

EpiToolKit—a web server for computational immunomics

Magdalena Feldhahn^{1,*}, Philipp Thiel¹, Mathias M. Schuler², Nina Hillen², Stefan Stevanović², Hans-Georg Rammensee² and Oliver Kohlbacher¹

¹Simulation of Biological Systems, Center for Bioinformatics and ²Department of Immunology, Institute for Cell Biology, Eberhard Karls University Tübingen, Germany

Received January 31, 2008; Revised April 4, 2008; Accepted April 11, 2008

ABSTRACT

Predicting the T-cell-mediated immune response is an important task in vaccine design and thus one of the key problems in computational immunomics. Various methods have been developed during the last decade and are available online. We present EpiToolKit, a web server that has been specifically designed to offer a problem-solving environment for computational immunomics. EpiToolKit offers a variety of different prediction methods for major histocompatibility complex class I and II ligands as well as minor histocompatibility antigens. These predictions are embedded in a user-friendly interface allowing refining, editing and constraining the searches conveniently. We illustrate the value of the approach with a set of novel tumor-associated peptides. EpiToolKit is available online at www.epitoolkit.org.

INTRODUCTION

Prediction of T-cell epitopes is a key problem in Immunoinformatics (1). Identifying peptides with high binding affinity to major histocompatibility complex (MHC) molecules is generally considered the best way to predict epitopes (2). Many different methods and tools for peptide–MHC binding have been developed in the recent past, most of them are accessible online (3–12). Most of the individual web servers, however, were designed to quickly offer online access to a newly developed method with usability not being the major concern. Nonexpert users may be overwhelmed by the variety of layouts, formats and options.

The main focus in the development of the EpiToolKit web server was usability. The purpose of EpiToolKit is to facilitate immunological research by providing a consistent and user-friendly interface for different methods

from computational immunomics. The prediction pipeline is organized in four main steps: sequence input, sequence information, model selection and display of prediction results. Each page contains hints and short comments to guide the user through the pipeline. In addition, a detailed help and documentation is available through direct links on every page.

The service provides most of the commonly accepted prediction methods and allows their simultaneous application. Thereby, prediction results can be compared without the need to access the individual web-based services separately. The combination of different prediction methods also increases the number of available allelic models.

In addition to epitope prediction, EpiToolKit offers the functionality to examine the influence of sequence polymorphisms or mutations on potential T-cell epitopes. This feature is useful for the identification of minor histocompatibility antigens (mHags) and for the development of peptide-based vaccines against highly variable pathogens such as HCV and HIV.

EpiToolKit is based on a flexible and modular framework for predictions related to epitope prediction (13). The framework and EpiToolKit can easily be extended with new methods—e.g. new methods for MHC binding, or for the prediction of the epitope processing pathway.

WEB INTERFACE

The web interface is divided into two parts, the epitope prediction and the prediction on polymorphic proteins (SNEPv2). Both predictions use the same layout and are based on a common pipeline. To facilitate the use of EpiToolKit, the prediction pipeline has been broken down into four intuitive steps:

- (1) Sequence input. Sequences can be retrieved from the most important resources for protein sequences, namely Swiss-Prot (14) and NCBI RefSeq (15). In case of polymorphic prediction, all reported

*To whom correspondence should be addressed. Tel: +49 7071 29 70460; Fax: +49 7071 29 5152; Email: feldhahn@informatik.uni-tuebingen.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

polymorphisms for the specified sequences are provided automatically. Alternatively, the user can paste sequences directly or upload a FASTA file. In case of user-defined polymorphic or mutated sequences, the changes can be specified in the FASTA header.

- (2) Sequence information. The purpose of this step is to present the query results for the requested sequences. For all returned sequences, additional information such as sequence length, GeneID or RefSeq accession is displayed. Furthermore, SNEPv2 returns all annotated polymorphisms for each sequence. The sequences as well as the polymorphisms can be selected or deselected individually for further processing.
- (3) Allele selection. In this step, the user can select allelic models for prediction. The model selection is organized in an expandable model/allele tree, sorted by allele name. The tree can be customized to contain only models for selected peptide lengths, prediction methods or alleles in the advanced options section.
- (4) Prediction results. Results are displayed as tables. For epitope prediction, a single table is created for every peptide length. In case of polymorphic predictions using SNEPv2 for each polymorphism a separate table is created. Different methods for discrimination between predicted binders and nonbinders are available. Filter and display options can be changed in the advanced options section. Additionally, the prediction results can be exported in CSV (comma separated values) or XLS (Microsoft Excel) format.

DATASETS

EpiToolKit provides access to the Swiss-Prot database (14) and the NCBI RefSeq database (15). To improve the reliability of the service and to accelerate access time, both databases are kept as local copies. They are updated monthly. Current release information is displayed on the sequence input site.

Polymorphism information is provided from two different sources depending on the database used for sequence retrieval. For RefSeq sequences, polymorphisms are obtained from dbSNP (16). SNP entries contain links to protein sequences if the corresponding variation has been mapped onto coding regions. For requested RefSeq sequences, nonsynonymous coding SNPs are yielded by EpiToolKit. Furthermore, to limit the search to relevant polymorphisms the set of reported polymorphisms can be restricted to a user-defined heterozygosity range. Polymorphisms for Swiss-Prot sequences are extracted directly from the Swiss-Prot entries. Polymorphism data from dbSNP as well as from Swiss-Prot are also kept in local databases.

IMPLEMENTATION

This section briefly describes the prediction methods included in EpiToolKit and the implementation of the web server.

Prediction on polymorphic proteins

In addition to epitope prediction EpiToolKit provides the possibility to perform predictions on proteins containing polymorphisms. This tool is called SNEPv2. It is a new version of the SNP-derived Epitope Prediction program (SNEP) developed by Schuler *et al.* (11). One major improvement of SNEPv2 is the use of a second resource to retrieve sequence polymorphisms [dbSNP (16)]. Additionally, due to the embedding into EpiToolKit, SNEPv2 offers a comfortable user interface. The latter enables users to perform predictions for multiple polymorphic sequences from several sources, to apply a variety of prediction methods and to use different filtering methods to restrict the result set.

SNEPv2 creates a set of polymorphic peptides for each reported sequence polymorphism. These peptide sets are generated by extracting the peptides around the polymorphic position using a sliding window of specified length and subsequently mutating these peptides to all observed variants. The prediction results are displayed separately for each peptide set. On the result page, the user can additionally switch between the epitope prediction results for the source protein and the polymorphic predictions.

Epitope prediction methods

A major challenge in epitope prediction is that the manifold MHC alleles display a wide spectrum of binding specificities. Prediction methods must therefore provide models for different alleles. Not all prediction methods have a model for every allele or peptide length.

Five different methods for the prediction of peptides binding to MHC class I and two methods for MHC class II binding are currently available in EpiToolKit. These methods are described briefly in the following. For details on the prediction methods refer to the original publications. A table of all available allelic models and methods can be found in Supplementary Material 1. The description of the available methods in the documentation of EpiToolKit will be updated when new methods are included into EpiToolKit.

- SYFPEITHI (3) is based on position-specific scoring matrices. The matrices are manually generated based on expert knowledge and the occurrence of amino acids in naturally processed MHC ligands from the SYFPEITHI database.
- BIMAS/HLA_BIND (4) was developed at the BioInformatics and Molecular Analysis Section (BIMAS) at the NIH. The prediction method uses position specific scoring matrices that are derived from experimentally determined relative binding affinities. Dissociation rates of peptide:MHC: β_2 -microglobulin complexes are used to measure binding affinities relative to a reference peptide. The original values in the matrices are log-transformed to obtain an additive scoring scheme.
- SVMHC (5) uses support vector machine classification to predict MHC-binding peptides. The method is trained on known MHC-binding peptides from

the SYFPEITHI database and randomly generated nonbinders.

- Epidemix (13) is based on position-specific scoring matrices. The matrices are statistically computed based on the positive training set of SVMHC. Sequence weighting and pseudo-count correction are applied to obtain the frequencies used to generate the matrices.
- UniTope (Feldhahn, Toussaint, Ziehm, and Kohlbacher, manuscript in preparation) is a support vector classification method recently developed in our group. UniTope combines structural and sequence information in a machine-learning framework. Based on a decomposition of the MHC-binding groove into distinct pockets, the correlation between the physico-chemical properties of these pockets and peptide binding is learned. The allele encoding uses pocket profiles derived from crystal structures of peptide:MHC complexes. The peptides are also encoded using physico-chemical properties. This enables binding prediction even for alleles where no experimental binding data are available.
- Hammer (17) is based on position-specific scoring matrices and predicts binding peptides for MHC class II. The virtual matrices were published by Sturniolo *et al.* (17) and are used by the TEPITOPE software.
- MHCIIIMulti (18) is a new method based on multiple instance learning and support vector classification, which can be used to predict peptide binding for MHC class II. Therefore, a new kernel is introduced, which also takes similarities between alleles into account. The method can even be used to predict binding peptides for alleles without available binding data.

Filtering methods

Most prediction methods predict a score that represents the binding affinity of a peptide to an MHC allele. In order to discriminate binding peptides from nonbinding peptides, a threshold can be used to separate binders from nonbinders.

EpiToolKit uses thresholds to filter the results for potential epitopes. Only peptides classified as binders are displayed if a filter is activated. The following filter options are available:

- (1) No filtering. All predicted scores are displayed. The peptides are not classified as binders or nonbinders.
- (2) Filtering using *halfmax*-scores. For all matrix-based methods, the *halfmax*-score is defined as half of the maximal value obtainable from the matrix. The *halfmax*-scores are used as thresholds. For SVM-based predictions *halfmax*-scores are not defined. The 2%-thresholds are used instead (see 'Filtering by percentage' subsequently). The *halfmax* filter is the default filter method in EpiToolKit.
- (3) Filtering by percentage. The thresholds are determined based on a background score distribution that was computed on a large set of peptides derived from natural proteins. The thresholds can be interpreted as follows: using as example a 2%-threshold 2% of the peptides used to compute the background

distribution would be classified as binders. The advantage of this filtering method is that—in contrast to the *halfmax*-filtering—thresholds for different allelic models are comparable. The input format for the percentage thresholds is a float in the interval of [0, 1], e.g. 0.02 for a 2%-threshold.

Note that the current version of UniTope performs a binary classification. For consistency, all filter methods are also available for UniTope predictions but do not have an influence on the classification—peptides with score 1 are always classified as binders, whereas peptides with score 0 are always classified as nonbinders.

Web server

EpiToolKit utilizes the open source content management framework Plone (<http://plone.org>) based on the application server Zope (<http://www.zope.org>). Dynamic HTML pages provide forms for user input and prediction results. The pages use CSS and JavaScript and the service was tested for compatibility with the two most widely used web browsers, namely Mozilla Firefox (version 2.0) and MS Internet Explorer (version 7). Input data is pre-processed by Python scripts for data validation. Entered user data are temporarily stored on the server and automatically deleted after the corresponding session has expired. The program logic for data processing and for performing epitope prediction (13) is written in Python. Several tasks of the program logic use the Biopython package (<http://biopython.org>).

VALIDATION AND APPLICATION

To validate the correct function of EpiToolKit, the prediction results for 1000 randomly chosen proteins were compared to the results of the original methods/web servers. In all cases EpiToolKit produced the same results as the original methods.

In addition, a set of 27 novel HLA-A*0201 ligands characterized from tumor cells and in part derived from tumor antigens were used for validation. These peptides are listed in Supplementary Material 2. All available allelic models of appropriate length were used for prediction and default filtering (*halfmax*) was applied. Eighteen out of 20 nonameric peptides and six out of seven decameric peptides were correctly classified as binders.

To validate SNEPv2, we tried to retrospectively identify the mHags reported by Goulmy (19). The paper gives an introduction to the clinical relevance of mHags and contains a list of all then known human mHags. We refer to the single mHags with the names used in (19). The dataset is available in Supplementary Material 3.

Eight out of nine autosomally encoded mHags reported could directly be found using SNEPv2. The HLA-A*2902 restricted mHag UGT2B17 results from a gene-deletion and therefore no allelic counterpart is available, so epitope prediction was used. Since no decameric model was available for HLA-A*2902, the nonameric UniTope model was used to score the two nonameric substrings. Both substrings were predicted to bind to HLA-A*2902.

The Y-chromosomal encoded mHags do not have allelic counterparts, except for the HLA-A*01 restricted mHag DFFRY. Both versions of DFFRY were predicted to bind to HLA-A*0101. For all other Y-chromosomal mHags, epitope prediction was used. RPS4Y/HLA-DRB3*0301 was correctly classified as binder. No allelic model was available to predict the mHag DBY/HLA-DQ. For two gonosomal mHags (UTY/HLA-B60 and SMCY/HLA-B7), no model of the appropriate length was available, so all related models of shorter length were used. For both mHags, at least one nonameric substring was predicted to bind by the respective model. The remaining three mHags were correctly predicted to bind to the restricting allele.

These examples demonstrate the usefulness and applicability of EpiToolKit for the identification of potential epitopes and mHags.

CONCLUSION

As a convenient and user-friendly program, EpiToolKit enables the direct comparison of different epitope prediction software packages and thus allows for precise selection of HLA-presented peptides from any protein of choice. In addition, its unique feature of screening polymorphic proteins for HLA ligands with SNEPv2 will promote and facilitate the identification of mHags, including tissue-specific peptides, as well as epitopes from quickly mutating pathogens. We expect that EpiToolKit with its epitope prediction and in particular with the SNEPv2 screening will provide valuable support in the future identification of T cell epitopes of major clinical relevance.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

This work was supported by Deutsche Forschungsgemeinschaft (SFB 685). Funding to pay the Open Access publication charges for this article was provided by Deutsche Forschungsgemeinschaft (SFB685).

Conflict of interest statement. None declared.

REFERENCES

- DeLuca, D.S. and Blasczyk, R. (2007) The immunoinformatics of cancer immunotherapy. *Tissue Antigens*, **70**, 265–271.
- Rötzschke, O., Falk, K., Stevanović, S., Jung, G., Walden, P. and Rammensee, H.G. (1991) Exact prediction of a natural T cell epitope. *Eur. J. Immunol.*, **21**, 2891–2894.
- 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.
- 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.
- Dönnes, P. and Kohlbacher, O. (2006) SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res.*, **34**, W194–W197.
- Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S.L., Lamberth, K., Buus, S., Brunak, S. and Lund, O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.
- Peters, B. and Sette, A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinform.*, **6**, 132.
- Honeyman, M.C., Brusic, V., Stone, N.L. and Harrison, L.C. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, **16**, 966–969.
- Bui, H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K., Mothé, B.R., Chisari, F.V., Watkins, D.I. and Sette, A. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, **57**, 304–314.
- Peters, B., Bui, H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W. *et al.* (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.*, **2**, e65.
- Schuler, M.M., Dönnes, P., Nastke, M.D., Kohlbacher, O., Rammensee, H.G. and Stevanović, S. (2005) SNEP: SNP-derived epitope prediction program for minor H antigens. *Immunogenetics*, **57**, 816–820.
- Halling-Brown, M., Quartey-Papafio, R., Travers, P.J. and Moss, D.S. (2006) SiPep: a system for the prediction of tissue-specific minor histocompatibility antigens. *Int. J. Immunogenet.*, **33**, 289–295.
- Feldhahn, M. (2006) FRED: a framework for T-cell epitope detection. Diploma thesis, University of Tuebingen, Tuebingen, Germany.
- Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
- The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D65.
- Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
- Sturniolo, T., Bono, E., Ding, J., Radrizzani, L., Tuercio, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M.P., Sinigaglia, F. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555–561.
- Pfeifer, N. and Kohlbacher, O. (2008) Multiple Instance Learning allows MHC class II epitope predictions for alleles without experimental data. Manuscript submitted.
- Goulmy, E. (2004) Minor histocompatibility antigens: allo target molecules for tumor-specific immunotherapy. *Cancer J.*, **10**, 1–7.

Supplementary material

Supplementary material 1: A list of all allelic models currently available in EpiToolKit.

Supplementary material 2: Novel ligands for HLA-A*0201 used for validation of the epitope prediction.

Supplementary material 3: A list of all mHags used for the validation of SNEPv2.

Supplementary material 1

The allelic models available in EpiToolKit are grouped by prediction method. The alleles per group are arranged alphabetically. The enclosed values are the available peptide lengths for the corresponding allele. For the methods *Hammer* and *Unitope* only nonameric models are available. The method *MHCIIIMulti* is not restricted to fixed peptide lengths.

Bimas alleles						
A*0101 (9)	A*0201 (9)	A*0205 (9)	A*0301 (9)	A*1101 (9)	A*2402 (9)	A*3101 (9)
A*3302 (9)	A*6801 (9)	B*0401 (9)	B*0702 (9)	B*0801 (8,9)	B*1501 (9)	B*2702 (9)
B*2705 (8,9)	B*3501 (8,9)	B*3701 (9)	B*3801 (9)	B*3901 (8,9)	B*3902 (9)	B*40 (9)
B*4001 (9)	B*4006 (8,9)	B*4403 (9)	B*5101 (8,9)	B*5102 (8,9)	B*5103 (9)	B*5201 (8,9)
B*5801 (9)	Cw*0301 (9)	Cw*0401 (9)	Cw*0602 (9)	Cw*0702 (9)		

Epidemix alleles						
A*01 (9)	A*0201 (9,10,11)	A*03 (9,10)	A*1101 (9,10)	A*24 (9)	A*2402 (9)	A*25 (9,10)
B*07 (9)	B*08 (8,9)	B*1801 (8,9)	B*27 (9,10)	B*2705 (9)	B*3701 (9)	B*44 (9)
B*5101 (8,9)						

Hammer alleles						
DRB1*0101	DRB1*0102	DRB1*0301	DRB1*0305	DRB1*0306	DRB1*0307	DRB1*0308
DRB1*0309	DRB1*0311	DRB1*0401	DRB1*0402	DRB1*0404	DRB1*0405	DRB1*0408
DRB1*0410	DRB1*0421	DRB1*0423	DRB1*0426	DRB1*0701	DRB1*0703	DRB1*0801
DRB1*0802	DRB1*0804	DRB1*0806	DRB1*0813	DRB1*0817	DRB1*1101	DRB1*1102
DRB1*1104	DRB1*1106	DRB1*1107	DRB1*1114	DRB1*1120	DRB1*1121	DRB1*1128
DRB1*1301	DRB1*1302	DRB1*1304	DRB1*1305	DRB1*1307	DRB1*1311	DRB1*1321
DRB1*1322	DRB1*1323	DRB1*1327	DRB1*1328	DRB1*1501	DRB1*1502	DRB1*1506

MHCIIIMulti alleles						
DRB1*0101	DRB1*0102	DRB1*0103	DRB1*0104	DRB1*0105	DRB1*0106	DRB1*0107
DRB1*0108	DRB1*0109	DRB1*0110	DRB1*0111	DRB1*0112	DRB1*0113	DRB1*0114
DRB1*0115	DRB1*0116	DRB1*0117	DRB1*0118	DRB1*0119	DRB1*0301	DRB1*0302
DRB1*0303	DRB1*0304	DRB1*0305	DRB1*0306	DRB1*0307	DRB1*0308	DRB1*0309
DRB1*0310	DRB1*0311	DRB1*0312	DRB1*0313	DRB1*0314	DRB1*0315	DRB1*0316
DRB1*0317	DRB1*0318	DRB1*0319	DRB1*0320	DRB1*0321	DRB1*0322	DRB1*0323
DRB1*0324	DRB1*0325	DRB1*0326	DRB1*0327	DRB1*0328	DRB1*0329	DRB1*0330
DRB1*0331	DRB1*0332	DRB1*0333	DRB1*0334	DRB1*0335	DRB1*0336	DRB1*0337
DRB1*0338	DRB1*0339	DRB1*0401	DRB1*0402	DRB1*0403	DRB1*0404	DRB1*0405
DRB1*0406	DRB1*0407	DRB1*0408	DRB1*0409	DRB1*0410	DRB1*0411	DRB1*0412
DRB1*0413	DRB1*0414	DRB1*0415	DRB1*0416	DRB1*0417	DRB1*0418	DRB1*0419
DRB1*0420	DRB1*0421	DRB1*0422	DRB1*0423	DRB1*0424	DRB1*0425	DRB1*0426
DRB1*0427	DRB1*0428	DRB1*0429	DRB1*0430	DRB1*0431	DRB1*0432	DRB1*0433
DRB1*0434	DRB1*0435	DRB1*0436	DRB1*0437	DRB1*0438	DRB1*0439	DRB1*0440
DRB1*0441	DRB1*0442	DRB1*0443	DRB1*0444	DRB1*0445	DRB1*0446	DRB1*0447
DRB1*0448	DRB1*0449	DRB1*0450	DRB1*0451	DRB1*0452	DRB1*0453	DRB1*0454
DRB1*0455	DRB1*0456	DRB1*0457	DRB1*0458	DRB1*0459	DRB1*0460	DRB1*0461
DRB1*0462	DRB1*0463	DRB1*0464	DRB1*0465	DRB1*0466	DRB1*0467	DRB1*0468
DRB1*0469	DRB1*0470	DRB1*0471	DRB1*0472	DRB1*0473	DRB1*0701	DRB1*0703
DRB1*0704	DRB1*0705	DRB1*0706	DRB1*0707	DRB1*0708	DRB1*0709	DRB1*0711

DRB5*0105	DRB5*0106	DRB5*0107	DRB5*0108	DRB5*0109	DRB5*0111	DRB5*0112
DRB5*0113	DRB5*0202	DRB5*0203	DRB5*0204	DRB5*0205		

Svmhc alleles						
A*01 (9)	A*0201 (9,10)	A*03 (9,10)	A*1101 (9)	A*24 (9)	A*2402 (9)	A*25 (10)
B*07 (9)	B*08 (9)	B*1501 (9,10)	B*1801 (8,9)	B*27 (9)	B*2705 (9)	B*3501 (9)
B*3701 (9)	B*44 (9)	B*5101 (9)				

Syfpeithi alleles						
A*0101 (9,10,11)	A*0201 (9,10)	A*0301 (9,10)	A*1101 (9,10)	A*2402 (9,10)	A*2601 (9,10)	A*6801 (9,10)
B*0702 (9,10)	B*0801 (8,9)	B*1402 (8,9)	B*1501 (9,10)	B*1510 (9)	B*1801 (8,9)	B*2705 (9)
B*3701 (8,9)	B*3801 (9,10)	B*3901 (9)	B*3902 (9)	B*4001 (9)	B*4101 (9)	B*4402 (9,10)
B*4501 (9)	B*4701 (9)	B*4901 (9)	B*5001 (9)	B*5101 (8,9)		

Unitope alleles						
A*0101	A*0102	A*0103	A*0106	A*0107	A*0108	A*0109
A*0110	A*0112	A*0113	A*0114	A*0117	A*0119	A*0120
A*0201	A*0202	A*0203	A*0204	A*0205	A*0206	A*0207
A*0208	A*0209	A*0210	A*0211	A*0212	A*0213	A*0214
A*0216	A*0217	A*0218	A*0219	A*0220	A*0221	A*0222
A*0224	A*0225	A*0226	A*0227	A*0228	A*0229	A*0230
A*0231	A*0233	A*0234	A*0235	A*0236	A*0237	A*0238
A*0239	A*0240	A*0241	A*0242	A*0244	A*0245	A*0246
A*0247	A*0248	A*0249	A*0250	A*0251	A*0252	A*0254
A*0255	A*0256	A*0257	A*0258	A*0259	A*0260	A*0261
A*0262	A*0263	A*0264	A*0265	A*0266	A*0267	A*0268
A*0269	A*0270	A*0271	A*0272	A*0273	A*0274	A*0275
A*0276	A*0277	A*0278	A*0279	A*0280	A*0281	A*0284
A*0285	A*0286	A*0287	A*0289	A*0290	A*0291	A*0292
A*0293	A*0295	A*0296	A*0297	A*0299	A*0301	A*0302
A*0304	A*0305	A*0306	A*0307	A*0308	A*0309	A*0310
A*0312	A*0313	A*0314	A*0315	A*0316	A*0317	A*0318
A*0319	A*0320	A*0322	A*0323	A*0324	A*0325	A*0326
A*1101	A*1102	A*1103	A*1104	A*1105	A*1106	A*1107
A*1108	A*1109	A*1110	A*1111	A*1112	A*1113	A*1114
A*1115	A*1116	A*1117	A*1118	A*1119	A*1120	A*1122
A*1123	A*1124	A*1125	A*1126	A*1127	A*1128	A*1129
A*2301	A*2302	A*2303	A*2304	A*2305	A*2306	A*2309
A*2310	A*2312	A*2313	A*2314	A*2402	A*2403	A*2404
A*2405	A*2406	A*2407	A*2408	A*2410	A*2413	A*2414
A*2415	A*2417	A*2418	A*2419	A*2420	A*2421	A*2422
A*2423	A*2424	A*2425	A*2426	A*2427	A*2428	A*2429
A*2430	A*2431	A*2432	A*2433	A*2434	A*2435	A*2437
A*2438	A*2439	A*2441	A*2442	A*2443	A*2444	A*2446
A*2447	A*2449	A*2450	A*2451	A*2452	A*2453	A*2454
A*2455	A*2456	A*2457	A*2458	A*2459	A*2461	A*2462
A*2463	A*2464	A*2465	A*2466	A*2467	A*2468	A*2501
A*2502	A*2503	A*2504	A*2505	A*2506	A*2601	A*2602
A*2603	A*2604	A*2605	A*2606	A*2607	A*2608	A*2609
A*2610	A*2612	A*2613	A*2614	A*2615	A*2616	A*2617

A*2618	A*2619	A*2620	A*2621	A*2622	A*2623	A*2624
A*2626	A*2627	A*2628	A*2629	A*2630	A*2631	A*2632
A*2633	A*2634	A*2901	A*2902	A*2903	A*2904	A*2905
A*2906	A*2907	A*2909	A*2910	A*2911	A*2912	A*2913
A*2914	A*2915	A*2916	A*3001	A*3002	A*3003	A*3004
A*3006	A*3007	A*3008	A*3009	A*3010	A*3011	A*3012
A*3013	A*3014L	A*3015	A*3016	A*3017	A*3018	A*3019
A*3101	A*3102	A*3103	A*3104	A*3105	A*3106	A*3107
A*3108	A*3109	A*3110	A*3111	A*3112	A*3113	A*3115
A*3201	A*3202	A*3203	A*3204	A*3205	A*3206	A*3207
A*3208	A*3209	A*3210	A*3211Q	A*3212	A*3213	A*3214
A*3301	A*3303	A*3304	A*3305	A*3306	A*3307	A*3308
A*3309	A*3401	A*3402	A*3403	A*3404	A*3405	A*3406
A*3407	A*3408	A*3601	A*3602	A*3603	A*3604	A*4301
A*6601	A*6602	A*6603	A*6604	A*6605	A*6606	A*6801
A*6802	A*6803	A*6804	A*6805	A*6806	A*6807	A*6808
A*6809	A*6810	A*6812	A*6813	A*6814	A*6815	A*6816
A*6817	A*6819	A*6820	A*6821	A*6822	A*6823	A*6824
A*6825	A*6826	A*6827	A*6828	A*6829	A*6830	A*6831
A*6832	A*6833	A*6834	A*6835	A*6836	A*6901	A*7401
A*7402	A*7403	A*7404	A*7405	A*7406	A*7407	A*7408
A*7409	A*7410	A*7411	A*8001	A*9201	A*9202	A*9203
A*9204	A*9206	A*9207	A*9208	A*9209	B*0702	B*0703
B*0704	B*0705	B*0706	B*0707	B*0708	B*0709	B*0710
B*0711	B*0712	B*0713	B*0714	B*0715	B*0716	B*0717
B*0718	B*0719	B*0720	B*0721	B*0722	B*0723	B*0724
B*0725	B*0726	B*0727	B*0728	B*0729	B*0730	B*0731
B*0732	B*0733	B*0734	B*0735	B*0736	B*0737	B*0738
B*0739	B*0740	B*0741	B*0742	B*0743	B*0744	B*0745
B*0746	B*0747	B*0748	B*0750	B*0751	B*0801	B*0802
B*0803	B*0804	B*0805	B*0806	B*0807	B*0809	B*0810
B*0811	B*0812	B*0813	B*0814	B*0815	B*0816	B*0817
B*0818	B*0820	B*0821	B*0822	B*0823	B*0824	B*0825
B*0826	B*0827	B*0828	B*0829	B*0831	B*1301	B*1302
B*1303	B*1304	B*1306	B*1308	B*1309	B*1310	B*1311
B*1312	B*1313	B*1314	B*1315	B*1316	B*1317	B*1401
B*1402	B*1403	B*1404	B*1405	B*1406	B*1501	B*1502
B*1503	B*1504	B*1505	B*1506	B*1507	B*1508	B*1509
B*1510	B*1511	B*1512	B*1513	B*1514	B*1515	B*1516
B*1517	B*1518	B*1519	B*1520	B*1521	B*1523	B*1524
B*1525	B*1527	B*1528	B*1529	B*1530	B*1531	B*1532
B*1533	B*1534	B*1535	B*1536	B*1537	B*1538	B*1539
B*1540	B*1542	B*1543	B*1544	B*1545	B*1546	B*1547
B*1548	B*1549	B*1550	B*1551	B*1552	B*1553	B*1554
B*1555	B*1556	B*1557	B*1558	B*1560	B*1561	B*1562
B*1563	B*1564	B*1565	B*1566	B*1567	B*1568	B*1569
B*1570	B*1571	B*1572	B*1573	B*1574	B*1575	B*1576
B*1577	B*1578	B*1580	B*1581	B*1582	B*1583	B*1584
B*1585	B*1586	B*1587	B*1588	B*1589	B*1590	B*1591
B*1592	B*1593	B*1595	B*1596	B*1597	B*1598	B*1599
B*1801	B*1802	B*1803	B*1804	B*1805	B*1806	B*1807
B*1808	B*1809	B*1810	B*1811	B*1812	B*1813	B*1814
B*1815	B*1818	B*1819	B*1820	B*1821	B*1822	B*1824

B*2701	B*2702	B*2703	B*2704	B*2705	B*2706	B*2707
B*2708	B*2709	B*2710	B*2711	B*2712	B*2713	B*2714
B*2715	B*2716	B*2717	B*2718	B*2719	B*2720	B*2721
B*2723	B*2724	B*2725	B*2726	B*2727	B*2728	B*2729
B*2730	B*2731	B*2732	B*2733	B*2734	B*2735	B*2736
B*3501	B*3502	B*3503	B*3504	B*3505	B*3506	B*3507
B*3508	B*3509	B*3510	B*3511	B*3512	B*3513	B*3514
B*3515	B*3516	B*3517	B*3518	B*3519	B*3520	B*3521
B*3522	B*3523	B*3524	B*3525	B*3526	B*3527	B*3528
B*3529	B*3530	B*3531	B*3532	B*3533	B*3534	B*3535
B*3536	B*3537	B*3538	B*3539	B*3541	B*3542	B*3543
B*3544	B*3545	B*3546	B*3547	B*3548	B*3549	B*3550
B*3551	B*3552	B*3554	B*3555	B*3556	B*3557	B*3558
B*3559	B*3560	B*3561	B*3562	B*3563	B*3564	B*3565Q
B*3566	B*3567	B*3568	B*3569	B*3570	B*3571	B*3572
B*3701	B*3702	B*3704	B*3705	B*3706	B*3707	B*3708
B*3709	B*3710	B*3711	B*3712	B*3801	B*3802	B*3803
B*3804	B*3805	B*3806	B*3807	B*3808	B*3809	B*3810
B*3811	B*3812	B*3813	B*3814	B*3815	B*3901	B*3902
B*3903	B*3904	B*3905	B*3906	B*3908	B*3909	B*3910
B*3911	B*3912	B*3913	B*3914	B*3915	B*3916	B*3917
B*3918	B*3919	B*3920	B*3922	B*3923	B*3924	B*3926
B*3927	B*3928	B*3929	B*3930	B*3931	B*3932	B*3933
B*3934	B*3935	B*3936	B*3937	B*3938Q	B*3939	B*3941
B*4001	B*4002	B*4003	B*4004	B*4005	B*4006	B*4007
B*4008	B*4009	B*4010	B*4011	B*4012	B*4013	B*4014
B*4015	B*4016	B*4018	B*4019	B*4020	B*4021	B*4023
B*4024	B*4025	B*4026	B*4027	B*4028	B*4029	B*4030
B*4031	B*4032	B*4033	B*4034	B*4035	B*4036	B*4037
B*4038	B*4039	B*4040	B*4042	B*4043	B*4044	B*4045
B*4046	B*4047	B*4048	B*4049	B*4050	B*4051	B*4052
B*4053	B*4054	B*4055	B*4056	B*4057	B*4058	B*4059
B*4060	B*4061	B*4062	B*4063	B*4064	B*4065	B*4066
B*4067	B*4068	B*4069	B*4101	B*4102	B*4103	B*4104
B*4105	B*4106	B*4107	B*4108	B*4201	B*4202	B*4204
B*4205	B*4206	B*4207	B*4208	B*4209	B*4402	B*4403
B*4404	B*4405	B*4406	B*4407	B*4408	B*4409	B*4410
B*4411	B*4412	B*4413	B*4414	B*4415	B*4416	B*4417
B*4418	B*4420	B*4421	B*4422	B*4424	B*4425	B*4426
B*4427	B*4428	B*4429	B*4430	B*4431	B*4432	B*4433
B*4434	B*4435	B*4436	B*4437	B*4438	B*4439	B*4440
B*4441	B*4442	B*4443	B*4444	B*4445	B*4446	B*4447
B*4448	B*4449	B*4450	B*4451	B*4501	B*4502	B*4503
B*4504	B*4505	B*4506	B*4507	B*4601	B*4602	B*4603
B*4604	B*4605	B*4606	B*4608	B*4609	B*4701	B*4702
B*4703	B*4704	B*4705	B*4801	B*4802	B*4803	B*4804
B*4805	B*4806	B*4807	B*4808	B*4809	B*4810	B*4811
B*4812	B*4813	B*4814	B*4815	B*4816	B*4901	B*4902
B*4903	B*4904	B*4905	B*5001	B*5002	B*5004	B*5101
B*5102	B*5103	B*5104	B*5105	B*5106	B*5107	B*5108
B*5109	B*5110	B*5112	B*5113	B*5114	B*5115	B*5116
B*5117	B*5118	B*5119	B*5120	B*5121	B*5122	B*5123
B*5124	B*5126	B*5128	B*5129	B*5130	B*5131	B*5132

B*5133	B*5134	B*5135	B*5136	B*5137	B*5138	B*5139
B*5140	B*5142	B*5143	B*5145	B*5146	B*5201	B*5202
B*5203	B*5204	B*5205	B*5206	B*5207	B*5208	B*5209
B*5210	B*5301	B*5302	B*5303	B*5304	B*5305	B*5306
B*5307	B*5308	B*5309	B*5310	B*5311	B*5312	B*5401
B*5402	B*5403	B*5404	B*5406	B*5407	B*5409	B*5410
B*5501	B*5502	B*5503	B*5504	B*5505	B*5507	B*5508
B*5509	B*5510	B*5511	B*5512	B*5513	B*5514	B*5515
B*5516	B*5517	B*5518	B*5519	B*5520	B*5521	B*5522
B*5523	B*5524	B*5601	B*5602	B*5603	B*5604	B*5605
B*5606	B*5607	B*5608	B*5609	B*5610	B*5611	B*5612
B*5613	B*5614	B*5615	B*5616	B*5617	B*5618	B*5701
B*5702	B*5703	B*5704	B*5705	B*5706	B*5707	B*5708
B*5709	B*5710	B*5711	B*5801	B*5802	B*5804	B*5805
B*5806	B*5807	B*5808	B*5809	B*5811	B*5812	B*5813
B*5814	B*5901	B*5902	B*6701	B*6702	B*7301	B*7801
B*7802	B*7803	B*7804	B*7805	B*8101	B*8102	B*8201
B*8202	B*8301	B*9501	B*9502	B*9503	B*9504	B*9505
B*9506	B*9507	B*9508	B*9509	B*9510	B*9512	B*9513
B*9514	B*9515	B*9516	B*9517	B*9518	B*9519	B*9520
B*9521	B*9522	Cw*0102	Cw*0103	Cw*0104	Cw*0105	Cw*0106
Cw*0107	Cw*0108	Cw*0109	Cw*0110	Cw*0111	Cw*0112	Cw*0113
Cw*0202	Cw*0203	Cw*0204	Cw*0205	Cw*0206	Cw*0207	Cw*0208
Cw*0209	Cw*0210	Cw*0211	Cw*0212	Cw*0213	Cw*0214	Cw*0215
Cw*0216	Cw*0217	Cw*0302	Cw*0303	Cw*0304	Cw*0305	Cw*0306
Cw*0307	Cw*0308	Cw*0309	Cw*0310	Cw*0311	Cw*0312	Cw*0313
Cw*0314	Cw*0315	Cw*0316	Cw*0317	Cw*0318	Cw*0319	Cw*0321
Cw*0322	Cw*0323	Cw*0324	Cw*0325	Cw*0326	Cw*0327	Cw*0328
Cw*0329	Cw*0330	Cw*0331	Cw*0332	Cw*0333	Cw*0334	Cw*0335
Cw*0401	Cw*0403	Cw*0404	Cw*0405	Cw*0406	Cw*0407	Cw*0408
Cw*0410	Cw*0411	Cw*0412	Cw*0413	Cw*0414	Cw*0415	Cw*0416
Cw*0417	Cw*0418	Cw*0419	Cw*0420	Cw*0421	Cw*0423	Cw*0424
Cw*0501	Cw*0502	Cw*0503	Cw*0504	Cw*0505	Cw*0506	Cw*0508
Cw*0509	Cw*0510	Cw*0511	Cw*0512	Cw*0513	Cw*0514	Cw*0515
Cw*0602	Cw*0603	Cw*0604	Cw*0605	Cw*0606	Cw*0607	Cw*0608
Cw*0609	Cw*0610	Cw*0611	Cw*0612	Cw*0613	Cw*0614	Cw*0701
Cw*0702	Cw*0703	Cw*0704	Cw*0705	Cw*0706	Cw*0707	Cw*0708
Cw*0709	Cw*0710	Cw*0711	Cw*0712	Cw*0713	Cw*0714	Cw*0715
Cw*0716	Cw*0717	Cw*0718	Cw*0719	Cw*0720	Cw*0721	Cw*0722
Cw*0723	Cw*0724	Cw*0725	Cw*0726	Cw*0727	Cw*0728	Cw*0729
Cw*0730	Cw*0731	Cw*0734	Cw*0735	Cw*0736	Cw*0737	Cw*0738
Cw*0739	Cw*0740	Cw*0741	Cw*0742	Cw*0743	Cw*0744	Cw*0745
Cw*0801	Cw*0802	Cw*0803	Cw*0804	Cw*0805	Cw*0806	Cw*0807
Cw*0808	Cw*0809	Cw*0810	Cw*0811	Cw*0812	Cw*0813	Cw*0814
Cw*1202	Cw*1203	Cw*1204	Cw*1205	Cw*1206	Cw*1207	Cw*1208
Cw*1209	Cw*1210	Cw*1211	Cw*1212	Cw*1213	Cw*1214	Cw*1215
Cw*1216	Cw*1217	Cw*1218	Cw*1219	Cw*1402	Cw*1403	Cw*1404
Cw*1405	Cw*1406	Cw*1408	Cw*1502	Cw*1503	Cw*1504	Cw*1505
Cw*1506	Cw*1507	Cw*1508	Cw*1509	Cw*1510	Cw*1511	Cw*1512
Cw*1513	Cw*1514	Cw*1515	Cw*1516	Cw*1517	Cw*1601	Cw*1602
Cw*1604	Cw*1606	Cw*1607	Cw*1608	Cw*1609	Cw*1701	Cw*1702
Cw*1703	Cw*1704	Cw*1801	Cw*1802			

Supplementary material 2

Twenty-seven novel HLA-A*0201 ligands were used for the validation of the EpiToolKit epitope prediction. For each peptide the corresponding GeneID is given. The last column states if the peptide was correctly predicted to bind to HLA-A*0201.

Peptide Sequence	GeneID	Identified as Binder
------------------	--------	----------------------

Nonameric peptides:

HTANIQTLI	57082	Yes
MLERVIKNY	4103	No
TMLARLASA	1464	Yes
ALLGKIEKV	2590	Yes
ALNEKLVNL	8665	Yes
ETFEHPNVV	1021	No
EVSNTFVRL	10623	Yes
GLNDFIQKI	6446	Yes
HVVDFFNRV	126308	Yes
LMVDHVTEV	10011	Yes
NMLPQIFRV	57102	Yes
RIIDVVYNA	6202	Yes
SLDQPTQTV	8663	Yes
SLIKQIPRI	4736	Yes
VLQTRKYLA	5251	Yes
YLADTVQKL	54937	Yes
HACGVIATI	23585	Yes
KAFQKIVVL	9689	Yes
SLLDKIIGA	284119	Yes
TIPEAFRGV	1192	Yes

Decameric peptides:

QLLSVDGRSL	4301	Yes
DALDGLLQQV	8178	Yes
DVYSGTPTKV	7884	No
ETIPGPDAKV	84153	Yes
FLDPNNIPKA	79053	Yes
GILCLQAHSA	400073	Yes
LSLLLGYLKL	9675	Yes

Supplementary material 3

A list of the mHags used for the validation of SNEPv2. These mHags are taken from a publication by Goulmy¹. The list contains all human mHags that were identified by the publication date of this paper. The polymorphic positions – if available – are marked in red.

Name (by Goulmy)	Sequence	MHC Restriction
------------------	----------	-----------------

Autosomally encoded:

HA-1	VL[R H]DDLLEA	HLA-A2
HA-1	KECVL[H R]DDL	HLA-B60
HA-2	YIGEVLS[V M]	HLA-A2
HA-3	V[M T]EPGTAQY	HLA-A1
HA-8	[R P]TLDKVLEV	HLA-A2
HB-1	EEKRGS[L H Y]VW	HLA-B44
BCL2A1	DYLQ[C Y]VLQI	HLA-A24
BCL2A1	KEFED[G D]IINW	HLA-B44
UGT2B17	AELLNIPVLY	HLA-A29

Y-chromosome encoded:

SMCY	FIDSYICQV	HLA-A2
SMCY	SPSVDKARAEL	HLA-B7
UTY	LPHNHTDL	HLA-B8
UTY	RESEEVSVSL	HLA-B60
DFFRY	IVD[C S]LTEMY	HLA-A1
DBY	HIENFSDIDMGE	HLA-DQ5
RPS4Y	VIKVNDTVQI	HLA-DRB3*0301

¹ Goulmy, E. (2004) Minor Histocompatibility Antigens: Allo Target Molecules for Tumor-Specific Immunotherapy. *Cancer J*, 10, 1-7. C1.