Efficient identification of neoantigen-specific T-cell responses in advanced human ovarian cancer. *J. Immuno Ther Cancer* 7, 156. doi: 10.1186/s40425-019-0629-6
- Löffler, M. W., Chandran, P. A., Laske, K., Schroeder, C., Bonzheim, I., Walzer, M., et al. (2016). Personalized peptide vaccine-induced immune response associated with long-term survival of a metastatic cholangiocarcinoma patient. *J. Hepatol* 65, 849–855. doi: 10.1016/j.jhep.2016.06.027
- Löffler, M. W., HEPVAC Consortium Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., et al. (2019). Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med.* 11, 28. doi: 10.1186/s13073-019-0636-8
- Łuksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovytov, A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 551 (7681), 517–520. doi: 10.1038/nature24473
- Lundegaard, C., Lund, O., and Nielsen, M. (2008). 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* 24, 1397–1398. doi: 10.1093/bioinformatics/btn128
- Madden, D. R. (1995). The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.* 13, 587–622. doi: 10.1146/annurev.iy.13.040195.003103

- Malecek, K., Grigoryan, A., Zhong, S., Gu, W. J., Johnson, L. A., Rosenberg, S. A., et al. (2014). Specific Increase in Potency via Structure-Based Design of a TCR. *J. Immunol.* 193, 2587–2599. doi: 10.4049/jimmunol.1302344
- Mamitsuka, H. (1998). Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 33, 460–474. doi: 10.1002/(sici)1097-0134(19981201)33:4<460::aid-prot2>3.0.co;2-m
- Marino, F., Chong, C., Michaux, J., and Bassani-Sternberg, M., (2019). “High-throughput, fast, and sensitive immunopeptidomics sample processing for mass spectrometry,” in *Immune Checkpoint Blockade*. Ed. Pico de Coaña, Y. (New York, NY: Springer New York), 67–79. doi: 10.1007/978-1-4939-8979-9_5
- Martin, S. D., Wick, D. A., Nielsen, J. S., Little, N., Holt, R. A., and Nelson, B. H. (2018). A library-based screening method identifies neoantigen-reactive T cells in peripheral blood prior to relapse of ovarian cancer. *Oncot Immunology* 7, e1371895. doi: 10.1080/2162402X.2017.1371895
- Marty, R., Kaabinejadian, S., Rossell, D., Sliker, M. J., van de Haar, J., Engin, H. B., et al. (2017). MHC-I genotype restricts the oncogenic mutational landscape. *Cell* 171, 1272–1283.e15. doi: 10.1016/j.cell.2017.09.050
- McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351, 1463–1469. doi: 10.1126/science.aaf1490
- McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., et al. (2017). Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* 171, 1259–1271.e11. doi: 10.1016/j.cell.2017.10.001
- Mendes, M. F. A., Antunes, D. A., Rigo, M. M., Sinigaglia, M., and Vieira, G. F. (2015). Improved structural method for T-cell cross-reactivity prediction. *Mol. Immunol.* 67, 303–310. doi: 10.1016/j.molimm.2015.06.017
- Menegatti Rigo, M., Amaral Antunes, D., Vaz de Freitas, M., Fabiano de Almeida Mendes, M., Meira, L., Sinigaglia, M., et al. (2015). DockTope: a web-based tool for automated pMHC-I modelling. *Sci. Rep.* 5, 18413. doi: 10.1038/srep18413
- Miller, A., Asmann, Y., Cattaneo, L., Braggio, E., Keats, J., Auclair, D., et al. (2017). High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood Cancer J.* 7, e612. doi: 10.1038/bcj.2017.94
- Morgan, R. A., Chinnasamy, N., Abate-Daga, D., Gros, A., Robbins, P. F., Zheng, Z., et al. (2013). Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immunother.* 36, 133–151. doi: 10.1097/CJI.0b013e3182829903
- Moutafsi, M., Peters, B., Pasquetto, V., Tschärke, D. C., Sidney, J., Bui, H.-H., et al. (2006). A consensus epitope prediction approach identifies the breadth of murine T_{CD8+}-cell responses to vaccinia virus. *Nat. Biotechnol.* 24, 817–819. doi: 10.1038/nbt1215
- Nathanson, T., Ahuja, A., Rubinsteyn, A., Aksoy, B. A., Hellmann, M. D., Miao, D., et al. (2017). Somatic mutations and neoepitope homology in melanomas treated with CTLA-4 blockade. *Cancer Immunol. Res.* 5, 84–91. doi: 10.1158/2326-6066.CIR-16-0019
- Neeffes, J., Jongsma, M. L. M., Paul, P., and Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* 11 (12), 823–836. doi: 10.1038/nri3084
- Nielsen, M., and Andreatta, M. (2017). NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions. *Nucleic Acids Res.* 45, W344–W349. doi: 10.1093/nar/gkx276
- Nielsen, M., Justesen, S., Lund, O., Lundegaard, C., and Buus, S. (2010). NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res.* 6, 9. doi: 10.1186/1745-7580-6-9
- Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinf.* 8, 238. doi: 10.1186/1471-2105-8-238
- Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57, 33–41. doi: 10.1007/s00251-005-0781-7
- Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., et al. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12, 1007–1017. doi: 10.1110/ps.0239403
- O’Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018b). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7 (1), 129–132.e4. doi: 10.1016/j.cels.2018.05.014
- O’Donnell, T., Christie, E. L., Ahuja, A., Buross, J., Aksoy, B. A., Bowtell, D. D. L., et al. (2018a). Chemotherapy weakly contributes to predicted neoantigen expression in ovarian cancer. *BMC Cancer* 18, 87. doi: 10.1186/s12885-017-3825-0
- Ogishi, M., and Yotsuyanagi, H. (2019). Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* 10, 827. doi: 10.3389/fimmu.2019.00827
- Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. doi: 10.1038/nature22991
- Park, M.-S., Park, S. Y., Miller, K. R., Collins, E. J., and Lee, H. Y. (2013). Accurate structure prediction of peptide–MHC complexes for identifying highly immunogenic antigens. *Mol. Immunol.* 56, 81–90. doi: 10.1016/j.molimm.2013.04.011
- Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 152, 163–175.
- Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., et al. (2006). A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* 2, e65. doi: 10.1371/journal.pcbi.0020065
- Peters, B., Bulik, S., Tampe, R., van Endert, P. M., and Holzhutter, H.-G. (2003). Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* 171, 1741–1749. doi: 10.4049/jimmunol.171.4.1741
- Pierce, B. G., and Weng, Z. (2013). A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Sci.* 22, 35–46. doi: 10.1002/pro.2181
- Rajasagi, M., Shukla, S. A., Fritsch, E. F., Keskin, D. B., DeLuca, D., Carmona, E., et al. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* 124, 453–462. doi: 10.1182/blood-2014-04-567933
- Rammensee, H. G., Falk, K., and Rötzschke, O. (1993). Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.* 11, 213–244. doi: 10.1146/annurev.iy.11.040193.001241
- Rammensee, H. G., Friede, T., and Stevanović, S. (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics* 41, 178–228. doi: 10.1007/bf00172063
- Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50, 213–219. doi: 10.1007/s002510050595
- Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A. B., Nielsen, I. K., Nielsen, M., et al. (2016). Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* 197, 1517–1524. doi: 10.4049/jimmunol.1600582
- Reche, P. A., Glutting, J.-P., and Reinherz, E. L. (2002). Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* 63, 701–709. doi: 10.1016/s0198-8859(02)00432-9
- Reuben, A., Gittelman, R., Gao, J., Zhang, J., Yusko, E. C., Wu, C.-J., et al. (2017). TCR Repertoire intratumor heterogeneity in localized lung adenocarcinomas: an association with predicted neoantigen heterogeneity and postsurgical recurrence. *Cancer Discovery* 7, 1088–1097. doi: 10.1158/2159-8290.CD-17-0256
- Ritz, D., Gloger, A., Neri, D., and Fugmann, T. (2017). Purification of soluble HLA class I complexes from human serum or plasma deliver high quality immunopeptidomes required for biomarker discovery. *PROTEOMICS* 17, 1600364. doi: 10.1002/pmic.201600364
- Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., and Fugmann, T. (2016). High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients’ sera. *PROTEOMICS* 16, 1570–1580. doi: 10.1002/pmic.201500445
- Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124–128. doi: 10.1126/science.aaa1348
- Robbins, P. F., Lu, Y.-C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., et al. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* 19, 747–752. doi: 10.1038/nm.3161

- Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43, D423–D431. doi: 10.1093/nar/gku1161
- Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160, 48–61. doi: 10.1016/j.cell.2014.12.033
- Rosenthal, R., Cadieux, E. L., Salgado, R., Moore, D. A., Lund, T., Tanić, M., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature* 567, 479–485. doi: 10.1038/s41586-019-1032-7
- Rothbard, J. B., and Taylor, W. R. (1988). A sequence pattern common to T cell epitopes. *EMBO J.* 7, 93–100. doi: 10.1002/j.1460-2075.1988.tb02787.x
- Sadelain, M., Brentjens, R., and Rivière, I. (2013). The basic principles of chimeric antigen receptor design. *Cancer Discovery* 3, 388–398. doi: 10.1158/2159-8290.CD-12-0548
- Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222–226. doi: 10.1038/nature23003
- Schischlik, F., Jäger, R., Rosebrock, F., Hug, E., Schuster, M. K., Holly, R., et al. (2019). Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood*, blood.2019000519. doi:10.1182/blood.2019000519
- Schubert, B., Walzer, M., Brachvogel, H.-P., Szolek, A., Mohr, C., and Kohlbacher, O. (2016). FRED 2: an immunoinformatics framework for Python. *Bioinformatics* 32, 2044–2046. doi: 10.1093/bioinformatics/btw113
- Schumacher, T. N., and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74. doi: 10.1126/science.aaa4971
- Segal, N. H., Parsons, D. W., Peggs, K. S., Velculescu, V., Kinzler, K. W., Vogelstein, B., et al. (2008). Epitope landscape in breast and colorectal cancer. *Cancer Res.* 68, 889–892. doi: 10.1158/0008-5472.CAN-07-3095
- Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., et al. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 46, D419–D427. doi: 10.1093/nar/gkx760
- Shukla, S. A., Rooney, M. S., Rajasagi, M., Tiao, G., Dixon, P. M., Lawrence, M. S., et al. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* 33, 1152–1158. doi: 10.1038/nbt.3344
- Singh, H., and Raghava, G. P. S. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17, 1236–1237. doi: 10.1093/bioinformatics/17.12.1236
- Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., et al. (2018). Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.* 36 (11), 1056–1058. doi: 10.1038/nbt.4239
- Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New Engl. J. Med.* 371, 2189–2199. doi: 10.1056/NEJMoa1406498
- Sonntag, K., Hashimoto, H., Eyrich, M., Menzel, M., Schubach, M., Döcker, D., et al. (2018). Immune monitoring and TCR sequencing of CD4 T cells in a long term responsive patient with metastasized pancreatic ductal carcinoma treated with individualized, neoepitope-derived multi-peptide vaccines: a case report. *J. Trans. Med.* 16, 23. doi: 10.1186/s12967-018-1382-1
- Spranger, S., Luke, J. J., Bao, R., Zha, Y., Hernandez, K. M., Li, Y., et al. (2016). Density of immunogenic antigens does not explain the presence or absence of the T-cell-inflamed tumor microenvironment in melanoma. *Proc. Natl. Acad. Sci.* 113, E7759–E7768. doi: 10.1073/pnas.1609376113
- Storkus, W. J., Zeh, H. J., Salter, R. D., and Lotze, M. T. (1993). Identification of T-cell epitopes: rapid isolation of class I-presented peptides from viable cells by mild acid elution. *J. Immunother. Emphasis Tumor Immunol.* 14, 94–103.
- Stranzl, T., Larsen, M. V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62, 357–368. doi: 10.1007/s00251-010-0441-4
- Strønen, E., Toebes, M., Kelderman, S., van Buuren, M. M., Yang, W., van Rooij, N., et al. (2016). Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* 352, 1337–1341. doi: 10.1126/science.aaf2288
- Sugawara, S., Abo, T., and Kumagai, K. (1987). A simple method to eliminate the antigenicity of surface class I MHC molecules from the membrane of viable cells by acid treatment at pH 3. *J. Immunol. Methods* 100, 83–90. doi: 10.1016/0022-1759(87)90175-x
- Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 30, 3310–3316. doi: 10.1093/bioinformatics/btu548
- Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812–830.e14. doi: 10.1016/j.immuni.2018.03.023
- Tran, E., Ahmadvadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P. F., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350, 1387–1390. doi: 10.1126/science.aad1253
- Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaefer, T., et al. (2016). The length distribution of class I-restricted t cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J. Immunol.* 196, 1480–1487. doi: 10.4049/jimmunol.1501721
- Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 350, 207–211. doi: 10.1126/science.aad0095
- van Buuren, M. M., Calis, J. J., and Schumacher, T. N. (2014). High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. *Oncot Immunology* 3, e28836. doi: 10.4161/onci.28836
- van Gool, I. C., Eggink, F. A., Freeman-Mills, L., Stelloo, E., Marchi, E., de Bruyn, M., et al. (2015). POLE proofreading mutations elicit an antitumor immune response in endometrial cancer. *Clin. Cancer Res.* 21, 3347–3355. doi: 10.1158/1078-0432.CCR-15-0057
- van Rooij, N., van Buuren, M. M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., et al. (2013). Tumor exome analysis reveals neoantigen-specific T-Cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* 31, e439–e442. doi: 10.1200/JCO.2012.47.7521
- Veatch, J. R., Lee, S. M., Fitzgibbon, M., Chow, I.-T., Jesernig, B., Schmitt, T., et al. (2018). Tumor-infiltrating BRAF^{v600E}-specific CD4⁺ T cells correlated with complete clinical response in melanoma. *J. Clin. Invest.* 128, 1563–1568. doi: 10.1172/JCI98689
- Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2018). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006
- Vrecko, S., Guenat, D., Mercier-Letondal, P., Faucheu, H., Dosset, M., Royer, B., et al. (2018). Personalized identification of tumor-associated immunogenic neoepitopes in hepatocellular carcinoma in complete remission after sorafenib treatment. *Oncotarget* 9, 35394–35407. doi: 10.18632/oncotarget.26247
- Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., et al. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinf.* 11, 568. doi: 10.1186/1471-2105-11-568
- Warren, R. L., and Holt, R. A. (2010). A census of predicted mutational epitopes suitable for immunologic cancer control. *Hum. Immunol.* 71, 245–254. doi: 10.1016/j.humimm.2009.12.007
- Warren, R. L., Choe, G., Freeman, D. J., Castellarin, M., Munro, S., Moore, R., et al. (2012). Derivation of HLA types from shotgun sequence datasets. *Genome Med.* 4, 95. doi: 10.1186/gm396
- Wood, M. A., Paralkar, M., Paralkar, M. P., Nguyen, A., Struck, A. J., Ellrott, K., et al. (2018). Population-level distribution and putative immunogenicity of cancer neoepitopes. *BMC Cancer* 18, 414. doi: 10.1186/s12885-018-4325-6
- Wu, J., Zhao, W., Zhou, B., Su, Z., Gu, X., Zhou, Z., et al. (2018). TSNAdb: a database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics Proteomics Bioinf.* 16, 276–282. doi: 10.1016/j.gpb.2018.06.003
- Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* 515, 572–576. doi: 10.1038/nature14001
- Yanover, C., and Bradley, P. (2011). Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6981–6986. doi: 10.1073/pnas.1018165108
- Zacharakis, N., Chinnasamy, H., Black, M., Xu, H., Lu, Y.-C., Zheng, Z., et al. (2018). Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat. Med.* 24, 724–730. doi: 10.1038/s41591-018-0040-8

- Zhang, G. L., Petrovsky, N., Kwoh, C. K., August, J. T., and Brusic, V. (2006). PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res.* 2, 3. doi: 10.1186/1745-7580-2-3
- Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 1293–1299. doi: 10.1093/bioinformatics/btp137
- Zhang, X., Kim, S., Hundal, J., Herndon, J. M., Li, S., Petti, A. A., et al. (2017). Breast cancer neoantigens can induce CD8⁺ T-cell responses and antitumor immunity. *Cancer Immunol. Res.* 5, 516–523. doi: 10.1158/2326-6066.CIR-16-0264
- Zoete, V., Irving, M., Ferber, M., Cuendet, M. A., and Michielin, O. (2013). Structure-based, rational design of T cell receptors. *Front. Immunol.* 4, 268. doi: 10.3389/fimmu.2013.00268

Conflict of Interest: AM, SR, and MW are employees and DS is a Managing Director of Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, Planegg, Germany.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Mösch, Raffeggerst, Weis, Schendel and Frishman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.