The handle http://hdl.handle.net/1887/35349 holds various files of this Leiden University dissertation.

**Author:** Hassan, Chopie  
**Title:** New strategies in identification and quantification of minor histocompatibility antigens  
**Issue Date:** 2015-09-16
CHAPTER 2

The Human Leukocyte Antigen-Presented Ligandome of B-Lymphocytes

Molecular & Cellular Proteomics (2013) 12 (7):1829-43
Based on:

Chopie Hassan*
Michel G.D. Kester*
Arnoud H. de Ru
Pleun Hombrink
Jan Wouter Drijfhout
Harm Nijveen
Jack A.M. Leunissen
Mirjam, H.M. Heemskerk
J.H. Frederik Falkenburg
Peter A. van Veelen

* equal contribution
The Human Leukocyte Antigen-Presented Ligandome of B-Lymphocytes
ABSTRACT

Peptides presented by human leukocyte antigen (HLA) molecules on the cell surface play a crucial role in adaptive immunology, mediating the communication between T cells and antigen presenting cells. Knowledge of these peptides is of pivotal importance in fundamental studies on T cell action, and in cellular immunotherapy and transplantation. In this study we present the in-depth identification and relative quantification of 14,500 peptide ligands constituting the HLA-ligandome of B-cells. This large number of identified ligands provides a general insight in the presented peptide repertoire and antigen presentation. Our uniquely large set of HLA-ligands allowed us to characterize in detail the peptides constituting the ligandome in terms of relative abundance, peptide length distribution, physicochemical properties, binding affinity to the HLA molecule and presence of post-translational modifications. The presented B-lymphocyte ligandome is shown to be a rich source of information by the presence of minor histocompatibility antigens, virus-derived epitopes and post-translationally modified HLA ligands and can be a good starting point to solve a wealth of specific immunological questions. These HLA ligands can form the basis for reversed immunology approaches to identify T cell epitopes, not based on in silico predictions, but based on the bona fide eluted HLA-ligandome.
INTRODUCTION

Peptides presented by human leukocyte antigen (HLA) molecules on the cell surface play a crucial role in immunology, and mediate the communication between T cells and antigen presenting cells. Knowledge of these peptides is of pivotal importance in fundamental studies on T cell action and in design of T cell-mediated therapy, like tumor immunotherapy [1], and treatment of hematological malignancies by a combination of hematopoietic stem cell transplantation (HSCT) and donor lymphocyte infusion (DLI) [2]. In addition, T cells can play an important role in organ rejection following transplantation.

The presented HLA class I ligands are the products of the intracellular processing machinery with its continuous cycle of protein synthesis and degradation [3]. Much is known about the proteins involved in antigen processing, but high fidelity ligand/epitope predictions are at present not possible. The complexity of antigen processing has been shown to be even more complex by the discovery of new enzymes involved in antigen processing [3, 4] and the exciting discovery of peptide splicing [5]. Moreover, gene expression studies have shown many non-standard, unexpected protein products, including the production of antigens derived from aberrant protein fragments as a result of expression in alternative reading frames [6]. Several studies report on the identification of HLA ligands [7-10]. Many results have been collected and discussed in a recent review on large-scale analysis of HLA class I ligands [11]. Collectively, these reports illustrate the need for in-depth elucidation of the HLA-ligandome.

Elucidation of T cell epitopes has traditionally been achieved by a forward immunological approach, as pioneered by Hunt and coworkers [12, 13]. In this approach, the cognate peptide of T cells with the appropriate activity profile is elucidated by repeated rounds of chromatographic separation in combination with T cell recognition assays. Since T cells are not always available from the start, reverse immunological approaches [14-17] have been developed to predict T cell epitopes by a combination of bioinformatics and e.g. in vitro proteasome digests. Predicted epitopes are synthesized and tested for their capability to activate T cells. The main disadvantage of this approach is that less than 0.1% of the peptides that survive intracellular processing, are presented on HLA- class I molecules [3].

Therefore, we developed a large scale peptidomics approach, which is a reverse immunology approach, not based on algorithms, but based on the bona fide eluted ligandome, which means that the identified peptides are known to have survived processing and are bona fide HLA-ligands. Following the identification of the ligandome as comprehensively as possible, T cells can subsequently be selected on the basis of the immunological question at hand, as will be illustrated in a separate paper. The development of MHC exchange tetramers for finding relevant T cell epitopes is instrumental to this approach [18, 19].

To improve ligandome coverage we applied and compared three off-line first dimension separation techniques,
followed by on-line nanoHPLC-tandem MS.

The tandem mass spectra were interrogated by matching against the IPIhuman database [20]. In a second step post translation modifications (phosphorylation, cysteinylation) were allowed in the database search. In a third step the tandem mass spectra were matched against a newly in-house developed database for optimal finding of polymorphic ligands, for finding potential minor histocompatibility antigens [21]. This led to the identification of approximately 14,000 HLA class I ligands, the majority of which was also relatively quantitated. Next, we analyzed the peptides constituting our ligandome as detailed as possible to confirm the correct identification of the vast majority of the ligands. We achieved this by a combination of several physicochemical and biological checks and comparison with existing ligand and epitope databases.

Finally, as an additional quality check, we illustrate the functional relevance of the ligandome by the identification of previously known and new minor histocompatibility antigens, virus-derived epitopes and post-translationally modified HLA ligands (phosphorylated ligands, and cysteinylated ligands) [22-24]. Since this is the largest ligandome reported to date, it allows a general insight in the presented peptide repertoire. This study supports the building of the ‘immunopeptidome’ as has recently been suggested [25]. A proteomics approach as a starting point to contribute to immunology by providing a peptidome landscape can be used in many immunological studies, both fundamental and applied.

**EXPERIMENTAL PROCEDURES**

**Sample preparation**

The Epstein-Barr virus (EBV) transformed B lymphoblastic cell lines B-LCL-HHC (typing: HLA-A*0201, B*0702, B*4402, Cw*0501 & Cw*0702) and B-LCL-JY pp65 (typing: HLA-A*0201, B*0702 & Cw*0702) were used as source of HLA-class I molecules. The CMV-derived pp65 transduced cell line was used to introduce an internal control for the ligandome since the CMVpp65-derived T cell epitopes are known [26]. Cells were expanded in roller bottles using IMDM supplemented with 10% heat-inactivated fetal bovine serum (FBS), penicillin/streptomycin and L-glutamine, were collected, washed with ice cold PBS, and stored at -80°C until use.

The hybridoma cell line was expanded in roller bottles to obtain W6/32 (anti HLA-Class I) antibody using protein free hybridoma medium supplemented with penicillin/streptomycin, and L-glutamine. Antibodies produced by the hybridoma cell lines were purified from the supernatant using Prot-A sepharose beads, and eluted from the Prot-A beads with Glycine pH 2.5. The eluted antibodies were used to produce immunoaffinity column (W6/32- Prot-A sepharose 2.5 mg/ml). The W6/32 antibodies were covalently bound to Prot-A sepharose beads using dimethylpimelimidate (DMP). The columns were stored in PBS pH 8.0 and 0.02% NaN₃ at 4°C.
Isolation of HLA class I-presented peptides

Extraction of peptides associated with HLA-class I molecules was performed as described previously [13, 27]. Briefly, pellets from 60 x10^9 B-LCL-JYpp65 cells and 40 x10^9 B-LCL-HHC cells were lysed in 50 mM Tris-HCl, 150 mM NaCl, 5 mM EDTA, and 0.5% Nonidet-P40, (pH 8.0) and supplemented with Complete® protease inhibitor (Sigma Aldrich). The total concentration of the cells in the lysis buffer was 0.1x 10^9 cells/ml.

After 2 hours incubation with tumbling of the cells in the lysis buffer at 4 °C, the preparation was centrifuged for 10 minutes at 2,500 rpm and 4 °C. The supernatant was transferred to a new tube and centrifuged for 35 minutes at 11,000 rpm and 4°C. The supernatant was pre-cleared with CL4B beads and subjected to the immunoaffinity column with a flow rate of 2.5 ml/min. After washing, bound HLA-class I/peptide complexes were eluted from the column, and dissociated with 10% acetic acid. Peptides were separated from the HLA-class I molecules by passage through a 10 kDa membrane (Pall macropore centrifuge devices). The filtrate was freeze dried. If an oily sample remained after freeze drying, the sample was dissolved and the peptides were further purified by solid phase extraction (C18 Oasis, 100 µl bed volume, Waters). The peptides were eluted from the C18 Oasis column with 500 µl 50/50/0.1 water/ACN/FA, v/v/v. The eluted peptides from B-LCL-HHC cell lines were divided into two equal portions, freeze dried and dissolved in 95/3/0.1 water/ACN/FA, v/v/v. The eluted peptides from B-LCL-JYpp65 cell lines were divided into three equal portions, freeze dried, and dissolved in 95/3/0.1 water/ACN/FA, v/v/v.

Peptide separation

For peptide IEF separations, the OFFGEL Agilent 3100 fractionator (Agilent Technologies, Waldbronn, Germany) was used. A modified method was applied by addition of 1 M urea to the buffer sample and rehydration buffer, instead of 5% glycerol only. The commercially available 13 cm IPG dry strips with a linear pH gradient ranging from 3-10 (GE-Health care) were used. The strips were rehydrated with 40 µL/well rehydration solution in the assembled device for 30 min. 150 µL of the prepared samples were loaded on each well, the cover fluid (mineral oil, Agilent technologies) was added onto both ends of the gel strip. The focusing methods, OG12PE01, as supplied by the manufacturer was applied for 12 well fractionations. The performance of the 3100 OFFGEL fractionator was checked under similar conditions in a separate run by determination of the pH using a pH indicator pH 3-10 (Fluka Analytical, Germany). Fractions were recovered and desalted by solid phase extraction, to desalt the fractions and to avoid the presence of oil and gel pieces in the samples, using C18 Oasis columns. The column was prewetted with 10/90 water/ACN v/v and equilibrated with 95/3/0.1 water/ACN/FA v/v/v. The samples were eluted with 50/50/0.1 water/ACN/FA v/v/v, freeze dried and dissolved in 100 µL 95/3/0.1 water/ACN/FA v/v/v. The twelve fractions generated from IEF were analyzed in triplicate
CHAPTER 2

and duplicate for B-LCL-HHC and B-LCL-JYpp65 cell lines, respectively, with nanoLC-MS/MS.

Next to the peptide IEF separation, two chromatographic separation techniques were applied, strong cation chromatography (SCX) and RP-C18 chromatography. For SCX separations, one portion of the eluted peptides from the W6/32 column was fractionated with a home-made SCX column (320 µm ID, 15 cm, polysulfoethyl A 3 µm, Poly LC), run at 4 ul/min. Gradients were run for 10 min at 100% solvent A (100% water/0.1% TFA), after which a linear gradient started to reach 100% solvent B (250 mM KCl, 35% ACN/0.1% TFA) in 15 min, followed by 100 % solvent C (500 mM KCl, 35% ACN/0.1% TFA) in the next 15 min and remained at 100 % solvent C for 5 min, then switched again to 100 % solvent A. Twenty 4 ul-fractions were collected in vials prefilled with 100 µl 95/3/0.1 water/ACN/FA v/v/v.

One portion of the eluted peptides from B-LCL-JY pp65 cell lines was fractionated on a home-made RP Reprosil-Pur C18-AQ column (200 µm ID, 3 µm x 15 cm) (Dr. Maisch, GmbH, Ammerbuch, Germany). The sample was loaded in solvent A (10/90/0.1 water/ACN/FA v/v/v), and the gradient was run from 0-50% B (10/90/0.1 water/ACN/FA v/v/v) in 30 min at a flow rate of 3 ul/min. The samples were taken up in a make-up flow of 50/50/0.1 water/ACN/FA at 100 ul/min supplied via a T-piece through the annular space between the separation capillary and an auxiliary capillary. In this way 45 half a minute wide fractions were collected, subsequently freeze dried, and dissolved in solvent A for analysis by nanoLC-MS/MS.

**LC-MS/MS analysis**

The dissolved fractions were analyzed by on-line nano-HPLC mass spectrometry with a system, consisting of a conventional Agilent 1100 gradient HPLC system (Agilent, Waldbronn, Germany), described by Meiring et al [28], and a LTQ-FT Ultra mass spectrometer (Thermo, Bremen, Germany). Fractions were injected onto a home-made pre-column (100 um×15 mm; Reprosil-Pur C18-AQ 3 µm, Dr. Maisch, Ammerbuch, Germany) and eluted via in home-made analytical nano-HPLC column (15 cm×50 µm; Reprosil-Pur C18-AQ 3 µm). The gradient was run from 0% to 50% solvent B (10/90/0.1 water/ACN/FA v/v/v) in 90 min. The nano-HPLC column was drawn to a tip of approximately 5 µm and acted as the electrospray needle of the MS source. The mass spectrometer was operated in data dependent mode, automatically switching between MS and MS/MS acquisition. Full scan mass spectra were acquired in the FT-ICR with a resolution of 25,000 at a target value of 5,000,000. The two most intense ions were then isolated for accurate mass measurements by a selected ion monitoring scan in FT-ICR with a resolution of 50,000 at a target accumulation value of 50,000. The selected ions were then fragmented in the linear ion trap using collision-induced dissociation at a target value of 10,000. In a post analysis process, raw data were converted to peak lists using Bioworks Browser software, Version 3.2.0.
**Data analysis**

The tandem mass spectra were matched against the International Protein Index (IPI) human database version 3.87, using the mascot search engine version 2.2.04 (Matrix Science, London, UK), with a precursor mass tolerance of 2 ppm, with methionine oxidation as a variable modification, and a product ion tolerance of 0.5 Da. For finding post-translationally modified HLA-ligands phosphorylation on serine, threonine and tyrosine were allowed, and cysteinylation of cysteine in separate searches. Scaffold software version 3 ([www.proteomesoftware.com](http://www.proteomesoftware.com)) was subsequently used to process the mascot output files and generate spectrum reports. Duplicates were removed, and peptides with a best mascot ion score >35 and 8-11 amino acids long, were selected for the production of supplemental Table 1. For the immunological examples a best mascot score of >20 was selected, and the length was restricted to 8-18 amino acids. Next to the above mentioned procedure, Proteome Discoverer 1.3 (Thermo, Bremen, Germany) was used to extract all identified peptides from the input *.RAW-files, using the mascot server mentioned above, and calculate their intensity, as reported in supplemental Table 1. False discovery rates were as determined by Proteome Discoverer, for the Homo Sapiens-extracted Uniprot/SWISSProt database [29] (release 2010_11) ([www.ebi.ac.uk](http://www.ebi.ac.uk)) containing 20,259 protein sequences, and the IPIhuman 3.87 database ([www.ebi.ac.uk/IPI/IPIhuman.html](http://www.ebi.ac.uk/IPI/IPIhuman.html)) see also supplemental Table 2. IceLogo version 1.2 was used to generate the binding motifs as presented in Figure 5 and supplemental Figure 1 [30]. The GRAVY index was calculated using [http://www.bioinformatics.org/sms2/protein_gravy.html](http://www.bioinformatics.org/sms2/protein_gravy.html). The pIs of the identified peptides were calculated using [http://www.expasy.org/tools/pi_tool.html](http://www.expasy.org/tools/pi_tool.html). NetMHC 3.2 ([http://www.cbs.dtu.dk/services/NetMHC](http://www.cbs.dtu.dk/services/NetMHC)) was used to predict the binding affinity (nM) of the identified peptides to HLA-A*0201, HLA-B*0702 and HLA-B*4402. NetMHCpan 2.4 ([http://www.cbs.dtu.dk/services/NetMHCpan](http://www.cbs.dtu.dk/services/NetMHCpan)) was used to predict the binding affinity of the identified peptides to HLA-C*0501 and HLA-C*0702. Overall protein turnover values were taken from Cambridge et al. [31]. Phosphosite.org [32] ([www.phosphosite.org](http://www.phosphosite.org)) and Phospho.ELM 8.3 [33] ([http://phospho.elm.eu.org](http://phospho.elm.eu.org)) were used to find known phosphosites in the identified peptides. NetPhos 2.0 [34] ([http://www.cbs.dtu.dk/services/NetPhos](http://www.cbs.dtu.dk/services/NetPhos)) was used to predict phosphosites in identified phosphorylated ligands. For the identification of polymorphic peptides, the tandem mass spectra were matched against the HSPVdb, a database optimized for finding polymorphic peptides [21]. For searching CMV pp65-derived ligands/epitopes in B-LCL-JYpp65, a separate database was constructed only containing the DNA sequence of the pp65 protein (NCBI; pp65_AD169_seq with intron Human herpesvirus 5, complete genome). For searching EBV-derived epitopes, a separate database was constructed containing the DNA sequence of EBV selected from RefSeq database [35] (>gi|82503188|ref|NC_007605.1) Human herpesvirus 4 type 1, complete genome).
Peptide synthesis

Peptides were synthesized according to standard fluorenylmethoxycarbonyl (F-moc) chemistry using a SyroII peptide synthesizer (MultiSynTech, Witten, Germany). The integrity of the peptides was checked using RP-HPLC and MS. Phospho peptides were synthesized using the building blocks Fmoc-Ser(PO(OBzl)OH)-OH or Fmoc-Thr(PO(OBzl)OH)-OH. Couplings of these building blocks were performed identical to normal couplings with one exception. A 3-fold excess of N-methylmorpholine was used instead of the routinely applied 2-fold excess.

RESULTS

The peptides constituting the HLA-ligandome

The pool of HLA-peptides eluted from two Epstein-Barr virus-transformed B lymphoblastoid cell lines (EBV-B-LCL), either B-LCL-JYpp65 or B-LCL-HHC, is quite complex. From previous experiments a simple count of the total number of peaks in the MS1 spectra in all fractions following a first dimension C18-separation resulted in a number of approximately 30,000. Because of this complexity, multidimensional separations were performed to reduce the complexity before mass spectrometric analysis. Three first dimension separations were chosen, reverse phase C18 (RP-C18) chromatography, strong cation exchange (SCX) chromatography, and peptide isoelectric focusing (IEF), since these are based on a different separation mechanism, and most commonly used. The second dimension separation was RP-C18 chromatography coupled on-line to the mass spectrometer in all cases. Below the results of the three first dimension separations in combination with the second dimension on-line HPLC-MS are compared and discussed.

Peptide isoelectric focusing as the 1st dimension

Before starting our IEF-experiments with HLA-eluted peptides, the separation performance of the OFFGEL 3100 system was tested using a trypsin-digested cell lysate, using 13 cm IPG strips, pH range of 3-10 (GE-Healthcare). Separation efficiency was studied as a function of loading. The results were as reported in [36], with approximately 83% of peptides present in a unique fraction and 96% present in one or two unique fractions, with peptide amounts up to 100 μg, confirming the high separation efficiency of the IEF process. Peptide loading below 10 μg results in considerable sample losses, which precludes use of IEF for HLA-ligandome studies of smaller cell amounts. Application of a pH indicator (pH 3 - pH 10) after IEF fractionation showed that the low pH side of the strip was not pH 3, but was actually close to pH 4.

The 12 fractions obtained from the IEF fractionation were analyzed by on-line nanoHPLC tandem mass spectrometry, and the tandem mass spectra were matched against the IPI human protein database. The results for both B-LCLs are summarized as Venn diagrams in Figure 1(A & B). As can be seen, analysis of the
B-LCL-HHC ligandome yielded 6878 unique 8-11 mer peptides with BMI>35, and the B-LCL-JYpp65 ligandome yielded 3531 unique peptides. The large difference in the number of peptides found for B-LCL-JYpp65 and B-LCL-HHC can be largely explained by the additional HLA-molecules expressed on B-LCL-HHC compared to B-LCL-JY cells. B-LCL-JY is homozygous for the B-allele (HLA-B7*0702) and C allele (HLA-Cw*0702), while B-LCL-HHC is heterozygous for the B-allele (HLA-B*0702 and HLA-B*4402) and C-allele (HLA-Cw*0501 and HLA-Cw*0702). The additional HLA-molecules can present two additional sets of peptides.

**Strong cation exchange (SCX) chromatography as the 1st dimension**

Twenty fractions were collected after separation on the SCX-column. These fractions were analyzed, like the IEF-fractions, by on-line nanoHPLC tandem mass spectrometry and they were matched against the IPI human protein database. The results for both B-LCLs are summarized as Venn diagrams in Figure. 1 (C& D) . Analysis of the B-LCL-HHC ligandome yielded 7766 unique peptides (BMI>35) and the B-LCL-JYpp65 ligandome yielded 3410 unique peptides (BMI>35).

**Reverse Phase-C18 Chromatography as the 1st dimension**

Forty five fractions were collected after separation on the RP-C18-column. These fractions were analyzed, like the IEF- and SCX-fractions, by on-line nanoHPLC tandem mass spectrometry and the tandem mass spectra were again matched against the IPI human protein database. Analysis of the B-LCL-JYpp65 ligandome yielded 2725 unique peptides (BMI>35). The results for B-LCL-JY are summarized as Venn diagrams in Figure. 1 (E). The B-LCL-HHC ligandome was not measured with the combination of RP-C18 and on-line nanoHPLC.

**Combination of the results of IEF, SCX and RP-C18**

To get an overview, a non-redundant list of all peptides obtained via the above-mentioned first dimension separation techniques was produced. The results are shown as Venn diagrams in Figure. 2. The analysis of the B-LCL-HHC ligandome yielded 10,867, Figure. 2 (A) unique peptides (BMI>35) and the B-LCL-JYpp65 ligandome yielded 6,493, Figure. 2 (B) unique peptides (BMI>35). Combination of the two B-LCL-derived peptide lists yields a non-redundant peptide list containing 14,065 members, Figure. 2 (C), of which 11,511 could be relatively quantified. This is by far the most comprehensive list of HLA-ligands reported to date. A complete listing of the HLA-presented peptides, sorted by intensity, is given in supplemental Table S1, including the protein name and IPI accession number they are derived from and their intensity in B-LCL-JY and/or B-LCL-HHC, the best mascot ion score, the predicted binding affinity by NetMHC, the GRAVY hydrophobicity value, the protein half-life, the copy number per cell calculated from the intensity and work-up yield, the peptide length and if peptides are present in the IEDB [37] and/or SYFPEITHI database [38], and the calculated FDR using the Uniprot/SWISSProt database and IPIhuman database. In addition, we compared
our data with those reported by [7-10].

The NetMHC algorithm was used to predict the HLA-binding affinity of the ligands. It is clear from the results that the majority of peptides is a binder in its particular HLA-molecule, see supplemental Table S1. In the table a NetMHC cut-off score of 1000 was used.

The number of peptides found in every experiment, independent of the work up procedure, is 835 of which 770 could be quantified, Figure 2 (B). Because of the shotgun nature of the experimental approach these 835 peptides are expected to be the most abundant on average, the peptides detected in either overlap of 2 experiments to be less abundant, and the peptides detected only in a single first dimension experiment to be the lowest. This was checked for the peptides presented on B-LCL-JY and the result displayed in Figure. 3.

Peptide length distribution

Our large set of peptides allows a closer look at the length distribution of peptides constituting the HLA-ligandome, as presented in Figure. 4. As can be seen, most ligands have a length of 9 amino acids, while a substantial part is either 10 or 11 amino acids long. The 8-mers, and 12-14 mers are all below 5%. However, still 708 long (>11 amino acids) peptides (amounting to 5% of the ligandome) are present, having the correct anchors for binding to the HLA-A and B alleles. Figure. 5 shows the HLA-A2 binding motif for the 9, 12 and 14-mers. All different peptide lengths display the same binding motif, and a binding motif can still be discerned for 15-mers.

The dynamic range of the HLA-ligandome

To get an impression of the dynamic range of the presented peptide repertoire, the intensities of the presented peptides in the ligandome were determined. To this end, the *.RAW-files were processed in Proteome Discoverer.

A complete listing of all peptides, including their intensity, presentation on the particular B-LCL, and its best mascot ion score is given in supplemental Table S1. Intensity extremes, expressed as peak area values in PD, range from 4E9 to 1E5, but the majority ranges from 1E9 to 1E6, a range of a factor of 1000. So the dynamic range of the peptide repertoire is quite high.

The summed intensity (integration value) of all peptides eluted from 2E10 B-LCL-JY or B-LCL-HHC cells is approximately 3E11, as reported by Proteome Discoverer (PD), see supplemental Table S1. To determine to what amount of peptide this number corresponds, we followed the same analytical procedure with a known amount of a mix of 20 synthetic peptides. An integration value of 1E8 as reported by PD corresponds to 1 pmol. From this we determine the total eluted peptide amount to be 3 nmol (3E11/1E8).

The composition of the ligandome can also be represented in terms of relative intensity of the peptides, as shown in Figure. 6 (A). The intensities have been translated to copy number in Figure. 6 (B).
The 14,065 HLA-ligands are derived from 7,059 different proteins, a graphical representation is shown Figure 7 (A). In Figure 7 (B) the average protein length is plotted against the number of peptides/protein. The higher the protein mass, the higher the number of derived HLA-ligands. 61 proteins are even represented by 10 peptides or more on the cell surface, with the exceptional case of 29 and 41 peptides/protein. The intensity of the peptides derived from the same protein varies greatly, as presented in Table 2.

**Physicochemical properties of the peptides comprising the HLA-ligandome**

To check if peptide detection depended on the peptide physicochemical properties, with either one of the employed first dimension separation techniques, we studied the influence of isoelectric point and hydrophobicity. To study the influence of the isoelectric point of the peptide we calculated the theoretical pIs of all identified ligands. The calculated pIs ranged from 3.2 to 10.4. When the peptides were assigned to pI-bins as shown in Figure 8 (A), it was immediately clear that ligands with pI extremes migrate out of the pI-strip (since these were not detected with IEF as the 1st dimension separation technique, but in contrast were detected using SCX and C18 as the 1st dimension separation technique) and are, therefore, lost for detection using IEF. In addition, the actual pH at the low pH side of the strip is not 3, but almost 4, which accounts for extra losses experienced by application of IEF. Since peptides binding to HLA-B44 are on average more acidic, they are more prone to be lost in IEF. Next, we studied the influence of peptide hydrophobicity on the peptide’s chance to be detected. To this end we calculated the gravy index of all identified peptides and plotted them in hydrophobicity bins as shown in Figure 8 (B). As can be seen, the distribution of the peptides is evenly spread independent of the 1st dimension separation technique employed. The more hydrophilic peptides are slightly disfavored by RP-C18 as the first dimension separation technique. In summary, peptides with pI extremes might be lost during IEF, but, generally the majority of peptides was detected independent of the 1st dimension separation technique used.

**The overall quality of the presented HLA-ligandome data**

To illustrate that the majority of the peptides reported here was indeed derived from the expected HLA-molecules, we showed above that these peptides carry the correct anchor residues required for binding. Another way to verify the overall quality of the ligandome reported, is to check the predicted binding affinity for the HLA-molecule. In addition, the overall physicochemical characteristics of the peptides bound depend on the particular HLA-molecule. The HLA-binding affinity can be approximated theoretically by the NetMHC algorithm. In Figure 9, the hydrophobicity distribution for HLA-A2, HLA-B7 and HLA-B44 predicted peptides is plotted. The HLA-A2 presented peptides are centered around a GRAVY value of +1, those from HLA-B7 around 0 and those from HLA-B44 around -1.
In addition to the overall quality check as described above, we compared our data on the individual peptide level with the two main public sources of HLA-ligands, which are the SYFPEITHI-database and the Immune Epitope Database (IEDB). These databases are collections of ligands and/or epitopes that result from different immunological experiments on a variety of cell types, and generally not fed by proteomics-type of experiments but by individual immunological reports. To compare our peptide lists with the SYFPEITHI-database we selected the reported HLA-A2, HLA-B7 and HLA-B44 ligands and compared it to our eluted peptide sequences. In addition, we compared our results with the peptides presented in the papers by K.E. Scull et al. on B-cells [7], N. Hillen et al. on HLA-B44 supertypes on B-cells [8], D. Granados et al. on B-cells [9] and C. Bade-Doeding et al. on B-cells [10]. The results of this comparison are shown in supplemental Table S1.

We found 253 of the 784 ligands described in the SYFPEITHI database, and 72 out of 263 ligands described in the IEDB. Out of cumulative 1179 (8-11 mer) selected HLA-A2 ligands recently published by Scull et al. [7], 889 were found in our dataset.

Immunological and biological values

Below a number of immunologically relevant ligands are described. These include minor histocompatibility antigens and virus-derived epitopes.

**MiHA**

Minor histocompatibility antigens (MiHA) are peptides derived from polymorphic proteins and are relevant in allogeneic hematopoietic stem cell transplantation (HSCT) as targets for immunotherapy for the treatment of hematological malignancies [2, 13, 39-42].

To find and identify polymorphic peptides in the eluted peptides from B-LCL-HHC and B-LCL-JYpp65 cell lines the tandem mass spectra were matched with a dedicated database, HSPVdb [21]. We found 1439 polymorphic peptides of 8-11 mer length and mascot ion score >35, accounting for 10% of the number of peptides in the ligandome. 6 out of 16 MiHA and 4 allelic counterparts of known minor antigens, which could potentially be present on our cells, considering the SNP-typing of the B-LCL were found in our eluted dataset which was confirmed by matching the MS/MS spectra of the synthetic peptides with the eluted peptides. These peptides are listed in Table 1. It should be noted that 3 out 10 of the peptides listed have a BMI<35, but were correctly identified.

**Virus-derived peptides**

Since EBV-transformed cell lines were used in this study, our tandem mass spectra were matched with a specific database containing the genomic EBV information. Four EBV peptides were found in our data, see supplemental Table S3. The binding affinity of the newly identified peptides was predicted using NetMHC. Two peptides displayed high binding affinity for HLA-A2 and the other two showed a high binding affinity for
HLA-B7. The correct identification of these four peptides was confirmed by matching the tandem mass spectra of the eluted peptides with their synthetic counterparts. These four EBV-derived peptides have not been reported before.

The B-LCL-JY cell line used in this study was transduced with the CMVpp65 protein, which gave us the opportunity to study the presentation of CMVpp65-derived ligands. To identify pp65-derived ligands, we matched the tandem mass spectra with a specific database containing the pp65 sequence. We could find two known pp65-derived epitopes, see supplemental Table S3. One epitope is presented in HLA-A2, the other

**Figure 1.** Venn diagrams of identified HLA-ligands per separation. The number of identified unique peptides per mass spectrometric replicate for IEF (A), SCX (C) and RP-C18 (E) for cell line B-LCL JYpp65. IEF (B) and SCX (D) for cell line B-LCL HHC. The area of the circles corresponds with the number of peptides. Total, fractional and overlapping numbers of unique peptides are indicated.
epitope is presented in HLA-B7 [26]. Again the identification of these two epitopes, with BMIs of 28 and 12, was confirmed by matching the tandem mass spectra of the eluted peptides with their synthetic counterparts.

In summary, we identified four new EBV-derived ligands peptides and two known epitopes derived from the transduced CMVpp65 protein. The EBV peptides might be candidates for new EBV epitopes. It is important to note that low scoring peptides should not be discarded lightly in this field of application, without checking their functional relevance.

Independent immunological testing in combination with the synthesis of good candidates provides a good strategy to exclude false positive identifications without losing the immunologically relevant false negative identifications.

Together, the above clearly illustrates that our list of ligands harbours immunologically relevant peptides, suggesting that many more relevant peptides are in this list.

Post-translationally modified HLA ligands

The presentation of post-translationally modified peptides has been reported before [22, 24, 43], although not in large numbers. Therefore, we investigated our tandem mass spectra for the presence of phosphorylation of serine, threonine and tyrosine, or cysteinylation of cysteine residues, since both modifications have been
Aberrant phosphorylation has been implied in cancer. Inclusion of phosphorylation in our database matching process yielded 451 phosphopeptide hits of lengths 8 to 11 amino acids with a BMI>35 (phosphorylation on anchor positions was not allowed), of which 221 were estimated to be correctly identified, see supplemental Table S4 and discussion. Similarly, a cysteinylated cysteine as PTM was set as a modification in the database matching, which resulted in 1221 identified peptides with a cysteinylated cysteine (cysteinylation on anchor positions was not allowed). The peptides are listed in supplemental Table S5.

**Phosphorylated HLA ligands**

Inclusion of phosphorylation in the database matching process yielded 451 phosphopeptide hits of length 8 to 11 amino acids with a best mascot score of greater or equal to 35, as listed in supplemental Table S4. The distribution was as expected: 267 on serine, 154 on threonine and 30 on tyrosine. Most peptides were singly phosphorylated, 13 peptides were doubly phosphorylated.

To evaluate if these phosphopeptides were properly assigned we compared our results with previously reported phosphopeptides at Phosphosite.org ([http://www.phosphosite.org](http://www.phosphosite.org)). This yielded 72 hits. Phospho.ELM 8.3 ([http://phospho.elm.eu.org](http://phospho.elm.eu.org)) yielded 38 hits, all of which were also present in phosphosite.org. Next we used...
NetPhos 2.0 to predict phosphosites in the peptides, which resulted in 83 phosphosites, of which 28 overlapped with the 72 phosphosites detected with the two other sites mentioned above. In total 127 out of 451 phosphopeptides with either a known or predicted phosphosite were found.

As shown in Figure 7 proteins can be represented by several peptides in the HLA-ligandome. Here three proteins were chosen that are represented by 7, 5 and 5 peptides in HLA in B-LCL HHC. As can be seen these peptides can be presented by different alleles, have greatly differing hydrophobicity. Importantly, peptides derived from that same protein have greatly differing abundance. So, the abundance of a HLA-ligand derived from a certain protein does not simply correlate with its protein intensity.

To confirm the identification of the phosphorylated peptides in our dataset we chose to synthesize a selection of peptides and compare the synthetic phosphopeptides with their eluted counterparts. 26 peptides were selected, of which 9 peptides with a known phosphosite or predicted phosphosite and 17 peptides without a known or predicted phosphosite with varying phosphorylation position in the peptide sequence and different mascot ion scores. These were synthesized and the MS/MS spectra of these peptides were matched with their eluted counterparts. The 9 peptides with a known or predicted phosphosite appeared to be correctly identified.
Of the other 17 peptides without known of predicted phosphosite, 29% (5 out of 17) were found to be correctly identified.

First the 127 hits found in our study and which were also either known from previously identified proteomics studies or predicted, were extrapolated to be correctly identified, leaving (451-127)= 324 hits. Of these 324 hits 29\% (=94 hits) was extrapolated to be correct. Therefore, we estimate that 221 out of 451 (49\%) have been correctly identified.

Cysteinylation of HLA ligands

To find peptides with a cysteinylated cysteine as PTM in our eluted dataset we searched the dataset obtained from B-LCL-HHC and B-LCL-JYpp65 cell lines with variable modification on cysteinylation (+119 Da). We found 1221 identified peptides with cysteinylation, see supplemental Table S5. Since the binding affinity of the peptides probably does not change due to cysteinylation we used NetMHC server to check the binding affinity of these peptides. The vast majority was predicted to bind to either HLA-A2 (594), HLA-B7 (143) or HLA-B44 (294).

Figure 5. Binding motif of various HLA-A2 length variants. Icelogo plots of HLA-A2 binding peptides of 9-mer (N=3302), 12-mer (N=137), and 14-mer (N=17) length. Peptides of different length clearly display a similar binding motif.
DISCUSSION

The quality of the data

The quality of the elucidated ligandome was checked on three independent levels. First, by application of false discovery rates. Second, by consideration of chromatographic criteria in combination with the correct distribution over physical parameters. Third by the predicted binding strength. Additionally, the presence of biologically relevant peptides illustrates the good quality of our dataset, but these peptides will be discussed separately.

Mass spectrometry in HLA-ligandomics and choice of false discovery rates

HLA-presented peptides arise from the complex antigen processing process, involving many enzyme specificities. Therefore, enzyme restriction during the database matching process is not possible, leading to a large increase in database search space. The best way to reduce the number of falsely identified peptides is to apply a strict precursor mass tolerance of 2 ppm, which was achieved by including a SIM scan in the FTMS-measurements. Strict application of a tight FDR of 1%, or even 5%, as in trypsin-based proteomics experiments, leads to loss of a wealth of valuable peptides (false negatives) as we will illustrate below. To our opinion this is unacceptable in immunological workflows. We have chosen for a BMI>35, which we know from experience

Figure 6. Occupation of the ligandome. The intensity distribution of peptides constituting the ligandome is shown for B-LCL JYpp65. (A) The outer rim of the pie chart displays the percentage occupation of the HLA-ligandome and the number within each pie segment is the number of unique peptides in that segment. Of the total number of 6,500 identified peptides, the top 5 abundant proteins occupy 6% of all HLA molecules. The top 50 abundant peptides occupy 27% of all HLA-molecules, etc. Note that the lowest abundant 5,500 peptides occupy only 17% of all HLA-molecules. (B) The numbers in (A) have been converted to copy number, i.e. the number of ligands per cell. The outer rim of the pie chart displays the number of unique peptides and the number within each pie segment is the average copy number for the peptides in that segment. The top 5 abundant peptides are on average present with 16500 copies on the cell surface. Peptides 5-50 are on average present with 5700 copies, etc. The average copy number of the lower abundant peptides is 43, the very lower abundant peptides are present at copy numbers close to 1 or even less.
is a decent starting cut-off. Many known immunologically relevant peptide ligands are still found far below this limit. In supplemental Table S1, we listed peptides with a BMI>35, but also indicated below what FDR percentage the peptides are, after searching either the smaller SWISSProt (homo sapiens) or the larger IPIhuman database. We included 1%, 5% and 10% FDR. The results of these searches and the effect of the different search conditions is summarized in supplemental Table S2. A 10% FDR has e.g. been applied by Dudek et al. in a HLA class I-peptidomics-based study [44]. An FDR of even 37% has been used in a HLA-class II-based peptidomics study by Chornoguz et al. [45]. Importantly, in targeted immunological approaches, i.e. searching for ligands/epitopes or post-translationally modified epitopes derived from a specific antigen, we advise not to use a strict statistical cut-off score, be it a BMI or a false discovery rate during the first discovery phase. In the next step first round ‘hits’ should be tested independently, be it by MS/MS analysis of the synthetic candidate epitope or by a biological assay, e.g. a T cell activity assay. As illustrated by the BMI scores < 35 for 3 out of 10 validated MiHA (and 6 out of 10 with BMI<39), and both CMV peptides, with one CMV peptide having a BMI of 12 (!), low scoring peptides should be prevented from being assigned as false negatives. It is important to note that initially wrongly assigned peptides will immediately be filtered out by immunological follow-up experiments. Next to the MS-methodology available, additional non-MS checkpoints were applied to estimate the value of a peptide, see below.

Figure 7. Representation of proteins in the HLA-ligandome. (A)This figure depicts the number of proteins as a function of the number of peptides identified per protein. HLA-ligands derived from 7059 unique proteins are in our dataset. In our dataset about 50% of the proteins is represented by 1 peptide, 20% is represented by 2 peptides, 10% is represented by 3 peptides, and so on. As can be seen there is a gradual decrease in the number of peptides by which a protein is represented. Up to 41 peptides/protein are present in our dataset. (B) The protein length is depicted as a function of the number of peptides by which the protein is represented in HLA. The number of peptides in HLA of a protein correlates well with its protein mass.
Chromatography and physicochemical properties

Sample (pre)treatment might inadvertently favor peptides with certain physicochemical properties. Therefore, we checked the distribution of our peptides over the hydrophobicity scale (GRAVY index), see Figure 8 (B) and pH scale, see Figure 8 (A). The peptides were evenly distributed, so sample pretreatment has not favored peptides on the basis of their physicochemical properties. However, peptides at the pH extremes of the IEF strips were lost. It appeared that the real pH range of the strips is not 3-10, but instead are closer to pH 4-9, which explains their loss. The hydrophobicity distribution of the peptides is shown in Figure 9 for peptides eluted from each of the three alleles used in this study. The centre of the hydrophobicity distribution is clearly shifted to the hydrophobic side for HLA-A2, and shifted to the hydrophilic side for HLA-B44. The HLA-B7 peptides are at intermediate hydrophobicity. This is in perfect agreement with the general notion that HLA-A2 presented peptides are more hydrophobic [46], while HLA-B44 derived peptides are hydrophilic and HLA-B7 peptides have intermediate hydrophobicity [47-50]. The relative position of the hydrophobicity distributions supports the correct overall assignment of the ligandome.

In addition to the identification of the peptides we determined their relative intensity. From the shotgun nature of the experiments, peptides found in three experiments (the 835 peptides found Figure 2 (B) are expected to have a higher intensity than those peptides present in only one experiment. The results as shown in Figure 3 clearly illustrate that this is the case in our experiments.

Figure 8. HLA-ligandome physicochemical properties. (A) Distribution of HLA-peptides over the pH range as a function of the first dimension separation technique used. As can be seen IEF, SCX en RP-C18 do not discriminate for peptides based on their theoretical pH. The only exception is that peptides with low pH are clearly underrepresented in the IEF process. This is caused by the actual pH in the pH 3-10 strips, which in fact appear to be pH 4-10, as also independently checked with a pH-indicator after isoelectric focusing. (B) Distribution of HLA-peptides over the hydrophobicity range depending on the first dimension separation technique used. As can be seen IEF, SCX en RP-C18 do not discriminate for peptides based on their hydrophobicity. The above plots illustrate our HLA-ligandome is a good representation of the real HLA-ligandome.
Binding prediction

Because of the large number of peptides found, peptide identities cannot be verified by MS/MS of their synthetic counterpart. Alternatively, the NetMHC algorithm was used to predict binding affinity. As can be seen in supplemental Table S1 the large majority of our reported peptides has a low NetMHC score, i.e. is predicted to be a (strong) binder. Although at the NetMHC-website a score <50 is considered a strong binder and between 50 and 500 a weak binder, from our experience with known epitopes, we know this value is too strict.

From the mass spectral quality, the distribution of physicochemical parameters, the expected shotgun distribution, in which the overlapping regions in the Venn diagram represent the most abundant peaks and the excellent binding affinity we are convinced that the sample pretreatment and mass spectrometric analysis have not resulted in skewing of the identified ligandome as a whole. The fact that many known biologically relevant peptides are found is another strong indication of the high quality of our ligandome.

Ligandome characteristics

Our large dataset opens up opportunities to discuss a number of properties of the HLA-ligandome. It is important to realize that the peptides we report here are the survivors of the intracellular processing. Therefore, this listing is valuable for studies on antigen processing, to all in silico studies on ligand/epitope definition and the definition of antigen processing rules. In addition, it can be used to look up if peptides from specific protein antigens can survive the intracellular processing and are presented on the cell surface. Thus, it provides an indirect view of cellular processes. Many more peptides are probably presented in HLA-molecules, but may escape detection because of low abundance and/or a relatively low sensitivity in the electrospray ionization process.

The Venn diagrams of Figure. 2 still display relatively large regions unique to either 1st dimension separation technique, which shows that many peptides are just detected, or may still escape detection. Although the
ligandome presented here is the most comprehensive report to date, we interpret the occurrence of many
unique peptides in a single type of ‘precursor dimension’ as a strong indication that the HLA-ligandome must
contain many more peptides than we report. This notion is supported by our experience with 3- and 4-dimensional
separations before MS-analysis in the classical forward approach to find T cell epitopes. Every additional
chromatographic dimension yields peptides, not detected in ‘precursor’-dimensions, indicating that a significant
part of the presented HLA-ligandome still goes undetected. Based on the current listing of 14,000 peptides,
the many false negatives, i.e. correct peptides with a BMI<35 and the peptides going undetected, we estimate
the HLA-ligandome to comprise at least 3 times more peptides, i.e. at least 50,000 members.

**Binding motif and peptide length**

The ligandome contains a sizeable fraction of unusually long peptides for HLA class I binding. We found 5%
of the peptides to be longer than 11 amino acids. This is in line with the results of Scull et al. on a smaller set
of HLA-A2 peptides [7].

In supplemental Figure 1 the Icelogo graphs show the HLA groove can accommodate at least up to
14-mers. Peptides need to bulge out of the groove considerably to accommodate the extra amino acids between
the fixed anchors at position 2 and the C-terminus, as beautifully shown by crystallographic data [51].

<table>
<thead>
<tr>
<th>Gene name</th>
<th>HLA type</th>
<th>Peptide sequence</th>
<th>Eluted cell line</th>
<th>BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA-2V</td>
<td>A*02:01</td>
<td>YIGEVLVSV</td>
<td>B-LCL-JY, B-LCL-HHC</td>
<td>50</td>
</tr>
<tr>
<td>HA-8R</td>
<td>A*02:01</td>
<td>RTLDKVLEV</td>
<td>B-LCL-JY</td>
<td>34</td>
</tr>
<tr>
<td>LB-SSR1-1S</td>
<td>A*02:01</td>
<td>VLFRGGPRGSLAVA</td>
<td>B-LCL-HHC</td>
<td>29</td>
</tr>
<tr>
<td>SSR1-1L*</td>
<td>A*02:01</td>
<td>VLFRGGPRGLLAVA</td>
<td>B-LCL-JY</td>
<td>37</td>
</tr>
<tr>
<td>LB-ERAP1-1R</td>
<td>B*07:02</td>
<td>HPRQEQL</td>
<td>B-LCL-JY, B-LCL-HHC</td>
<td>36</td>
</tr>
<tr>
<td>LB-GEMIN4-1V</td>
<td>B*07:02</td>
<td>FPALRFVEV</td>
<td>B-LCL-JY</td>
<td>43</td>
</tr>
<tr>
<td>SMCY-B7</td>
<td>B*07:02</td>
<td>SPSVDKARIEL</td>
<td>B-LCL-JY</td>
<td>58</td>
</tr>
<tr>
<td>EBI3-1V*</td>
<td>B*07:02</td>
<td>RPRARYVVQV</td>
<td>B-LCL-JY, B-LCL-HHC</td>
<td>26</td>
</tr>
<tr>
<td>PDCD11-1L*</td>
<td>B*07:02</td>
<td>GPDDSKTLLCL</td>
<td>B-LCL-JY</td>
<td>39</td>
</tr>
<tr>
<td>ACC-2G *</td>
<td>B*44</td>
<td>KEFEDGIINW</td>
<td>B-LCL-HHC</td>
<td>81</td>
</tr>
</tbody>
</table>

* allelic counterpart of known MiHA
data on the peptide length distribution, Figure 4, underline the need to take into account longer peptides in the search for epitopes, both in practical MS-based strategies and in silico prediction approaches.

**Intensity, dynamic range and number of peptides of the ligandome**

The total peptide intensity (sum of all identified peptide intensities) of both cell lines was similar, and amounts to 3 nmol of total eluted peptide), which corresponds to approximately 3 ug of peptides. Our experience from previous elutions is that the amount of isolated HLA after immunoaffinity precipitation (IP) is 150-200 ng HLA/1E6 B-LCL-JY cells, which is in line with the independently reported amount of 150 ng/1E6 cells [52]. Therefore, assuming 150 ng/1E6 cells, a yield of 3 mg HLA/2E10 cells, corresponding to 60 ug of peptides was expected. Therefore, the overall work-up efficiency is about 5%. From the above the copy number (ligands per cell) can easily be calculated. These numbers are shown in supplemental Table 1. Intensity extremes, expressed as peak area values in PD, range from 4E9 to 1E5, so the dynamic range of the peptide repertoire is quite high. Considering that the lowest intensity ligands are most prone to go undetected, the actual dynamic range is probably even higher. This translates to a range from 24,000 copies/cell to (<)1 copy/cell. Clearly the chance of a protein being represented by a peptide in the ligandome varies enormously.

Of cells available in limited amount, e.g. DCs, the ligandomes have been studied of only (!) a few tens of millions of cells [53, 54]. The number of identified ligands, however, is low (a few hundreds). Considering the large dynamic range as discussed above, it is intrinsically impossible to find the relatively low intensity ligands using small cell populations for full ligandomics experiments. An inventory of the possible ligandome on a large scale followed by MRM [55] on the selected small population might be a fruitful approach. However, ‘more is better’ definitely seems to apply to HLA-ligandomics.

The composition of the ligandome in terms of relative intensity of the peptides is shown in Figure 6. Interestingly, it appears that in B-LCL JYpp65 the 5 most abundant peptides already ‘fill’ 6% of the HLA molecules. The top 50 abundant peptides ‘fill’ 27%, the top 300 ‘fill’ 59% and the top 1000 ‘fill’ 83%. Importantly, this leads to the observation that all other ligands, i.e. approximately 5,500 peptides, fill the final 17%, as shown in Figure 6 (A). The intensities have been translated to copy number in Figure 6 (B).

The 14,065 HLA-ligands are derived from 7,059 different proteins, a graphical representation is shown in Figure 7 (A). Half the proteins is represented by 1 peptide while the other proteins are represented by more peptides. A gradual decline in peptide number/protein is seen. In Figure 7 (B) the average protein length is plotted against the number of peptides/protein. 61 proteins are even represented by 10 peptides or more on the cell surface, with the exceptional case of 29 and 41 peptides/protein. The longer the protein, the more ligands are derived from it. It also appeared that the position of the ligands in their ‘precursor’ proteins is evenly distributed (data not shown). In addition, the intensities of the ligands derived from the same protein can differ
greatly, see Table 2 for a few examples, which makes it impossible to predict ligand copy number from overall protein copy number. This is in line with previous work on correlation between mRNA/protein expression and ligand density on the cell surface [56]. Furthermore, no correlation could be found between peptide intensity and binding strength to HLA. The former observations may seem logical, but these imply that there is no clear selection on the basis of protein properties during antigen processing. In fact, every part of the proteome seems equally suitable for sampling by the HLA-molecules. Selection of particular ligands takes place at another level, namely the binding affinity to the available HLA-molecules.

From our peptide listing it is tantalizing to speculate on the origin of HLA-presented ligands and the cellular processing machinery and distill a glimpse of the cellular protein processing landscape. A courageous effort based on a rather limited data set has been made to describe the presence of peptides in relation to their corresponding protein turnover in terms of protein stability, DRiPs and SLIPs [57]. Studies on the scale as presented here might be required to provide a better view at the origin of HLA-presented peptides. Protein turnover times, as in supplemental Table S1 cannot be correlated with HLA-ligand intensity. This lack of correlation might be seen as reinforcement of the DRiPs hypothesis [58]. Pulse chase experiments have shown that some HLA-presented peptides make it to the cell surface within 30 minutes of the pulse [57]. A general predictive description of the route from protein to its presented peptides in the HLA-ligandome is disturbed by the many underlying processes leading to antigen presentation, (transient) protein abundance, HLA-peptide complex stability and re-internalization of HLA-peptide complexes. Compartmentalization and multiple sources of peptides, and a subset of rapidly degraded polypeptides (RDPs) have been suggested to be needed to get a grip on the intricate antigen processing pathways [58]. However, the final result of the complex cellular protein machinery is the representation of virtually all proteins on the cell membrane. In this way all cellular proteins are under the scrutiny of T cell surveillance.

The biological value

Our reversed immunology approach based on the bona fide eluted ligandome was started to explore its possibilities for finding T cell epitopes. The above were global observations and considerations concerning the ligandome. Below the focus is on the immunological value of the data contained in our data set.

Comparison with sources of known epitopes

First we looked for known epitopes contained in the International Immuno Epitope Database, and the SYFPEITHI database for HLA-A2, HLA-B7 and HLA-B44. Comparison showed that 50% and 29% listed in the IEDB and SYFPEITHI database, respectively, were found in our ligandome, as indicated in supplemental Table S1. Considering that both the IEDB and the SYFPEITHI databases are a collection of ligands from many studies on a variety of cells, the biological value of our data set is clearly demonstrated. In addition,
there are probably many more epitopes in our list, which are to be explored in the field of immunology.

**MiHA**

The first category of biologically important T cell epitopes are the minor histocompatibility antigens, important in the treatment of hematological malignancies. Analysis of the data set, using our in house developed human short peptide variation database (HSPVdb) [21], led to the identification of approximately 1,400 polymorphic ligands. Among these were known MiHA, underlining the possibilities of our proteomics approach to find MiHA. From the set of 1400 polymorphic peptides 80 were selected for further study to investigate their potential for medical application. The details of the selection method and the immunological follow up will be described in detail elsewhere. The total length of all proteins in Refseq release 46, expressed in base pairs (bp), is 56,113,216 bp. The total number of non-synonymous SNPs and in/del within a coding sequence (CDS) is 296,901. Thus the chance that a nucleotide is polymorphous is 0.5%. Therefore, for a 10-mer peptide (30 nucleotides), the chance of not containing a SNP is \((0.995)^{30}\) is 0.86. Therefore, the chance of a 10-mer peptide containing at least one polymorphic amino acid is 14%. The number we experimentally found is 10%, so of the same order as the theoretical number. From our data there is no reason to suspect that polymorphic peptides represent a ‘special case’. The same conclusion was drawn in the recent paper by D. Granados et al. on B-cells [9].

**Virus-derived peptides**

The antigen presenting cells in this study were EBV-transformed, and B-LCL-JY was also CMVpp65 transduced. Therefore, the presence of EBV- and CMVpp65-derived peptides was checked in our ligandome, and the results listed in supplemental Table S3. Four EBV-derived ligands and 2 CMVpp65 derived epitopes were found. All identifications were checked by comparison of the tandem mass spectra to those of their synthetic counterparts. All four EBV-derived ligands were newly found.

We expected that most of the previously reported EBV epitopes would not be detected, since HLA-A2 and HLA-B7 restricted T cell clones isolated from EBV seropositive individuals and specific for the different EBV proteins BZLF-1, BRLF-1, and EBNA-3A are not reactive or low reactive against HLA-A2 and HLA-B7 positive EBV-LCLs, whereas the T cell clones are reactive when the genes encoding these proteins are additionally introduced into EBV-LCLs. These results indicate that the published EBV epitopes are not or hardly presented by EBV transformed B cells, and therefore it is very well possible that these peptides were not detected in our elutions from HLA of the EBV-LCLs. The reason why these epitopes are not efficiently processed and presented is most likely due to the fact that only a minority of the cells in the EBV transformed cell lines is in a lytic cell cycle state. The four epitopes described in the manuscript are derived from other EBV proteins: BGLF4, RPMS1 (derived from an alternative reading frame) and EBNA3B/C and HS4ENVGP
(the latter two both derived from the UTR), and are indicated in supplemental Table S3. CMVpp65 specific HLA-A2 restricted T cells (peptide NLVPMVATV (found)) and CMV-pp65 specific HLA-B7 restricted T cells (peptides RPHERNGFTV (found) and TRPVTGGGAM (not found)) efficiently recognize pp65 transduced EBV-LCLs, indicating that the three pp65 peptides are processed and presented in HLA-A2 and HLA-B7. Two out of the three peptides were detected in our study, whereas one peptide was not. The exact reason why we did not detect this particular peptide is unknown, but this peptide may have a low expression or may have been missed due to the shotgun nature of the experiment.

Together, the presence of the IEDB matching peptides, the MiHA, and the virus-derived peptides in our list proves the relevance of our peptide listing for immunological research. The fact that not all previously reported epitopes were found in our analysis illustrates that still a considerable fraction of the ligandome goes undetected.

Post-translationally modified ligands

Inclusion of phosphorylation and cysteinylation as a PTM in the database matching process yielded approximately 221 phosphorylated ligands and 1221 cysteinylated ligands. Previously 36 and 18 HLA-A2 and 8 HLA-B7 phosphorylated peptides were reported by Zarling et al. [22, 43] and Meyer et al., [24] respectively. 15 peptides of these overlap with our dataset. In summary, 2 out of 150 ligands in our study are phosphorylated. Many cysteinylated ligands were identified. The total number of cysteine-containing (modified and unmodified) peptides is 1386, which represents 9% of all peptides (15286). The frequency of cysteine in the Swissprot human database is 2%. Therefore, we calculate the chance of a peptide ligand containing a cysteine residue to be 15% and 13% for a 10-mer ($1-(0.98)^{10}=0.15$) or 9-mer ($1-(0.98)^{9}=0.13$) ligand respectively (we excluded the two anchor positions for this calculation because cysteine is not favorable at anchor positions). 9% of our peptides contain a cysteine residue. Scull et al. find a cysteine residue in 4% (50 out of 1179 unique 8-11 mer peptides) of their identified peptides [7]. The numbers indicate that we are close to the theoretically expected number of cysteine-containing HLA-ligands. Whether the cysteinylation has occurred in vivo or in vitro, the identified cysteine-containing peptides are true identifications. Since there is no suitable criterion for (de) selecting biological relevance of these cysteinylated peptides we do not wish to discard these true identifications. Both phosphorylation and cysteinylation have been shown to play a role in T cell recognition [23, 24, 43]. The number of phosphorylated HLA-ligands described here is by far the largest reported to date. Considering we did not apply a specific phosphopeptide-directed work up procedure, such as described in [59], we expect a considerably larger number of phosphorylated ligands to be presented. Other PTM, not included in this study, will certainly raise the number of post-translationally modified ligands.

In summary, our in-depth HLA-ligandome study enabled a detailed look at antigen processing and provided relevant ligands. From the presented data we estimate the ligandome to comprise at least 50,000 ligands.
Considering the above number and the average number of peptides per protein being between 2 to 3, the intracellular peptide pool generated by protein breakdown, represents most, if not all, proteins, which seems the perfect way to present the cellular state on the outside of the cell for immune surveillance. Overall, every part of the proteome seems equally suited for sampling by the HLA-molecules, although even the abundance of peptides from the same protein varies greatly. The final composition of the ligandome is determined by the enzymes involved, the transporter associated with antigen presentation (TAP), the location of a particular protein, and the relative contribution of DriPs for a given protein. The available HLA-molecules select “what fits them best”. Overall this process leads to ‘delegates’ from many proteins (5,000 found for B-LCL-HHC) in our set.

A rich and diverse repertoire of ligands is presented to the immune system, including a considerable number of post-translationally modified ligands, an additional 10% is derived from polymorphic peptides and 5% is >11 amino acids.

We have shown that the ligandome presented here can be a good starting point to solve a wealth of specific immunological questions. Our large list of peptides presented here fits in the efforts towards a human immunopeptidome project [25].

ACKNOWLEDGEMENT

We would like to dedicate this paper to Jack Leunissen, one of the first Dutch bioinformaticians, whom we had to say goodbye to in the course of this work. N. Dolezal and R. Cordfunke are thanked for providing the synthetic peptides. Dr. G. Janssen and F. Koning are thanked for critical reading of the manuscript. This research was made possible by the financial assistance of the Landsteiner Foundation for Blood Transfusion Research (LSBR0713).
Table 2. Peptide intensities derived from the same protein. Intensity of different HLA-presented peptides from some selected proteins.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Peptide sequence</th>
<th>Intensity</th>
<th>HLA allele</th>
<th>BMI</th>
<th>NetMHC (nM)</th>
<th>Gravy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoform 1 of Clathrin heavy chain 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NENSLFKSL</td>
<td>7.6E+7</td>
<td>B44</td>
<td>59</td>
<td>359</td>
<td>-0.622</td>
</tr>
<tr>
<td></td>
<td>NPASKVIAL</td>
<td>2.6E+7</td>
<td>B7</td>
<td>51</td>
<td>13</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>AEELFARKF</td>
<td>2.3E+7</td>
<td>B44</td>
<td>39</td>
<td>351</td>
<td>-0.267</td>
</tr>
<tr>
<td></td>
<td>EEAFAI FRKF</td>
<td>6.4E+6</td>
<td>B44</td>
<td>47</td>
<td>17</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>AETGQVQKI</td>
<td>5.3E+6</td>
<td>B44</td>
<td>41</td>
<td>671</td>
<td>-0.556</td>
</tr>
<tr>
<td></td>
<td>FVHD LVLYL</td>
<td>3.3E+6</td>
<td>A2</td>
<td>41</td>
<td>13</td>
<td>1.622</td>
</tr>
<tr>
<td></td>
<td>LLMLSPRL</td>
<td>1.1E+6</td>
<td>A2</td>
<td>43</td>
<td>16</td>
<td>1.6</td>
</tr>
<tr>
<td>Mediator of RNA polymerase II transcription subunit 24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SLVEQLTMV</td>
<td>2.0E+7</td>
<td>A2</td>
<td>70</td>
<td>12</td>
<td>1.044</td>
</tr>
<tr>
<td></td>
<td>SPTDRS MSSL</td>
<td>1.9E+7</td>
<td>B7</td>
<td>56</td>
<td>10</td>
<td>-0.782</td>
</tr>
<tr>
<td></td>
<td>KERWS DYQW</td>
<td>6.4E+6</td>
<td>B4</td>
<td>36</td>
<td>123</td>
<td>-2.533</td>
</tr>
<tr>
<td></td>
<td>AEQMHKTF</td>
<td>3.6E+6</td>
<td>B44</td>
<td>53</td>
<td>594</td>
<td>-0.967</td>
</tr>
<tr>
<td></td>
<td>SLMD PPGLAL</td>
<td>1.3E+6</td>
<td>A2</td>
<td>35</td>
<td>31</td>
<td>0.27</td>
</tr>
<tr>
<td>Isoform 1 of Dedicator of cytokinesis protein 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KMWEEA ISL</td>
<td>3.4E+8</td>
<td>A2</td>
<td>58</td>
<td>9</td>
<td>-0.067</td>
</tr>
<tr>
<td></td>
<td>SPRL SQFLQL</td>
<td>5.3E+7</td>
<td>B7</td>
<td>52</td>
<td>21</td>
<td>-0.109</td>
</tr>
<tr>
<td></td>
<td>IPG M V PIL</td>
<td>5.2E+7</td>
<td>B7</td>
<td>36</td>
<td>89</td>
<td>1.656</td>
</tr>
<tr>
<td></td>
<td>LEAY IQ QHF</td>
<td>4.3E+6</td>
<td>B44</td>
<td>39</td>
<td>145</td>
<td>-0.233</td>
</tr>
<tr>
<td></td>
<td>ILL QVAA LKYI</td>
<td>1.7E+6</td>
<td>A2</td>
<td>49</td>
<td>63</td>
<td>1.773</td>
</tr>
</tbody>
</table>
REFERENCES


SUPPLEMENTAL DATA

**Supplemental Table 1:** Summary of the number of identified unique 8-11 mer peptides after searching the indicated database with the indicated FDR and/or best mascot ion score >35, and the number of overlapping peptides with the listing based on the use of a best mascot ion score of >35.

<table>
<thead>
<tr>
<th>Databases</th>
<th>FDR</th>
<th>8-11 unique peptides</th>
<th>overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swissprot homo</td>
<td>1%</td>
<td>2480</td>
<td>2314</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>7540</td>
<td>6483</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>12918</td>
<td>9770</td>
</tr>
<tr>
<td>IPI human 3.87</td>
<td>1%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>3790</td>
<td>3453</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>7044</td>
<td>6053</td>
</tr>
<tr>
<td>mascot ion score ≥35</td>
<td></td>
<td></td>
<td>14065</td>
</tr>
</tbody>
</table>

**Supplemental Table 2:** Virus-derived peptides identified, and confirmed by MS/MS of their synthetic counterpart, in our dataset. UTR=untranslated region. ARF=alternative reading frame.

<table>
<thead>
<tr>
<th>Source</th>
<th>Gene name</th>
<th>Peptide sequence</th>
<th>BMI</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBV</td>
<td>EBNA3B/C (UTR)</td>
<td>ITAPLLPAV</td>
<td>54</td>
<td>HLA-A2</td>
</tr>
<tr>
<td>EBV</td>
<td>BGLF4</td>
<td>GLKDAVYFL</td>
<td>46</td>
<td>HLA-A2</td>
</tr>
<tr>
<td>EBV</td>
<td>HS4ENVGP (UTR)</td>
<td>SPLSPmARL</td>
<td>40</td>
<td>HLA-B7</td>
</tr>
<tr>
<td>EBV</td>
<td>RPMS1 (ARF)</td>
<td>APGGGERAL</td>
<td>45</td>
<td>HLA-B7</td>
</tr>
<tr>
<td>CMV</td>
<td>PP65</td>
<td>NLVPMVATV</td>
<td>28</td>
<td>HLA-A2</td>
</tr>
<tr>
<td>CMV</td>
<td>PP65</td>
<td>RHERNGFTV</td>
<td>12</td>
<td>HLA-B7</td>
</tr>
</tbody>
</table>