Bioinformatics, 33(1), 2017, 104–111
doi: 10.1093/bioinformatics/btw567
Advance Access Publication Date: 10 September 2016
Original Paper

Systems biology

Assessment of cancer and virus antigens for cross-reactivity in human tissues

Victor Jaravine[1], Silke Raffegerst[2], Dolores J. Schendel[2] and Dmitrij Frishman[1,3,4,*]

[1]Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising 85354, Germany, [2]Medigene Immunotherapies GmbH, Martinsried/Planegg 82152, Germany, [3]Helmholtz Center Munich - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg 85764, Germany and [4]St Petersburg State Polytechnic University, St Petersburg 195251, Russia

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on March 22, 2016; revised on August 3, 2016; accepted on August 26, 2016

Abstract

Motivation: Cross-reactivity (CR) or invocation of autoimmune side effects in various tissues has important safety implications in adoptive immunotherapy directed against selected antigens. The ability to predict CR (on-target and off-target toxicities) may help in the early selection of safer therapeutically relevant target antigens.

Results: We developed a methodology for the calculation of quantitative CR for any defined peptide epitope. Using this approach, we performed assessment of 4 groups of 283 currently known human MHC-class-I epitopes including differentiation antigens, overexpressed proteins, cancer-testis antigens and mutations displayed by tumor cells. In addition, 89 epitopes originating from viral sources were investigated. The natural occurrence of these epitopes in human tissues was assessed based on proteomics abundance data, while the probability of their presentation by MHC-class-I molecules was modelled by the method of Keşmir *et al.* which combines proteasomal cleavage, TAP affinity and MHC-binding predictions. The results of these analyses for many previously defined peptides are presented as CR indices and tissue profiles. The methodology thus allows for quantitative comparisons of epitopes and is suggested to be suited for the assessment of epitopes of candidate antigens in an early stage of development of adoptive immunotherapy.

Availability and Implementation: Our method is implemented as a Java program, with curated datasets stored in a MySQL database. It predicts all naturally possible self-antigens for a given sequence of a therapeutic antigen (or epitope) and after filtering for predicted immunogenicity outputs results as an index and profile of CR to the self-antigens in 22 human tissues. The program is implemented as part of the iCrossR webserver, which is publicly available at http://webclu.bio.wzw.tum.de/icrossr/.

Contact: d.frishman@wzw.tum.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As the use of receptor-engineered T cells, using T-cell receptors (TCRs) has moved forward as a strategic vision for immunotherapy of cancer, viral diseases and even autoimmunity, the selection of target antigens has become the issue of central importance.

The clinical power of the adoptive transfer of receptor-engineered T cells is indicated by remarkable clinical efficacy in cases of patients with liquid and solid tumor indications (Maus *et al.*, 2014; Zang *et al.*, 2014). The early clinical success of adoptive T-cell therapy is tempered by the risk of significant, even lethal, toxicity whereby the specificity of antigen recognition is a deciding factor. The basis of the toxicities of immunotherapy for cancer for the majority of immune-related adverse events is a hyper-activated T-cell response with reactivity directed against normal tissue (Weber *et al.*, 2015). Immune cross-reactivity (CR) arises when the sequence similarity between foreign and self-peptides or between two different self-peptides is sufficient to result in binding of the peptide-MHC complex and the TCR, leading to cross-activation of unwanted autoimmune responses of T cells (Kohm *et al.*, 2003).

Cytotoxic T lymphocytes (CTL) or CD8+ human T cells fight intracellular infections. Each step of a CTL response directed against an antigen contributes to a certain degree to CR: the cellular digestion of antigen and native proteins, the transport of the processed peptides to the endoplasmic reticulum, the binding of peptides to MHC molecules and the binding of peptide-MHC complexes to the TCR.

The human TCR exhibits a very high CR level—one T cell reacts with approximately 10^6 different MHC-associated minimal peptide epitopes (Mason, 1998). If a TCR reacts with a specific peptide then the probability that it will react with another randomly chosen peptide is only $\sim 10^{-4}$. Thus, binding to TCR is both specific and highly cross-reactive (Frank, 2002). The balance for specificity that differentiates between antigens versus CR that occurs due to similarity in different antigens is critical for the ability of the immune system to provide effective responses against pathogens.

Here, we model three of the steps of the T-cell response using the established methods and available human proteomics data to obtain quantitative measures of potential side effects of many epitopes identified for cancer immunotherapy as well as for epitopes identified for most major human viral infections.

2 Materials and methods

2.1 Dataset of tumor antigenic peptides

A dataset of 431 tumor-specific antigens was obtained from the Cancer Immunity Peptide Database (Vigneron *et al.*, 2013). This database is intended for clinicians contemplating the development of immunotherapy trials. Thus it lists only comprehensively characterized and validated peptides, i.e. those with demonstrated immunogenicity and identification of T-cell recognition, known human leukocyte antigen (HLA) allele presentation, isolated CTL clones or lines and natural presentation of these antigenic peptides by tumor cells.

According to the common classification (Coulie *et al.*, 2014), the peptides are subdivided into two categories and into a total of four groups based on the potential of the immunotherapy to be tumor-specific or elicit side effects in normal tissues. From this dataset, we selected the peptides recognized by cytotoxic T cells, which are binding to MHC-I with a specified allele. Consequently, the first category, further referred to as the group A ‘Mutation’, includes 40 unique antigens having a high specificity to 1 tumor, resulting from point mutations in genes that are expressed ubiquitously (Supplementary Table S1A). The second category consists of ‘Shared’ antigens, which are present in many independent tumors, sub-divided into three groups:

- The group B ‘Tumor-specific’ (Supplementary Table S1B) includes 67 peptides encoded by ‘Cancer-germline’ [or cancer/testis (CT)] genes. They are expressed in many tumors but not in normal tissues, with the exception of placental trophoblasts and testicular germ cells. Because the latter cells do not express MHC class I molecules, peptide expression should not result in epitope presentation and therefore this group can be considered to be tumor-specific.
- The group C ‘Differentiation’ (Supplementary Table S1C) includes 57 peptides expressed both in tumors and in the normal tissue of origin of the malignancy. Thus, they are not tumor specific and their use as targets for cancer immunotherapy may result in autoimmunity towards normal tissues.
- The group D ‘Overexpressed’ (Supplementary Table S1D) includes 94 peptides expressed in a wide variety of normal tissues and overexpressed (high expression) in tumor cells; since a certain minimal amount of peptide is required for CD8+ T-cell recognition, the level of expression in normal tissues below that minimum may mean no autoimmune damage. However, this threshold is difficult to define.

2.2 Dataset of viral antigenic peptides

A number of viruses, such as the Epstein–Barr virus and the human papilloma virus, are associated with human malignancies, making peptides from viruses into potential immunotherapy antigens. We therefore included into the group E ‘Viruses’ 84 peptides from 47 viruses (Supplementary Table S1E) selected from the ProImmune Ltd database.

2.3 Protein abundance data

The analysis of epitope CR in multiple human tissues is based on experimental proteomics data. We obtained abundance data for human proteins from the PaxDB database (Wang *et al.*, 2015) (version 4.0) for 20 individual human tissues—brain, heart, lung, liver, kidney, prostate gland, pancreas, gall bladder, colon, esophagus, rectum, uterus, female gonad, testis, placenta, skin, plasma, platelet, saliva and urine. In addition, we used two further datasets: ‘whole organism’, containing data integrated over several whole-proteome measurements and ‘cell line’, containing data integrated over several proteomics experiments involving various cell lines. We sorted the list of tissues by descending importance for therapy survival. It is common knowledge that gross damage to brain, heart or lungs is life threatening. However, the exact sorting order is approximate, as there are no commonly accepted statistical data on tissue survival significance. Abundance values were expressed as parts per million (ppm), such that the sum of all protein abundances for each tissue is normalized to give a million.

2.4 A quantitative score for natural epitope presentation on MHC class I

An immunotherapeutic epitope can have CR with natural epitopes (NEs) presented by the major histocompatibility complex class I molecules (MHC-I) in various tissues. Such CR may arise if a protein normally expressed in cells is cleaved by the proteasome to produce a peptide that is close in amino acid sequence to the given epitope. We model this process by the method described by Keşmir *et al.* (2002). The quantitative score Q of epitope presentation on MHC-I is defined as:

$$Q = P_{\mathrm{CL}}/(A_{\mathrm{TAP}}{}^{*}A_{\mathrm{MHC}}), \qquad (1)$$

where P_{CL} is the proteasomal cleavage probability, while A_{TAP} and A_{MHC} are the IC_{50}-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

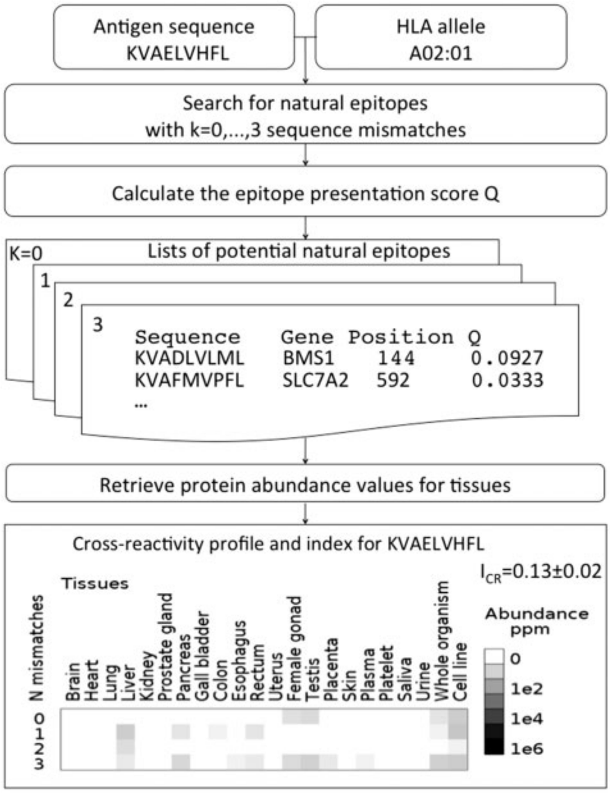

Fig. 1. Computational workflow for obtaining a CR profile $S(t,k)$ and an index I_{CR} from an input epitope sequence and its HLA-allele name. For details see *Methods*

affinities, as IC_{50} is defined as a dose of peptide that displaces 50% of a competitive ligand.

Segments of proteins in the database matching a given antigenic peptide with the allowed number of mismatches were considered as potential NEs. At first, the cleavage probabilities $P(n)$ were calculated for all positions n of the sequences of the NEs using NetChop 3.1 (Nielsen *et al.*, 2003), which employs a neural network to predict cleavage sites in the human proteasome. We used the network method 'C-term 3.0' and the threshold for the cleavage sites of 0.7. Then PCL was calculated as a product of $P(n\text{C})$ and all $(1-P(n))$: $P_{\text{CL}} = P(n\text{C})*\Pi(1-P(n))$, where nC is the C-terminal position of the peptide and n are all positions preceding the C-terminal position. Thus, the combined probability PCL represents the probability of being cleaved at C-terminus and not being cleaved at any other position. As suggested by the authors of NetChop, we do not take into account the possibility of cleavage at or before the N-terminal residue, as the molecular mechanism of such cleavage is more complex and not easily amenable to computational predictions.

To compute A_{TAP} we used a 20 × 9 consensus matrix obtained from experimental data by minimizing prediction errors (Peters *et al.*, 2003). The matrix contains log10(IC_{50}) affinities for all 20 amino acid types at every nonamer epitope position. For peptides of any length without precursors we use the accurate approach described here (Peters *et al.*, 2003), which involves summing the log-affinity for the C-terminal residue at position 9 and a 0.2-weighted sum of the log-affinities for the first three N-terminal amino acids. Raising the resulting sum to the power of 10 gives A_{TAP}, i.e. the IC_{50} value expressed in nM.

To predict the epitope affinity to MHC-I we used the program NetMHC 4.0 (Lundegaard *et al.*, 2008a,b), which is an artificial neural network trained on HLA alleles. For a given peptide sequence and an allele name (Supplementary Table S1A–S1E) the program returns the IC_{50} (nM) affinity, which is taken as A_{MHC} in the Equation (1).

2.5 Determination of CR scores

For each input epitope e we determined two quantitative measures of potential CR—its tissue profile $S(k,t)$ and its index I_{CR} (Fig. 1). The amino acid sequence of the epitope was matched against the RefSeq database (O'Leary *et al.*, 2016) of all naturally occurring human protein sequences, including all annotated isoforms, downloaded from the National Center for Biotechnology Information (NCBI). Between 0 and 3 mismatches k were allowed between a given epitope and any human peptide. This procedure yielded four lists of matched protein segments for $k = 0,1,2,3$, which we call here 'NEs'. Potential immunogenicity of each NE was calculated using the formula for epitope presentation score Q Equation (1). RefSeq (Pruitt *et al.*, 2012) transcript identifiers were mapped to STRING (Szklarczyk *et al.*, 2015) IDs based on the Biomart (Smedley *et al.*, 2009) mapping table. For each NE, abundance values of its parent protein expressed in ppm were obtained from PaxDB using STRING IDs.

The CR profile $S(k,t)$ for each of the input epitopes is a set of four ($k = 0,1,2,3$) arrays over the tissues $t = 1,T$ (i.e. a 4 × 22 matrix), where $T = 22$ is the number of tissues. The elements of the arrays are calculated as a sum of the abundances $a(i,t)$ in the tissue t, where i is the running index in the list of length $M(k)$ of matching NEs for each k. The sum only includes the abundances of the unique NEs that have the scores $Q(i)$ above the threshold of 0.02 (see below):

$$S(k,t) \;=\; \sum\nolimits_{i=1}^{M(k)} a(i,t). \tag{2}$$

Thus, large $S(k,t)$ values may indicate potential CR of the epitope e in the tissue t. The higher the number and abundances of different NEs that are close in sequence to a therapy peptide, the higher is the probability of cross-reaction. Higher thresholds for Q correspond to a higher probability of the selected NE 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 (Larsen *et al.*, 2005). In this work, we have chosen a threshold of 9% of sequence matches. The rationale for this choice was to ensure a low amount of false positives in the immunogenicity prediction, while the parameter of the number of mismatches controls the measure of closeness in sequence to the input immunogenic epitope.

To obtain a more concise measure of CR, we calculated the index I_{CR} as a weighted sum of the elements of the matrix $S(k,t)$, according to the following formula:

$$I_{\text{CR}} \;= \frac{1}{T}\sum\nolimits_{t=1}^{T}\sum\nolimits_{k=0}^{3}\log10[S(k,t)]*w(k), \tag{3}$$

where $k = 0,\ldots,3$, $t = 1,\ldots,T$ and $w(k) = (1/P(k))/\Sigma_k(1/P(k))$ are normalized weights, derived from the probability $P(k)$ of finding a random peptide of length l by matching with k-mismatches in our protein 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 weights are: $w(k = 1,2,3,4) = 0.95,\ 0.0475,\ 0.0023,\ 0.0002$. After taking log10 of the weighted sum of profiles $S(k,t)$ in ppm, summing over all tissues (for $S(k,t) > 1$), and normalizing by the total number of tissues T, the I_{CR} values span the range between 0 and 6.

The 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.

2.6 Determination of I_{CR}-threshold

To estimate a threshold value for the I_{CR} scores, above which toxicity would be expected, we used a clustering approach based on kernel density estimation (KDE), which is applicable to 1D data. KDE is a non-parametric way to estimate the probability density function of a random variable. We used the R implementation of the Gaussian KDE method (Botev *et al.*, 2010) to cluster the epitopes into groups based on their I_{CR} values without any assumptions or parameters.

3 Results and discussion

In this study we sought to obtain for the first time a comprehensive estimate of potential cross-reactions in human tissues for all MHC class I cancer immunotherapy antigens characterized so far, as well as for a representative selection of antigens from viruses. Antigens in the cancer immunotherapy dataset are subdivided into four groups dependent on their tumor-specificity. Although there is no reliable information on the actual use of these peptides in clinical practice and the quantitative degree of their CR measured in patients, our aim was to create a general approach for performing evaluation and validation of epitopes for potential cross-reactions at the tissue level.

3.1 Overview of the methodology

The complete process of obtaining CR measures is shown in Figure 1, with more details given in the Materials and Methods section. Using the sequence and the HLA allele name of each input antigen our software finds matching peptide segments in the database of human protein sequences, and calculates the quantitative score of the presentation on the MHC-I molecule. This step results in four tables for the number of mismatches $k = 0,\ldots,3$, containing NE sequences, their sequence positions and scores Q. Abundances of parent proteins in the human tissues are retrieved, unique entries are summed separately for each table, and plotted for each of the NEs in the form of a CR-profile. In addition, a CR-index is computed from the CR-profile, as described in the Methods. The CR-profiles thus comprise abundances of all matching peptides expressed in the respective tissues, which have high probability of being presented on MHC-I and therefore can potentially be reactive with the human CD8+ T cells. Since the sequences of the NEs are close to the given antigen sequence from the cancer database with demonstrated high immunogenicity, the matching natural peptides, particularly the top scoring ones, are also very likely to be immunogenic.

3.2 Epitope CR-profiles and CR-maps

Given that CR in non-target tissues can be invoked not only by exact peptides, but also by those with mismatching [e.g. one (Linette *et al.*, 2013) or four (Morgan *et al.*, 2013)] amino acids, in this study we allowed up to three mismatches between the therapeutic and native epitopes, and this parameter can be set to the desired value by the users of our server. The maps with four mismatches (data not shown) were qualitatively similar to the maps with three mismatches, albeit with somewhat higher intensities. Obviously, the number of matching native epitopes and thus the estimates of CR levels for the studied antigens quickly grow with the number of mismatches. The approach of including the cases where no exact hits were found, allows for a more comprehensive overview of potential CR.

The CR-profiles are computed according to Equation (2) separately for each epitope listed in the Supplementary Table S1A–S1E. The set of profiles for a list of peptides corresponding to a certain tumor specificity group constitutes a CR-map. Altogether we analyzed 342 epitopes in five groups: A—'Mutation', B—'Tumor-specific' (CT), C—'Differentiation', D—'Overexpressed' and E—'Viruses' (see Methods), and plotted the results in the form of five individual CR-maps (Fig. 2A–C and Supplementary Fig. S2D and S2E). In each figure the four CR-profiles for $k = 0,1,2,3$ are shown, with 'k' being the allowed number of mismatches between the amino-acid sequence of a given epitope and all matching proteins expressed in the human tissues. In each map, the 22 human tissues are listed from left to right in the order of diminishing importance for immunotherapy survival, but the importance is not factored into the calculations. The vertical cell numbers correspond (top–down) to the sequential numbers of antigens in the corresponding table. Thus, each cell in a CR-map contains log10 of the CR value $S(k,t)$ [calculated according to Equation (2) as the sum of abundances of the proteins harboring the epitope e in a particular tissue t with k mismatches] and with the combined score Q above a threshold of 0. 02, i.e. top-scored for proteasomal cleavage, TAP transport and MHC-I binding. This threshold is a user-defined parameter in the method, and 0.02 corresponds to selecting the top-scoring 9% of the matching peptides. The rationale for this choice was to ensure a low amount of false positives in the immunogenicity prediction, while the parameter of the number of mismatches controls the measure of closeness in sequence to the input immunogenic epitope. The median of Q-values of NEs found for all epitopes (all k, all tissues) was $1.8{*}10^{-4}$, while the average was 0.10. The log10-values of the CR-profiles for the five groups and for each $k = 0,1,2,3$ are given in Supplementary Table S2A–S2E. The range of log10-values from 0 to 6 (i.e. the range of ppm values from 0 to 10^6) was converted for plotting linearly to the 0–255 range of gray scale intensity values. Consequently, a larger intensity of a square corresponds to a larger sum of the abundances (more matching proteins and/or higher abundance).

3.3 Interpretation of the maps

We split our database of antigens under study into five groups, with the aim to uncover common patterns of CR in the tissues within the groups. Indeed, we found a significant variation between the groups with no mismatches allowed ($k = 0$), while for $k > 0$ the differences between the groups were less clear-cut, with varying CR levels in multiple tissues. As expected, the antigens from the group A 'Mutation' showed no CR at all (Fig. 2A), when matched exactly, as the antigens of this group have resulted from mutations of original genes. With 1–2 mismatches allowed, one of the antigens (20: SLFEGIDIYT) from the heat shock protein (HSP70.2) exhibited particularly strong CR and further four peptides had medium values across all tissues. Considering all mismatch numbers, a certain level of CR can be expected for about a third of them, if the assumption that CTL cells would exhibit CR to inexact peptides, differing by 1–3 amino acids, is correct.

For nine peptides of the Group B 'Tumor-specific' or CT antigens (Fig. 2B) there was only a low level of CR in testis, placenta and liver based on exact matches ($k = 0$). In theory these epitopes would be expected to be exclusively expressed in testis and nowhere else. Since there are no MHC-I molecules in the testis cells, no CTL response can occur there. For further four of the peptides this

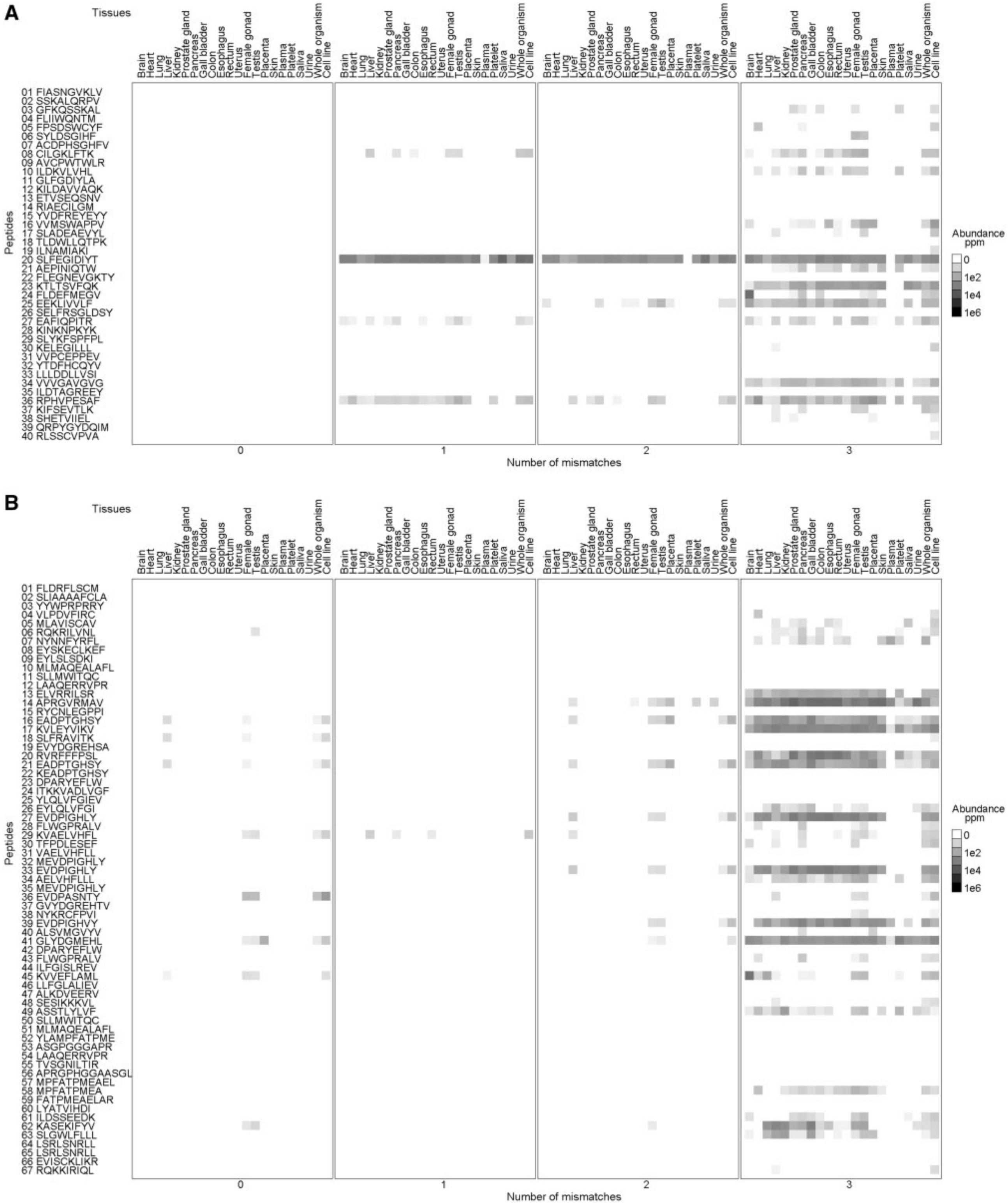

Fig. 2. (A) The CR-map for the 'Mutation' antigens (Group A) peptide sequences are given on the vertical axis. On the horizontal axis the number of sequence mismatches (bottom) and the names of 22 tissues (top) are shown. Intensity range from 0 (white) to 255 (black) corresponds to protein abundances in the range from 0 to 1 × 10^6 ppm. (B) The CR-map for the 'Tumor-specific' (CT) antigens (Group B)

pattern shows up at $k = 2$, while other peptides show no CR for up to two mismatches. Thus, the majority of the peptides in this group can be considered as potentially safe, meaning that for most of them no immunotherapeutic side effects would be expected in any tissues. The only exception is constituted by about 10 peptides showing medium level values for the tissues other than CT and for $k = 0,2$. Some epitopes also show a medium CR level based on the protein abundance levels for the whole organism and for cell lines.

Similarly, in the groups C 'Differentiation' and D 'Overexpressed' (Supplementary Fig. S2C and S2D) about 10% of the epitopes shows weak levels and a few epitopes show strong CR values, while approximately 90% of the peptides had no hits at all for $k = 0,1,2$.

Finally, and quite surprisingly, the epitopes of the group E 'Viruses' (Fig. 2E) display no CR for $k = 0,1$, while about 10% of them show weak to medium values for $k = 2$. This low CR level might be due to the genuine differences between the human and viral proteomes, which may imply a certain positive potential for safe immunotherapy or vaccination against common viruses.

3.4 Epitope CR-index

To obtain a 'compressed' measure of CR, suitable for comparing CR levels of various epitopes, we propose a novel I_{CR} index Equation (3), which is basically a weighted double sum of the CR-profiles overall numbers of mismatches and all tissues. The advantage of this index is that it provides a stable value that will only change slightly if the list of tissues considered or the abundance values get updated. Due to the weighting Equation (3) being biased toward smaller numbers of mismatches, the index will change only slightly when varying the allowed number of mismatches: e.g. for KVAELVHFL the I_{CR} equals 0.128 ± 0.015, 0.136 ± 0.016, 0.136 ± 0.022, 0.136 ± 0.021 for $k = 0,1,2,3$, respectively. In contrast to CR-profiles, which are meant to characterize CR in individual tissues, whole organisms or cell lines, the CR-index gives the overall measure. Note that the index can be non-zero even when there are no exact matches found for input epitope, as it also takes into account not-exact matches.

Overall, I_{CR} allows for a quick comparison of epitopes within and between groups, and gives a concise indication of the epitope's potential for CR in the entire human organism. The I_{CR} values, ranging from 0 to 6, are meant to be easily interpretable: 0—no CR and 6—cross-reacting to all proteins in all tissues.

We calculated I_{CR} for the epitopes belonging to our five datasets (Fig. 3). All groups A–E have epitopes with CR levels below 1.0 (equivalent to the total tissue abundance of below 10 ppm). The lowest overall values are for the groups A—'Mutation' below 0.15 and E—'Viruses' below 0.003. The highest overall values are for the groups C—'Differentiation' and D—'Overexpressed'. The epitopes thus exhibit differing degrees of variation with regard to I_{CR} within the groups. The index can be used in practical applications to rank variants of a therapy by their overall CR, to (in)validate a particular epitope, to compare the values from different therapies and to correlate them with clinical data.

3.5 Validation of the methodology

Due to the lack of systematic and comprehensive quantitative data on CR effects in clinical studies the development of a rigorously benchmarked CR prediction algorithm is currently not feasible. Here we propose a bioinformatics tool that, while not being able to predict the actual toxicity level, gives a quantitative indication of the spectrum of possible off-target effects based on experimental proteomics data. Namely, it provides information about tissue-specific expression of immunogenic epitopes from the normal human protein repertoire that are sequence-similar to therapeutic epitopes. In many cases, a low level of off-target peptide expression in normal tissues and suppressed autoimmunity to self-antigens may mean that the immune response to a therapy peptide is not triggered. However, when the immune response is not suppressed, it may depend on the peptide expression level, since a certain minimal amount of peptide is required for CTL recognition (Van den Eynde and van der Bruggen, 1997; Vigneron *et al.*, 2013). The toxicity threshold is difficult to define, as it depends on the normal gene expression level for each cell type (Vigneron *et al.*, 2013), but in any case it is very low. The study of Schodin *et al.* (1996) showed that lysis of target cells occurs for concentrations of peptides in the range 10^{-11} to $3*10^{-55}$ M. For the smallest concentration a large number of unbound TCRs per cell were required ($\geq 50\,000$, i.e. nearly the entire TCR complement), while for the highest concentration only 1000 TCRs were needed. Thus, from the point of view of off-target toxicity, even a very small amount of peptide can be problematic, if it can bind strongly to the therapeutic CTLs, while the presence of multiple peptides of this kind increases the probability of lysis. At the same time, if there are no peptides matching a given therapeutic peptide, even at high mismatch numbers, off-target toxicity by the CTL will be highly unlikely, simply because such peptides will not be presented in the normal human tissues. This is the rationale behind our approach.

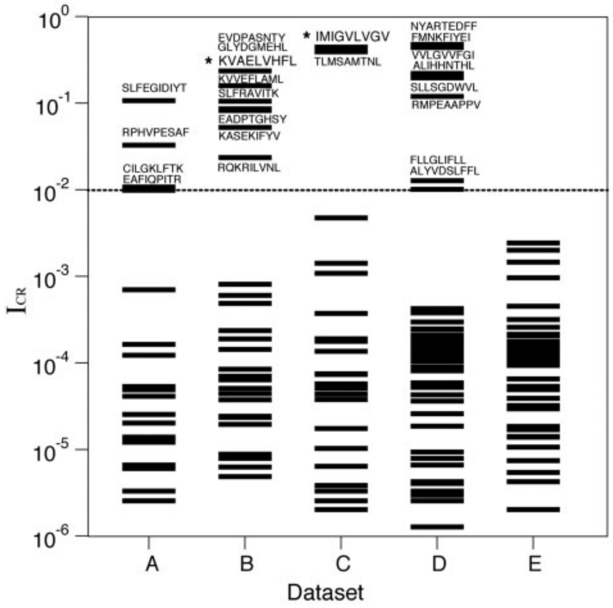

Fig. 3. CR indices I_{CR} of the individual peptides from the five datasets (Groups A–E) considered in this work. The index values correspond to thin dashes plotted on a logarithmic scale vertically. Symbol '*' denotes the epitopes (KVAELVHFL, IMIGVLVGV) that exhibited high toxicity in clinical studies (Table 1). Sequences of the epitopes with the CR-indices above the CR-threshold of 10^{-2} (dash line) are shown

We were able to find four studies that provide qualitative description of immunotherapy toxicity effects for the total of eight peptides, of which two invoked toxicity and six did not (Table 1). In the first study, the peptide KVAELVHFL, which perfectly matches ($k = 0$) melanoma associated antigens MAGEA3 and MAGEA9 with the Q-scores 0.07 and 0.03, respectively, caused severe and lethal neurologic toxicity because of the unexpected expression of other members of the MAGE CT family in the central nervous system (Morgan *et al.*, 2013). The TCR used in this study recognized epitopes in MAGE-A3/A9/A12. The study indicates that expression of MAGE-A12 in human brain, which was previously unrecognized, possibly initiated TCR-mediated inflammatory response that resulted in neuronal cell destruction. For this peptide iCrossr returned the CR-index (I_{CR}) of 0.136 ± 0.02 ($k = 0$–3). Additionally, it identified six matching peptide sequences with k in the range of 1–3 and Q ranging between 0.02 and 5.29 (Table 1), including the closely matching ($k = 1$) peptide from MAGEA12 with a high Q score of

Table 1. Comparison of the results from published clinical studies and the I_{CR} values for several cancer immunotherapy epitopes, binding to HLA-A02:01

Study	Therapy peptide	Gene/position	Off-target effects	Matching natural peptides	Gene	No. of mismatches	Q	I_{CR}
Morgan *et al.* (2013)	KVAELVHFL	MAGE-A3 112-120	Neurologic toxicity/lethal	KVAELVHFL	MAGEA3	0	0.07	0.13 ± 0.02
				KVAELVHFL	MAGEA9	0	0.03	
				KMAELVHFL	MAGEA12	1	0.92	
				KVAELVHIL	DDX28	1	0.13	
				KMVELVHFL	MAGEA2B	2	5.29	
				KVADLVLML	BMC1	3	0.09	
				KVAFMVPFL	SLC7A2	3	0.03	
				KVIILVHYL	MAGEB10	3	0.017	
Parkhurst *et al.* (2011)	IMIGVLVGV	CEA 691-699	Severe colitis	IMIGVLVGV	CEACAM5	0	0.60	0.57 ± 0.03
				IMIGVLARV	CEACAM8	2	0.12	
				IMIGVLAGM	CEACAM7	2	0.12	
				ALIGVLLG	VSIG2	3	0.69	
				IMAGQLVAV	ABCC2	3	0.05	
Parkhurst *et al.* (2011)	LLGRNSFEV*	p53 264-272	None	GLFRNRFEV	TRMU	3	0.11	3e-5 ± 3e-6
	FLPSDYFPSV*	HBVc 23Y/18-27	None	LLMRNNFEY	DOCK11	3	0.11	0
	YLEPGPVTA*	gp100 280-288	None	ILPGNSFEV	TDRD15	3	0.03	0
Rosenberg (2010)	VSLLMWITQV	NY-ESO-1 157-165	None	None	—	—	—	0
Johnson *et al.* (2009)	AAGIGILTV	MART-1 27-35	None	ALVIGILVV	AQP7	3	0.03	1e-5 ± 2e-6
	KTWGQYWQV	gp100 154-162	None	None	—	—		0

*Negative control peptides.

0.92. The visualization of the abundances of the peptide ‘29’ in multiple tissues can be seen in Figures 1 and 2B. The peptides KVAFMVPFL (SLC7A2) and KVIILVHYL (MAGEB10) had a small abundance in the tissue ‘Brain’—0.01 and 0.09 ppm, respectively (not visible on the plot). Thus, in this case iCrossR reported a large overall I_{CR} for the therapy peptide and correctly identified the expression of multiple matching high- and low-scoring epitopes, which are likely the cause of the off-target toxicity observed in the patients.

In the second study (Parkhurst *et al.*, 2011), the instances of on-target off-tumor toxicity have been reported for the peptide IMIGVLVGV in all three patients, resulting in severe transient inflammatory colitis. The study concluded that observations of transient mucosal destruction by carcinoembryonic antigen (CEA)-reactive T cells represent an autoimmune colitis probably due to lymphocyte recognition of the normal levels of CEA genes expression in colonic mucosa. For this peptide iCrossR gave a very large I_{CR} of 0.571 ± 0.028 (Fig. 3, top) and matched ($k = 0$–3) to five peptides with Q values in the range 0.05–0.69. The study used three negative control peptides that did not bind TCR and gave no toxicity. They exhibit I_{CR} equal or close to 0, and no close matches to natural peptides [although there are three distantly matching ($k = 3$) peptides for LLGRNSFEV]. Since no toxicity was reported for the control peptides, the relatively high Q scores for the distant matches (ranging between 0.03 and 0.11) indicate their possible binding to HLA; but there was no binding to TCR, which is a pre-requisite for CTL-mediated toxicity. Thus, in this case as well, toxicity is presumed to be caused by off-target effects to the genes expressed normally, in this case, in colon tissues, where the therapy peptide closely matched ($k = 0$) the gene CEACAM5. iCrossR correctly identified the peptide with a very high score of $Q = 0.60$ and reported its high expression level in multiple tissues. Notably, the tissues with the highest expression levels—Colon, Esophagus, Rectum and Saliva—with the protein abundances of 66.7, 59.8, 41.3 and 44.4 ppm, respectively, contain mucosal cells.

Next, we discuss the examples of therapies that had no reported toxicities. The peptide VSLLMWITQV from another CT antigen (NY ESO-1) was reported to be safe and demonstrated clinically evident antitumor responses; there were no treatment-related deaths in any of the 106 patients (Rosenberg, 2010). Similarly, for the two peptides from another study (Johnson *et al.*, 2009) the following clinical course of reactivity against normal tissues was established: 29 of the 36 patients in the trial exhibited a widespread erythematous skin rash, which gradually subsided over several days without treatment. Because melanocytic cells expressing MART-1 and gp100 genes exist in the eye, some of the patients developed an anterior uveitis, but in all patients ocular findings reverted to normal; no deaths were reported. iCrossr found no matching NEs for KTWGQYWQV ($I_{\mathrm{CR}} = 0$). For AAGIGILTV one distantly matching ($k = 3$) NE was identified expressed at small levels in six tissues, with a presentation score of $Q = 0.03$, but I_{CR} close to 0. Thus, in these two cases I_{CR} values close to zero were in line with the lack of toxicity.

Taken together, based on the above results, we can propose the following conclusion: there is a risk for a therapy epitope to exhibit side-effects in normal tissues, if all four conditions are met: it has close sequence matches to the normal genes ($k = 0$–1), the matching protein abundances are above a certain low threshold, the matching peptides have high scores of HLA-presentation, and the peptides can bind to a TCR of CTLs. If one of the conditions is not met, the off-target effects are not likely. By comparing the CR-maps (Fig. 2A–C and Supplementary Fig. S2D–S2E) and the I_{CR} values (Fig. 3) it becomes evident that all peptides showing multiple hits on the maps with small mismatch numbers ($k = 0,1$) also occupy top positions in Figure 3 (high I_{CR} values). This applies to the peptides SLFEGIDIYT (‘20’) and RPHVPESAF (‘36’) (Fig. 2A, $k = 1$ and Fig. 3, group A) as well as to the peptides IMIGVLVGV (‘02’) and TLMSAMTNL (‘36’) (Supplementary Fig. S2C, $k = 0$ and Fig. 3, group C).

We merged the I_{CR} data of the five groups (Supplementary Table S1A–S1E, column ‘I_{CR}’) into a single 1D array, excluding zero values, and applied the KDE function to the array (see Methods). The resulting density profile of the I_{CR} scores (Supplementary Fig. S1) shows a simple pattern of two hills with a ‘valley’ between them. This can be interpreted that the I_{CR} data can be separated at around 10^{-2} into two groups of high and low I_{CR} values (shown by the horizontal line in Fig. 3). All epitopes shown in Table 1 causing off-target toxicity have large I_{CR} values and belong to the high value group, while the control peptides and all epitopes without off-target effects have zero or small I_{CR} values and thus belong to the low value group. Consequently, based on the very scarce clinical evidence summarized in Table 1, we can propose a tentative I_{CR} threshold of 10^{-2}, although using a higher value and carefully analyzing tissue distribution of the underlying abundance values would be safer.

4 Conclusions

We present a new methodology to compute two quantitative measures of CR for MHC-I epitopes—the CR-profile, reflecting the distribution of CR across tissues and the CR-index, giving an overall measure. The former measure is better suited for visual analysis while the latter one allows for comparison and ranking of epitopes.

In this work we exclusively focused on assessing the antigens binding to the MHC-I molecules, as this mechanism of immune response involves natural peptides originating inside the cell. In the future we intend to adapt our approach to assess exogenous peptides presented by the MHC-II complex that binds T helper cells.

One limitation of our model is that it does not consider data on CTL binding, which would significantly reduce the list of possible cross-reactive natural peptides. Computational approaches to predict CTL binding are currently lacking, but could be easily incorporated in our procedure once they become available.

We suggest using these results as a filter for early stages of epitope selection, and with a considerable degree of precaution when used at later stages of therapy development. To conclude, our protocol is aimed to help in development of better immunotherapies, to improve safety, and to avoid or minimize undesirable autoimmune side effects.

Conflict of Interest: none declared.

References

Botev,Z.I. *et al*. (2010) Kernel density estimation via diffusion. *Ann. Stat*., **38**, 2916–2957.

Coulie,P.G. *et al*. (2014) Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat. Rev. Cancer*, **14**, 135–146.

Frank,S.A. (2002) *Immunology and Evolution of Infectious Disease*. Princeton, NJ, Princeton University Press.

Johnson,L.A. *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.

Keşmir,C. *et al*. (2002) Prediction of proteasome cleavage motifs by neural networks. *Prot. Eng*., **15**, 287–296.

Kohm,A.P. *et al*. (2003) Mimicking the way to autoimmunity: an evolving theory of sequence and structural homology. *Trends Microbiol*., **11**, 101–105.

Larsen,M.V. *et al*. (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol*., **35**, 2295–2303.

Linette,G.P. *et al*. (2013) Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood*, **122**, 863–871.

Lundegaard,C. *et al*. (2008a) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*., **36**, W509–W512.

Lundegaard,C. *et al*. (2008b) 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.

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

Maus,M.V. *et al*. (2014) Adoptive immunotherapy for cancer or viruses. *Annu. Rev. Immunol*., **32**, 189–225.

Morgan,R.A. *et al*. (2013) Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immunother*., **36**, 133–151.

Nielsen,M. *et al*. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Prot. Sci*., **12**, 1007–1017.

O'Leary,N.A. *et al*. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*., **44**, D733–D745.

Parkhurst,M.R. *et al*. (2011) T Cells targeting carcinoembryonic antigen can mediate regression of metastatic colorectal cancer but induce severe transient colitis. *Mol. Ther*., **19**, 620–626.

Peters,B. *et al*. (2003) Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol*., **171**, 1741–1749.

Pruitt,K.D. *et al*. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*., **40**, D130–D135.

Rosenberg,S.A. (2010) Of mice, not men: no evidence for graft-versus-host disease in humans receiving T-cell receptor-transduced autologous T cells. *Mol. Ther*., **18**, 1744–1745.

Schodin,B.A. *et al*. (1996). Correlation between the number of T cell receptors required for T cell activation and TCR-ligand affinity. *Immunity*, **5**, 137–146.

Smedley,D. *et al*. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Szklarczyk,D. *et al*. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., **43**, D447–D452.

Van den Eynde,B.J., and van der Bruggen,P. (1997) T cell defined tumor antigens. *Curr. Opin. Immunol*., **9**, 684–693.

Vigneron,N. *et al*. (2013) Database of T cell-defined human tumor antigens: the 2013 update. *Cancer Immun*., **13**, 15.

Wang,M. *et al*. (2015) Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**, 3163–3168.

Weber,J.S. *et al*. (2015) Toxicities of immunotherapy for the practitioner. *J. Clin. Oncol*., **33**, 2092–2099.

Zang,Y.W. *et al*. (2014) Clinical application of adoptive T cell therapy in solid tumors. *Med. Sci. Monit*., **20**, 953–959.