Toward a Definition of Self: Proteomic Evaluation of the Class I Peptide Repertoire[1]

Heather D. Hickman,* Angela D. Luis,* Rico Buchli,† Steven R. Few,* Muthuraman Sathiamurthy,* Rodney S. VanGundy,† Christopher F. Giberson,† and William H. Hildebrand[2]*†

MHC class I molecules present host- and pathogen-derived peptides for immune surveillance. Much attention is given to the search for viral and tumor nonself peptide epitopes, yet the question remains, "What is self?" Analyses of Edman motifs and of small sets of individual peptides suggest that the class I self repertoire consists of thousands of different peptides. However, there exists no systematic characterization of this self-peptide backdrop, causing the definition of class I-presented self to remain largely hypothetical. To better understand the breadth and nature of self proteins sampled by class I HLA, we sequenced >200 endogenously loaded HLA-B*1801 peptides from a human B cell line. Peptide-source proteins, ranging from actin-related protein 6 to zinc finger protein 147, possessed an assortment of biological and molecular functions. Major categories included binding proteins, catalytic proteins, and proteins involved in cell metabolism, growth, and maintenance. Genetically, peptides encoded by all chromosomes were presented. Statistical comparison of proteins presented by class I vs the human proteome provides empiric evidence that the range of proteins sampled by class I is relatively unbiased, with the exception of RNA-binding proteins that are over-represented in the class I peptide repertoire. These data show that, in this cell line, class I-presented self peptides represent a comprehensive and balanced summary of the proteomic content of the cell. Importantly, virus- and tumor-induced changes in virtually any cellular compartment or to any chromosome can be expected to be presented by class I molecules for immune recognition. ***The Journal of Immunology,*** **2004, 172: 2944–2952.**

Major histocompatibility complex class I molecules are heterotrimeric cell surface glycoproteins that convey intracellular fitness to immune effector cells (1, 2). Class I molecules are assembled in the endoplasmic reticulum, where they become loaded with endogenously synthesized host peptides before egress to the cell surface (3–5). Cells displaying class I with healthy host-derived peptide ligands remain unharmed, while cells exhibiting peptide ligands unique to diseased cells become targets for the immune response (6–8). Immune effector discrimination of diseased and healthy cells is therefore contingent upon class I presentation of short (8–13 aa) endogenous peptides.

Understanding the breadth of the proteome sampled by class I HLA is vital to a number of immunologic fields. For instance, the development of vaccines that target CTL to infected or neoplastic cells requires the identification of Ags or epitopes that distinguish the unhealthy cell. Are viral Ags that localize to the nucleus less likely to be presented to CTL than a viral protein that resides preferentially in the Golgi apparatus? Will an up-regulated tumor Ag on chromosome 20 be better presented than a putative tumor Ag on chromosome 7? Understanding the range of peptide presentation is also pertinent to clinical transplantation: can we expect minor histocompatibility Ags to be derived from allogeneic polymorphisms throughout the human genome, or are person-to-person differences in select proteins more subject to presentation? Finally, as have studies of nucleotide and amino acid sequences increased our knowledge of evolution, understanding the breadth of the proteome sampled will further expose the evolutionary pressures that have shaped the immune function of MHC class I molecules.

MHC class I biologists have theorized that no self protein precursors are excluded from the class I peptide-presentation pathway, although this hypothesis has been pieced together from independent peptide-characterization experiments for multiple class I alleles (9). A more complete understanding of the boundaries of peptide sampling has been delayed by the relatively laborious techniques associated with the sequencing of a large number of peptides from any single class I molecule (for examples, see Ref. 10–13). For instance, the most well-characterized class I molecule, HLA-A*0201, has ~110 documented endogenous ligands, most of which were identified in different laboratories using different techniques (14, 15). It cannot be ruled out that the small numbers of peptide-source proteins (i.e., the proteins from which the peptides are derived) identified in each study are not a result of methodological differences (cell line, lysis method, affinity purification Ab, peptide-separation method, and ligand-sequencing method). Thus, no individual study has systematically analyzed a large number of endogenously loaded class I peptide ligands to ascertain the actual extent of peptide sampling of the human proteome.

To explore the repertoire of self proteins presented as peptides by MHC class I molecules, we sequenced >200 peptides from HLA-B*1801 (the largest number of peptides reported from a single experiment). A total of 10 mg of the B*1801 protein was isolated from the class I-deficient 721.221 B cell line that has been extensively used for class I protein expression. Peptides were

*Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104; and †Pure Protein LLC, Oklahoma City, OK 73104

Received for publication October 10, 2003. Accepted for publication December 16, 2003.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] This work was supported by National Institutes of Health Contract NO1-AI-95360 (to W.H.H.). H.D.H. was supported by National Institutes of Health Training Grant T32AI07633.

[2] Address correspondence and reprint requests to Dr. William H. Hildebrand, Department of Microbiology and Immunology, 975 NE 10th Street, Oklahoma City, OK 73104. E-mail address: william-hildebrand@ouhsc.edu

0022-1767/04/$02.00

eluted from B*1801, individual peptide ligands were selected for sequencing by mass spectrometry (MS), and proteomic techniques were applied for data analysis. In contrast to the close sequence relatedness of the bound peptides, the source proteins for these B*1801-bound peptides were found in almost every compartment of the cell and represented the spectrum of biological and cellular functions. Statistical analysis of peptide-source proteins in the context of the unbiased human proteome revealed a preference for RNA- and nucleic acid-binding proteins, ribosomal constituents, and cellular chaperones. These data provide the first direct evidence that a single class I molecule can access proteins from a large portion of, if not the entire, human proteome and present them as a sequence-related set of self peptides.

Materials and Methods

Cell line and transfectants

Soluble HLA (sHLA)[3]-B*1801 was produced by transfection of the class I-negative EBV-transformed B-lymphoblastoid cell line 721.221 (16) with a PCR-truncated cDNA cloned into the vector PCDNA 3.1$^-$ (Invitrogen, Carlsbad, CA), as previously described (17).

Class I production and purification

sHLA-B*1801-producing transfectants were cultured in a CP-2500 Cell Pharm (Biovest International, Minneapolis, MN), and the sHLA-containing supernatant was collected. Upon completing a bioreactor run, sHLA complexes were affinity purified from the harvests obtained using W6/32 (18) coupled to a Sepharose 4B matrix (Amersham, Piscataway, NJ). Harvests were applied to the column using a peristaltic pump system (Amersham) with a speed of 5 ml/min at 4°C. After the column was extensively washed with PBS, bound sHLA molecules were eluted with 0.1 M glycine (pH 11.0) and immediately neutralized by addition of 1 M Tris-HCl, pH 7.0. Purified molecules were buffer exchanged with PBS at pH 7.2 and concentrated using 10-kDa cutoff Macrosep centrifugal concentrators (Pall Filtron, Northborough, MA).

Intact B*1801 molecules were brought to a final concentration of 10% acetic acid and heated to boiling for 10 min. Peptides were then purified by passage through a 3-kDa Microcon Microconcentrator (Millipore, Bedford, MA) before loading onto a Jupiter Proteo 4-μm C18 reversed-phase HPLC (RP-HPLC) column (Phenomenex, Torrance, CA).

Mass-spectrometric peptide sequencing and analysis

Fractionated peptides were examined on a QStar QTOF mass spectrometer (PerSceptive Sciex, Foster City, CA) equipped with a NanoSpray nano-ESI ionization source (Protana, Odense, Denmark). Individual peptides were selected for tandem MS (MSMS) fragmentation and sequence analysis at random; however, the most abundant peptides were generally fragmented first and produced the best sequence data. Selected ions were analyzed manually and by automated sequence assignment using the programs BioMultiview and MASCOT (19).

Proteomic data analysis

Peptide sequences were analyzed for their putative derivation through protein-protein and protein-translated database BLAST searches (20). Source proteins for the peptides were assigned the appropriate LocusLink ID number (21) before automated annotation using the program Database for Annotation, Visualization, and Integrated Discovery (DAVID) (22), resulting in standardized nomenclature for each of the peptide-source proteins. Cellular protein locations, molecular functions, and biological functions were assessed using the DAVID software utilizing the GOcharts function based upon the Gene Ontology Annotation (GOA) protein hierarchical classifications (23). Statistically over-represented peptide categories were determined using the EASE program (included in the DAVID software package) by calculating the probability of being presented as a peptide vs the entire human proteome using Fisher's Exact Test and the Bonferroni correction for multiple queries. Genomic locations of genes corresponding to peptide-source proteins were compiled using gene location, as determined by LocusLink ID.

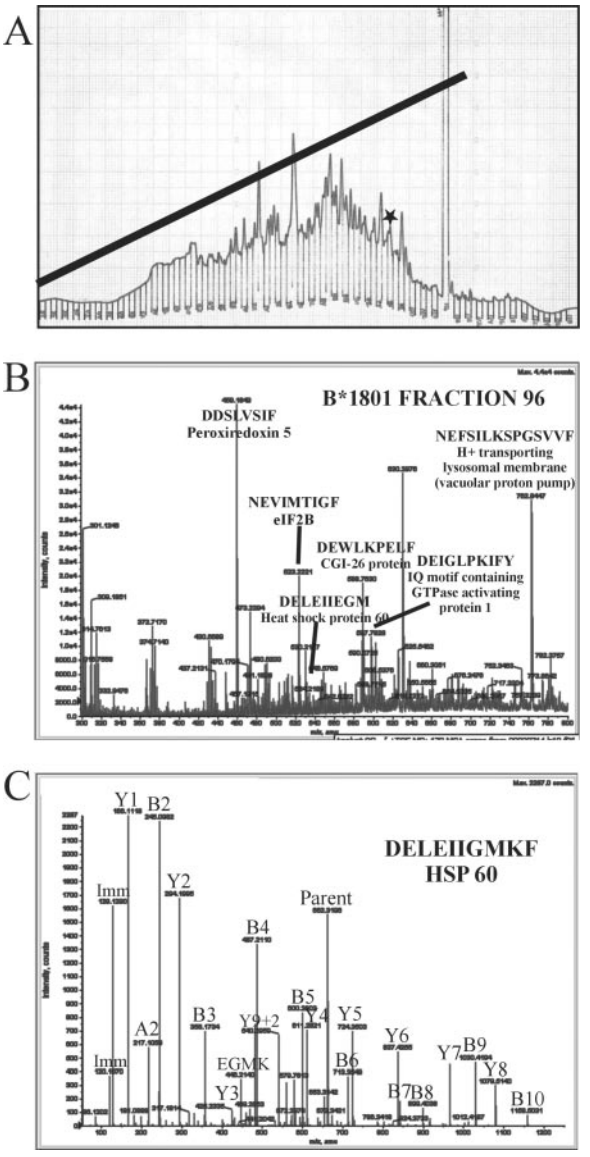

FIGURE 1. B*1801 peptide-sequencing strategy. *A*, B*1801 peptides were separated into fractions by RP-HPLC. A star indicates fraction 96, shown in *B*. *B*, Abundant peaks in each fraction (as determined by Nano-ESI MS) were selected for sequence determination. Fraction 96 is shown as a representative mass spectroscopy spectrum. *C*, Selected ions subjected to MSMS fragmentation and the spectra interpreted to yield sequence information.

A text file of all LocusLink ID numbers for the peptide-source proteins reported in this work is available in the supplemental material;[4] these are provided for independent data analyses.

Results

*Mass-spectrometric peptide sequencing identifies >200 B*1801 peptides*

Using sHLA secretion from hollow fiber bioreactors as previously described (17), we produced ~50 mg of B*1801, a molecule with only one peptide sequence described (24), from a single 4-wk-long experiment. Approximately 10 mg of B*1801 (roughly 500 μg of final peptide weight) was affinity purified after production in the EBV-transformed B cell line 721.221 (16). The peptide/MHC complexes were further subjected to acid denaturation to remove

[3] Abbreviations used in this paper: sHLA, soluble HLA; MS, mass spectrometry; MSMS, tandem ms; RP-HPLC, reversed-phase HPLC.

[4] The on-line version of this article contains supplemental material.

Table I. *B*1801 peptides*

LocusLink ID	Gene Name	Peptide Sequence
159	Adenylosuccinate synthase	DELQIPVKWI
211	Aminolevulinate, δ-, synthase 1	HEFGATTF
417	ADP-ribosyltransferase 1	EEVLIPPF
471	5-Aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	DENEVRTLF
595	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	EEEVFPLAM
608	TNFR superfamily, member 17	DEIILPRGL
639	PR domain-containing 1, with ZNF domain	EDFLKASLAY
701	BUB1 budding uninhibited by benzimidazoles 1 homologβ (yeast)	EEYEARENF
740	Mitochondrial ribosomal protein L49	DEYQFVERL
811	Calreticulin	DEFTHLYTL
908	Chaperonin-containing TCP1, subunit 6A (ζ1)	ALQFLEEVKV
1048	Carcinoembryonic Ag-related cell adhesion molecule 5	VFDKDAVAF
1105	Chromodomain helicase DNA-binding protein 1	FSDLESDE
1105	Chromodomain helicase DNA-binding protein 1	EEFETIERF
1355	Cyclooxygenase 15 homolog, cytochrome *c* oxidase assembly protein (yeast)	TEFKFIWY
1491	Cystathionase (cystathionine γ-lyase)	SEFGLKISF
1612	Death-associated protein kinase 1	ENKTDVILI
1653	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 1	DFFDGKVTY
1770	Dynein, axonemal, heavy polypeptide 9	DEFNIPELM
1915	Eukaryotic translation elongation factor 1 α1	FEKEAAEM
1915	Eukaryotic translation elongation factor 1 α1	TEVKSVEM
1915	Eukaryotic translation elongation factor 1 α1	SGKKLEDGPKF
1915	Eukaryotic translation elongation factor 1 α1	KLEDGPKF
1974	Eukaryotic translation initiation factor 4A, isoform 2	DEMLSRGF
1975	Eukaryotic translation initiation factor 4B	EENPASKF
2081	Endoplasmic reticulum to nucleus signaling 1	DEHPNVIRY
2222	Farnesyl-diphosphate farnesyltransferase 1	NELITNAL
3032	Hydroxyacyl-coenzyme A dehydrogenase/3-ketoacyl-coenzyme A thiolase/enoyl-coenzyme A hydratase (trifunctional protein)	DIDAFEFHEAF
3094	Histidine triad nucleotide-binding protein 1	DESLLGHLMIV
3122	MHC, class II, DRα	EEFGRFASF
3183	Heterogeneous nuclear ribonucleoprotein C (C1/C2)	VEAIFSKY
3187	Heterogeneous nuclear ribonucleoprotein H1 (H)	YEHRYVELF
3275	HMT1 hnRNP methyltransferase-like 1 (*Saccharomyces cerevisiae*)	DEVRTLTY
3276	HMT1 hnRNP methyltransferase-like 2 (*S. cerevisiae*)	DEVRTLTY
3315	Heat shock 27-kDa protein 1	NEITIPVTF
3317	Heat shock 27-kDa protein-like 1	GEDIVADSV
3320	Heat shock 90-kDa protein 1, α	EEVETFAF
3329	Heat shock 60-kDa protein 1 (chaperonin)	DELEIIEGM
3329	Heat shock 60-kDa protein 1 (chaperonin)	DELEIIEGMKF
3638	Insulin-induced gene 1	EEVIATIF
3801	Kinesin family member C3	SVELGPGLR
3929	LPS-binding protein	EEHNKMVY
4065	Lymphocyte Ag 75	DEIMLPSF
4130	Microtubule-associated protein 1A	NEAVKQQDKAL
4172	MCM3 minichromosome maintenance-deficient 3 (*S. cerevisiae*)	NAFEELVAF
4193	Mdm2, transformed 3T3 cell double minute 2, p53-binding protein	DEVYQVTVY
4261	MHC class II *trans* activator	DEVFSHIL
4436	mutS homolog 2, colon cancer, nonpolyposis type 1 (*Escherichia coli*)	HEFLVKPSF
4600	Myxovirus (influenza virus) resistance 2 (mouse)	FEIIVHQY
4673	Nucleosome assembly protein 1-like 1	NEVLTKTY
4683	Nijmegen breakage syndrome 1 (nibrin)	DEIPVLTL
4776	NF-AT, cytoplasmic, calcineurin-dependent 4	KVLEMTLLP
5042	Poly(A)-binding protein, cytoplasmic 3	DERLKDLF
5111	Proliferating cell nuclear Ag	YLAPKIEDEEGS
5307	Paired-like homeodomain transcription factor 1	EDVYAAGYSY
5394	Polymyositis/scleroderma autoantigen 2, 100 kDa	DEYDFYRSF
5478	Peptidylprolyl isomerase A (cyclophilin A)	DENFILKH
5591	Protein kinase, DNA-activated, catalytic polypeptide	NELKFYQFG
5591	Protein kinase, DNA-activated, catalytic polypeptide	DEFKIGELF
5690	Proteasome (prosome, macropain) subunit, β type 2	DEHEGPAL
5690	Proteasome (prosome, macropain) subunit, β type 2	DEHEGPALY
5708	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 2	DETELKDTY
5719	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 13	NEVELLVM
5902	RAN-binding protein 1	DEEELFKM
5931	Retinoblastoma-binding protein 7	EERVINEEY
6125	Ribosomal protein L5	DEDAYKKQF
6129	Ribosomal protein L7	MEDLIHEIY
6154	Ribosomal protein L26	DEVQVVRGHY
6154	Ribosomal protein L26	QVVRGHY

(*Table continues*)

Table I. *Continued*

LocusLink ID	Gene Name	Peptide Sequence
6154	Ribosomal protein L26	VQVVRGHY
6160	Ribosomal protein L31	NEVVTREY
6188	Ribosomal protein S3	NEFLTREL
6201	Ribosomal protein S7	DEFESGISQAL
6222	Ribosomal protein S18	VERVITIM
6228	Ribosomal protein S23	DEVLVAGF
6257	Retinoid X receptor, β	NELLIASF
6543	Solute carrier family 8 (sodium-calcium exchanger), member 2	VIPAGESRKI
6635	Small nuclear ribonucleoprotein polypeptide E	DEYMNLVL
6646	Sterol *O*-acyltransferase (acyl-coenzyme A: cholesterol acyltransferase) 1	DEGRLVLEF
6749	Structure-specific recognition protein 1	DEISFVNF
6772	STAT1, 91 kDa	EELEQKYTY
6772	STAT1, 91 kDa	SEVLSWQF
6774	STAT3 (acute-phase response factor)	EELQQKVSY
6897	Threonyl-tRNA synthetase	SKQAEEEF
7094	Talin 1	DEYSLVREL
7205	Thyroid hormone receptor interactor 6	LDAEIDLL
7332	Ubiquitin-conjugating enzyme E2L 3	IEINFPAEY
7353	Ubiquitin fusion degradation 1-like	YEFKLGKITF
7453	Tryptophanyl-tRNA synthetase	FDINKTFIF
7453	Tryptophanyl-tRNA synthetase	IEVLQPLI
7520	Double-strand-break rejoining; Ku autoantigen, 80 kDa	DEIALVLF
7706	Zinc finger protein 147 (estrogen-responsive finger protein)	DEFEFLEKA
8260	ARD1 homolog, *N*-acetyltransferase (*S. cerevisiae*)	DENGKIVGY
8512	Mannose-binding lectin (protein A) 1, pseudogene 1	VEVQLPELF
8723	Sorting nexin 4	SEFELLRSY
8826	IQ motif containing GTPase-activating protein 1	DEIGLPKIFY
8892	Eukaryotic translation initiation factor 2B, subunit 2β, 39 kDa	NEVIMTIGF
9263	Serine/Threonine kinase 17a (apoptosis inducing)	EELIVVTSY
9267	Pleckstrin homology, Sec7, and coiled-coil domains 1 (cytohesin 1) homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain	DEFNIQVL
9709	Member 1	DEINRDWLDWTY
9770	Ras association (RalGDS/AF-6) domain family 2	DEFIVEGL

bound peptides. Eluted peptides were purified by ultrafiltration and separated based on hydrophobicity by RP-HPLC to reduce complexity (Fig. 1*A*). Once separated, individual fractions were sprayed for analysis on a Q-Star Q-TOF mass spectrometer (Fig. 1*B*). Ions were then selected for sequence derivation through MSMS fragmentation (Fig. 1*C*) and automated and de novo sequence interpretation. Although peptides were chosen at random, abundant peptides produced better sequence data; more sequences were therefore identified from peptides with high to intermediate copy number than those present at low quantities.

Table II. *Amino acid frequency at specific positions of B*1801 peptides*[a]

	Residue Position in Peptide									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Amino Acid										
A	1	2	5	4	7	6	7	5	4	11
C	0	0	0	1	0	0	1	0	0	0
D	**38**	5	4	4	2	3	4	2	1	0
E	14	**81**	5	14	4	5	17	5	3	4
F	3	0	18	4	11	5	3	16	**31**	7
G	1	1	2	4	5	3	5	5	2	7
H	2	0	4	0	1	4	3	1	3	0
I	3	2	9	11	16	6	10	4	5	11
K	2	1	3	5	5	7	7	3	2	7
L	3	3	12	17	**20**	11	8	17	9	7
M	4	0	3	2	2	1	2	4	4	0
N	12	1	5	5	3	3	1	2	1	0
P	0	0	1	7	1	14	4	1	2	0
Q	1	1	3	6	3	2	4	3	2	0
R	1	1	3	4	2	5	4	5	1	0
S	7	1	3	4	4	7	5	10	2	4
T	3	2	2	3	7	4	7	5	2	0
V	6	2	18	6	8	12	11	4	3	11
W	0	0	2	1	1	1	1	0	1	7
Y	4	0	3	2	1	3	2	11	**11**	**22**

[a] Frequencies are indicated as percentage of total peptides possessing indicated residue. Bolded numbers indicate percentage above 20. Underlined numbers indicate peptide anchor preferences.

Using this approach, >200 individual HLA-B*1801 ligand sequences were obtained (100 representative ligand sequences are shown in Table I; the complete peptide list is present in supplemental data). A BLAST search (20) was performed on each peptide to determine the precise source protein from which the peptide was derived; each protein was assigned a LocusLink ID number (21), if available (Table I and supplemental data). B*1801 peptides were derived from a wide variety of proteins, alphabetically from actin-related protein 6 to zinc finger protein 147. In total, HLA-B*1801 peptides were found from 189 unique source proteins; 9 of the proteins possessed more than one identified peptide. When two or more peptides were present from the same protein, the peptides were exclusively derivatives of each other one-third of the time (for example, DEHEGPAL and DEHEGPALY from the proteasome β2 protein) (Table I and supplemental data). Thus, in most cases, multiple peptides were derived from separate areas of the protein, such as three of the four peptides derived from eukaryotic translation elongation factor 1α (Table I and supplemental data).

The previously reported B18 motif (25), as well as our motif data, specified an E at P2 (position 2) of bound peptides. As a secondary confirmation of the integrity of B*1801 peptides, we calculated the percentage of individual amino acid occurrence at each position of the eluted peptide (Table II). In accord with previously published data, B*1801-eluted peptides had a strong preference for E at P2. Additionally, we detected a preference for an aromatic F or Y at the C termini of the peptides; this correlates with a previously reported B18 peptide identified through synthetic peptide-binding assays (26). Moreover, many of these peptides initiated with a weak consensus sequence of DEL, DEV, or DEF. Thus, in terms of peptide sequence, the B*1801 ligands identified in this study represent a highly similar set of naturally loaded peptides that retain previously denoted B18 characteristics (25).

Source proteins for peptides are distributed throughout the cytoplasm, nucleus, and membranes of the cell

We next applied proteomic analysis to determine the locations of the peptide-source proteins within the cell. Source proteins were entered into the DAVID program (22), where they were sorted by cellular component according to their GOA classification (23). As expected, most (90%) of the peptide-source proteins were classified as intracellular, the main location for the generation and loading of MHC class I peptides (Fig. 2). Interestingly, many of the source proteins were not derived from the cytoplasm, the cellular

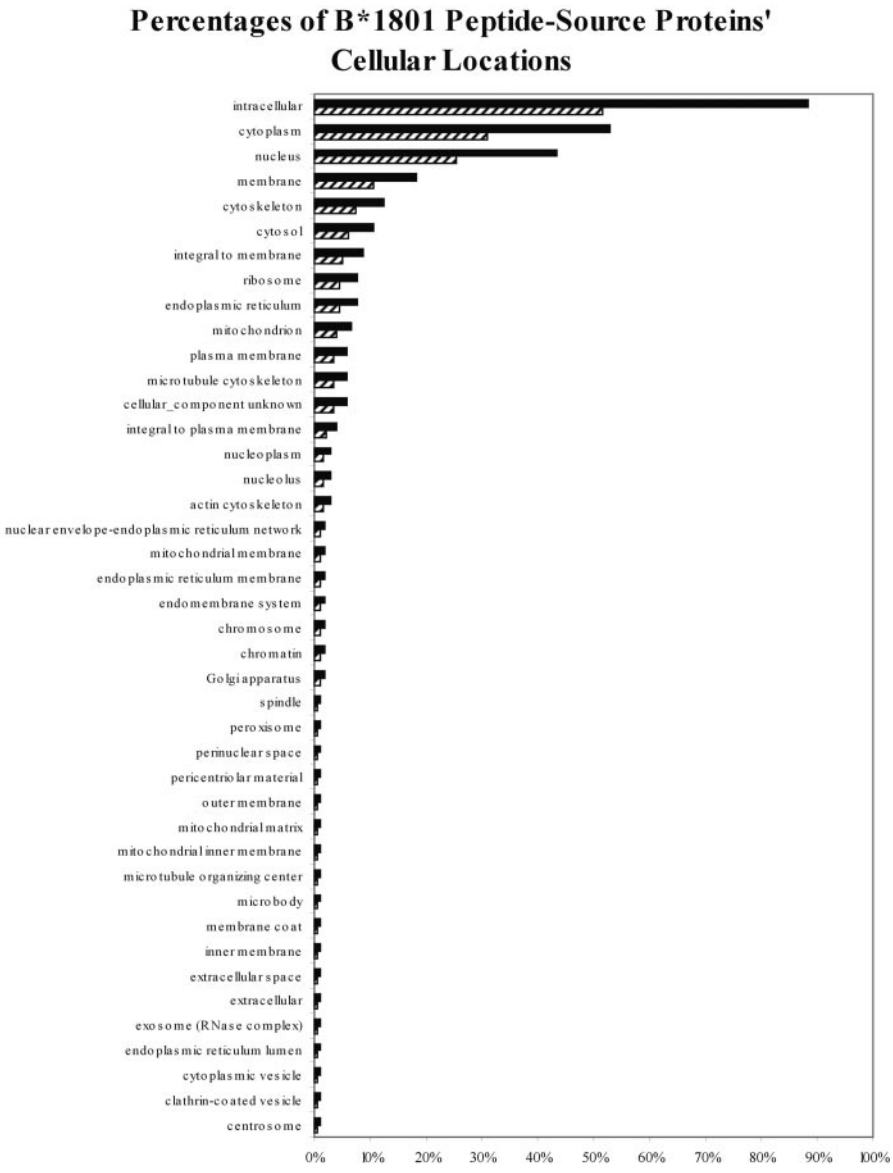

FIGURE 2. Source proteins for B*1801 peptides were distributed throughout the cellular compartments and membranes. ▨, Represents the percentage of total source proteins found in each compartment; ■, represents the percentage of proteins within a GOA-classified cellular component.

locale of peptide generation by proteasomal cleavage (27). Approximately 53% of peptide-source proteins are cytoplasmic proteins, while 43% are nuclear and 18% are membrane bound (note that some of the source proteins could be found in both the cytoplasm and nucleus, resulting in a total higher than 100%). Additionally, many of the peptide-source proteins had more precise location annotation: 8% ribosome, 8% endoplasmic reticulum, 7% mitochondrion, 6% plasma membrane, 3% nucleolus, 2% chromosome, and 2% Golgi. We found no peptides derived from proteins resident to lysosomal or endosomal compartments. Therefore, MHC class I peptides can be derived from proteins resident to almost every compartment in the cell and are not particularly biased toward the cytoplasmic compartment.

Source proteins are largely involved in normal cellular metabolism and maintenance

B*1801 peptide-source proteins were next evaluated for their biological and molecular functions, again according to GOA classification using the program DAVID. Although source proteins possessed multiple biological functions, a majority of the source proteins fell into two categories: 74% were involved in cellular metabolism, while 38% functioned in cell growth and maintenance (Fig. 3*A*). Minor categories of source-protein biological function included cell communication (16%), cell stress response (12%), and cell external stimuli response (10%). As seen with biological functions, a majority of the peptide-source proteins possessed two major molecular functions: 64% had binding activity, while 46% possessed catalytic activity (Fig. 3*B*). Aside from the two major molecular functions of source proteins, identified proteins were also involved in transcriptional regulation (11%), signal transduction (8%), molecular or solute transport (9%), and protein chaperoning (6%). Thus, at any time, a majority of the proteins serving as sources for MHC class I peptides are involved in normal cellular metabolism and maintenance through RNA-, DNA-, or protein-binding activities.

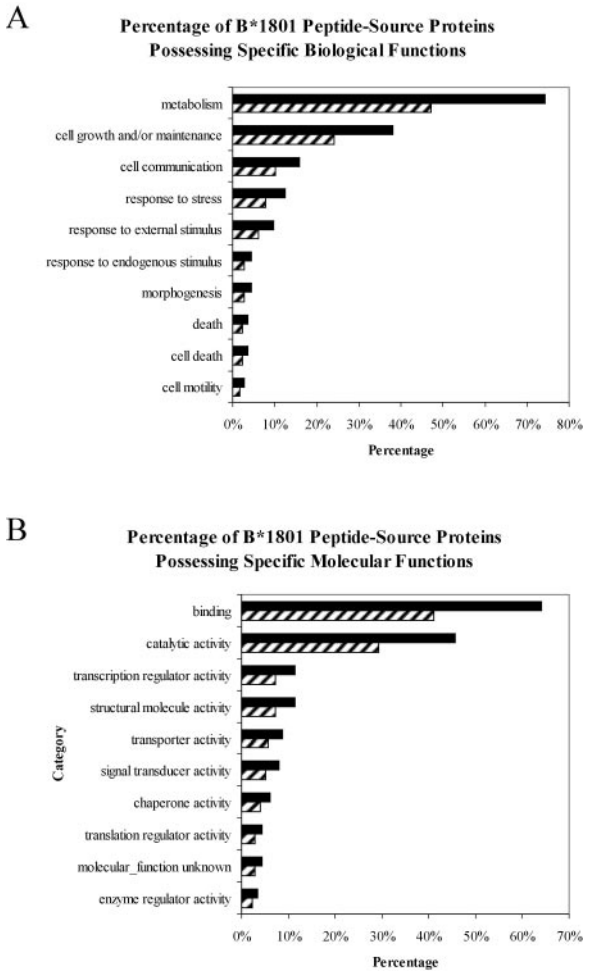

FIGURE 3. B*1801 peptide-source proteins were involved in cellular metabolism and possessed binding activity. *A*, Percentage of source proteins in each GOA category of biological function. *B*, Percentage of source proteins in each GOA category of molecular function. ■, Indicates percentage of classified proteins; ▨, indicates percentage of total proteins.

Class I peptide-source proteins are encoded by every chromosome

Each source protein was next assigned to a chromosomal location according to the location of the gene encoding it (Fig. 4). Multiple peptides were derived from source proteins on each chromosome; the largest number of source proteins (19) was encoded on chromosome 2, while the lowest number (2) was encoded by genes on chromosome 18. Although no peptides were found from the Y chromosome, this is to be expected from the 721.221 cell line that is of female etiology (16). Thus, no chromosome appears to be excluded from generating products for MHC class I peptides.

*RNA-binding proteins are preferentially presented by B*1801*

To compare the presentation of peptides with representation by the proteome (i.e., is the self repertoire a result of the specificity of class I and its loading pathway or a general result of protein abundance in a particular category), we performed statistical analysis comparing the representation of annotated source proteins of peptides with the representation in the annotated human proteome. Table III denotes the 10 most over-represented categories of B*1801 peptide-source proteins in terms of cellular location (cellular component) and biological and molecular function. As expected, intracellular proteins were significantly over-represented as class I peptides (as were the intracellular categories of cytoplasmic and nuclear). In terms of biological function, a majority of the over-represented proteins were involved in protein and macromolecule biosynthesis. Perhaps most interestingly, in the category of molecular function, RNA-binding proteins were significantly over-represented as peptide-source proteins when compared with the whole proteome. RNA-binding proteins therefore serve as a rich source for MHC class I ligands.

Discussion

The MHC class I self peptide repertoire is often discussed, but has never been systematically defined (9). Although a growing number of class I peptides is being added to the collective class I repertoire, these additions most likely represent a minority of those that remain to be revealed. Even with the gap in knowledge of class I

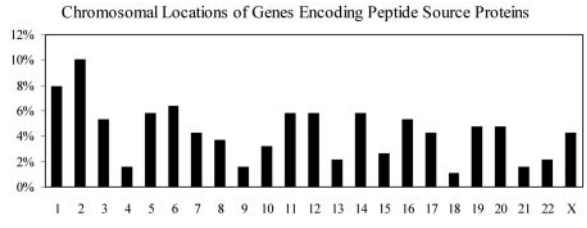

FIGURE 4. B*1801 peptide-source proteins were encoded on every chromosome. Percentages of unique source protein encoded on each chromosome are shown.

Table III. *B*1801 peptides were preferentially derived from RNA-binding proteins*[a]

GO Classification	Fisher's Exact Probability	Bonferroni's Correction
Cellular component		
Intracellular	0.00000000001	0.000000009
Cytoplasm	0.000002	0.002
Ribonucleoprotein complex	0.000008	0.006
Cytosolic ribosome (sensu Eukarya)	0.00001	0.01
Nucleus	0.00002	0.01
Cell	0.00003	0.02
Cytosol	0.00005	0.04
Large ribosomal subunit	0.0004	0.3
Small nuclear ribonucleoprotein complex	0.0005	0.4
Ribosome	0.001	0.7
Biological function		
Macromolecule biosynthesis	0.000003	0.002
Protein biosynthesis	0.000003	0.002
Protein metabolism	0.000004	0.003
Biosynthesis	0.00002	0.01
Physiological processes	0.00005	0.03
Regulation of translation	0.0001	0.08
mRNA splicing	0.0002	0.1
Metabolism	0.0002	0.2
RNA splicing	0.0002	0.2
RNA metabolism	0.0004	0.3
Molecular function		
RNA binding	0.000000001	0.0000008
Nucleic acid binding	0.00003	0.02
ATP binding	0.00008	0.06
Adenyl nucleotide binding	0.0001	0.07
Purine nucleotide binding	0.0001	0.09
Nucleotide binding	0.0001	0.1
Hemopoietin/IFN class cytokine receptor signal	0.0003	0.3
ATP-dependent helicase activity	0.0004	0.3
Chaperone activity	0.0005	0.4
Heat shock protein activity	0.0007	0.5

[a] A statistical comparison of peptide-source proteins vs the human proteome was performed using the program EASE. The 10 most over-represented subcategories of cellular component, biological function, and molecular function are shown.

ligands, tumor and vaccine immunologists strive to identify nonself and differentially expressed self to elicit an immune response against infected or tumor cells. It has yet to be determined, however, how distinct these peptides will be in the context of the normal cellular peptide environment. With the aim of providing a more complete understanding of self, we have performed a systematic proteomic analysis of class I peptides bound by the molecule B*1801.

Using sHLA as a means to gather a large amount of class I molecules and their cognate peptides, we eluted peptides endogenously bound in the B cell line 721.221 from ~10 mg of sHLA-B*1801 complexes and sequenced them randomly by mass spectrometry. Peptides were largely concordant with the previously identified B18 anchor residue of P2 E (25) and also possessed an aromatic C terminus. A majority of the peptides were nonamers, with several longer peptides also identified possessing canonical anchors. Individual B*1801 peptide ligands sequenced in this study therefore fit with existing B18 knowledge in terms of length and amino acid preferences.

The most striking observation obtained through analyses of peptide-source proteins was that peptides sampled by class I are relatively unbiased when compared with the human proteome. Statistically, the single most over-represented category in terms of class I peptide presentation are intracellular proteins, which reflects the main function of class I in the presentation of intracellular peptides. The statistical significance of other proteome categories sampled drops precipitously, except in the case of RNA-binding proteins, which class I molecules appear to be exceptionally adroit at presenting. It remains to be tested whether RNA-binding protein-derived peptides are overabundant due to a propensity to load into class I molecules or simply because they are among the most abundant proteins in the eukaryotic cell (28). In either instance, with the current exception of RNA-binding proteins, these data provide the first experimental evidence that class I molecules purified from B cells bind and present an accurate reflection of the human proteome.

Although the peptides identified were closely allied in amino acid sequence and length, genetically they were encoded by genes dispersed throughout the genome, being located to every chromosome without apparent preference. Furthermore, no telomeric or centromeric bias could be detected (data not shown). Immune surveillance mechanisms that review class I-presented self, at least through B*1801, can therefore monitor gene products irrespective of their chromosomal locale. Although they directly sample the proteome, these data indicate that class I molecules indirectly provide an unbiased view of the genome.

Current knowledge dictates that most class I peptides are created in the cytoplasm by the proteasome, although mechanisms for peptide generation and loading outside the cytoplasm have been identified, perhaps most notably in dendritic cells (4, 29–33). Theoretically, in the B cell line used in this study, peptides could be generated from proteins in cellular compartments accessory to the cytoplasm in a number of ways. First, peptide-source proteins could be retrieved from their cellular compartment for cytosolic degradation as a normal feature of cellular metabolism. Alternatively, source proteins may be degraded in their resident compartment; this may be especially relevant for nuclear protein degradation by the many nuclear proteasomes (29). Peptides generated in

cellular compartments such as the nucleus can freely diffuse from nuclear pores and enter the class I-processing pathway in the cytosol (34). Finally, Yewdell and colleagues (35, 36) have proposed that a large portion of newly generated class I products is derived from defective ribosomal products; along the same line, newly synthesized proteins have been identified as the major substrate for TAP, and thus the major source of class I peptides (37). Presentation of newly synthesized normal or defective proteins would occur in the cytoplasm before the egress of source proteins to ultimate cellular locations. Whatever the case, class I molecules appear to be extraordinarily adapted to present the entire cellular complement of proteins.

The equitable sampling of class I peptides from genetically, functionally, and compartmentally diverse proteins is most likely necessitated to comprehensively reflect the collective health of the cell. Genetically, presentation of all chromosomal products may allow NK or CTL detection of newly arising cancerous transformations regardless of location. Functionally, intracellular pathogens modify and usurp a wide range of host metabolic cycles (38, 39); distinct changes in proteins in multiple compartments of the cell may be necessary to report complex host-pathogen interactions to immune surveillance systems. Likewise, presentation of the full spectrum of the proteome as proteins are generated from ribosomes may allow early detection of replicating viral invaders. This comprehensive peptide presentation is certainly a more attractive mechanism of immune supervision than compartmentalized or biased presentation. Certainly, it will be important to compare these findings with those generated in other cell lines, such as professional APCs.

In summary, we have demonstrated that B*1801 samples an enormously complex proteome with great efficiency. Transcriptional regulators, chaperones, membrane proteins, and stress-response factors are all available for review by cellular immune mechanisms. As yet, we do not fully understand the mechanisms that enable class I HLA molecules to access such a vast array of protein products, nor do we fully understand the shaping of the innate and adaptive immune responses by this comprehensive view of the host proteome. It is emerging that class I HLA-presented self can both trigger and modulate immune responsiveness (40–42), and our understanding of the contribution of self protein presentation to human immunity will enhance the detection of disease-influenced changes therein.

Acknowledgments

We thank Kenneth Hatter and Dr. Kenneth W. Jackson of the Molecular Biology Resource Facility at the William K. Warren Medical Research Foundation for mass spectrometer use and expertise.

References

1. Terhorst, C., P. Parham, D. L. Mann, and J. L. Strominger. 1976. Structure of HLA antigens: amino-acid and carbohydrate compositions and NH2-terminal sequences of four antigen preparations. *Proc. Natl. Acad. Sci. USA 73:910*.
2. Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature 329:506*.
3. Heemels, M. T., and H. Ploegh. 1995. Generation, translocation, and presentation of MHC class I-restricted peptides. *Annu. Rev. Biochem. 64:463*.
4. Serwold, T., F. Gonzalez, J. Kim, R. Jacob, and N. Shastri. 2002. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature 419:480*.
5. Saric, T., S. C. Chang, A. Hattori, I. A. York, S. Markant, K. L. Rock, M. Tsujimoto, and A. L. Goldberg. 2002. An IFN-γ-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. *Nat. Immun. 3:1169*.
6. Townsend, A. R., F. M. Gotch, and J. Davey. 1985. Cytotoxic T cells recognize fragments of the influenza nucleoprotein. *Cell 42:457*.
7. Vose, B. M., and G. D. Bonnard. 1982. Human tumor antigens defined by cytotoxicity and proliferative responses of cultured lymphoid cells. *Nature 296:359*.
8. Muul, L. M., P. J. Spiess, E. P. Director, and S. A. Rosenberg. 1987. Identification of specific cytolytic immune responses against autologous tumor in humans bearing malignant melanoma. *J. Immunol. 138:989*.
9. Shastri, N., S. Schwab, and T. Serwold. 2002. Producing nature's gene-chips: the generation of peptides for display by MHC class I molecules. *Annu. Rev. Immunol. 20:463*.
10. Hickman, H. D., A. D. Luis, W. Bardet, R. Buchli, C. L. Battson, M. H. Shearer, K. W. Jackson, R. C. Kennedy, and W. H. Hildebrand. 2003. Cutting edge: class I presentation of host peptides following HIV infection. *J. Immunol. 171:22*.
11. Weinschenk, T., C. Gouttefangeas, M. Schirle, F. Obermayr, S. Walter, O. Schoor, R. Kurek, W. Loeser, K. H. Bichler, D. Wernet, et al. 2002. Integrated functional genomics approach for the design of patient-individual antitumor vaccines. *Cancer Res. 62:5818*.
12. Henderson, R. A., H. Michel, K. Sakaguchi, J. Shabanowitz, E. Appella, D. F. Hunt, and V. H. Engelhard. 1992. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science 255:1264*.
13. Hunt, D. F., R. A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. L. Cox, E. Appella, and V. H. Engelhard. 1992. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science 255:1261*.
14. Sathiamurthy, M., H. D. Hickman, J. W. Cavett, A. Zahoor, K. Prilliman, S. Metcalf, M. Fernandez Vina, and W. H. Hildebrand. 2003. Population of the HLA ligand database. *Tissue Antigens 61:12*.
15. Rammensee, H., J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. 1999. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics 50:213*.
16. Shimizu, Y., B. Koller, D. Geraghty, H. Orr, S. Shaw, P. Kavathas, and R. DeMars. 1986. Transfer of cloned human class I major histocompatibility complex genes into HLA mutant human lymphoblastoid cells. *Mol. Cell. Biol. 6:1074*.
17. Prilliman, K., M. Lindsey, Y. Zuo, K. W. Jackson, Y. Zhang, and W. Hildebrand. 1997. Large-scale production of class I bound peptides: assigning a signature to HLA-B*1501. *Immunogenetics 45:379*.
18. Barnstable, C. J., W. F. Bodmer, G. Brown, G. Galfre, C. Milstein, A. F. Williams, and A. Ziegler. 1978. Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens: new tools for genetic analysis. *Cell 14:9*.
19. Perkins, D. N., D. J. Pappin, D. M. Creasy, and J. S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis 20:3551*.
20. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol. 215:403*.
21. Pruitt, K. D., K. S. Katz, H. Sicotte, and D. R. Maglott. 2000. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet. 16:44*.
22. Dennis, G., Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol. 4:R60*.
23. Camon, E., M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler. 2003. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res. 13:662*.
24. Ogg, G. S., T. Dong, P. Hansasuta, L. Dorrell, J. Clarke, R. Coker, G. Luzzi, C. Conlon, A. P. McMichael, and S. Rowland-Jones. 1998. Four novel cytotoxic T-lymphocyte epitopes in the highly conserved major homology region of HIV-1 Gag, restricted through B*4402, B*1801, A*2601, B*70 (B*1509). *AIDS 12:1561*.
25. Papadopoulos, K. P., A. I. Colovai, A. Maffei, D. Jaraquemada, N. Suciu-Foca, and P. E. Harris. 1996. Tissue-specific self-peptides bound by major histocompatibility complex class I molecules of a human pancreatic β-cell line. *Diabetes 45:1761*.
26. Steven, N. M., N. E. Annels, A. Kumar, A. M. Leese, M. G. Kurilla, and A. B. Rickinson. 1997. Immediate early and early lytic cycle proteins are frequent targets of the Epstein-Barr virus-induced cytotoxic T cell response. *J. Exp. Med. 185:1605*.
27. Kessler, B., X. Hong, J. Petrovic, A. Borodovsky, N. P. Dantuma, M. Bogyo, H. S. Overkleeft, H. Ploegh, and R. Glas. 2003. Pathways accessory to proteasomal proteolysis are less efficient in major histocompatibility complex class I antigen production. *J. Biol. Chem. 278:10013*.
28. Alberts, B. 2002. *Molecular Biology of the Cell*. Garland Science, New York.
29. Rock, K. L., and A. L. Goldberg. 1999. Degradation of cell proteins and the generation of MHC class I-presented peptides. *Annu. Rev. Immunol. 17:739*.
30. Seifert, U., C. Maranon, A. Shmueli, J. F. Desoutter, L. Wesoloski, K. Janek, P. Henklein, S. Diescher, M. Andrieu, H. de la Salle, et al. 2003. An essential role for tripeptidyl peptidase in the generation of an MHC class I epitope. *Nat. Immun. 4:375*.
31. Saveanu, L., D. Fruci, and P. van Endert. 2002. Beyond the proteasome: trimming, degradation and generation of MHC class I ligands by auxiliary proteases. *Mol. Immunol. 39:203*.

32. Houde, M., S. Bertholet, E. Gagnon, S. Brunet, G. Goyette, A. Laplante, M. F. Princiotta, P. Thibault, D. Sacks, and M. Desjardins. 2003. Phagosomes are competent organelles for antigen cross-presentation. *Nature 425:402*.

33. Guermonprez, P., L. Saveanu, M. Kleijmeer, J. Davoust, P. Van Endert, and S. Amigorena. 2003. ER-phagosome fusion defines an MHC class I cross-presentation compartment in dendritic cells. *Nature 425:397*.

34. Reits, E., A. Griekspoor, J. Neijssen, T. Groothuis, K. Jalink, P. van Veelen, H. Janssen, J. Calafat, J. W. Drijfhout, and J. Neefjes. 2003. Peptide diffusion, protection, and degradation in nuclear and cytoplasmic compartments before antigen presentation by MHC class I. *Immunity 18:97*.

35. Schubert, U., L. C. Anton, J. Gibbs, C. C. Norbury, J. W. Yewdell, and J. R. Bennink. 2000. Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature 404:770*.

36. Yewdell, J. 2002. To DRiP or not to DRiP: generating peptide ligands for MHC class I molecules from biosynthesized proteins. *Mol. Immunol. 39:139*.

37. Reits, E. A., J. C. Vos, M. Gromme, and J. Neefjes. 2000. The major substrates for TAP in vivo are derived from newly synthesized proteins. *Nature 404:774*.

38. Perez, O. D., and G. P. Nolan. 2001. Resistance is futile: assimilation of cellular machinery by HIV-1. *Immunity 15:687*.

39. Garrus, J. E., U. K. von Schwedler, O. W. Pornillos, S. G. Morham, K. H. Zavitz, H. E. Wang, D. A. Wettstein, K. M. Stray, M. Cote, R. L. Rich, et al. 2001. Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding. *Cell 107:55*.

40. Michaelsson, J., C. Teixeira de Matos, A. Achour, L. L. Lanier, K. Karre, and K. Soderstrom. 2002. A signal peptide derived from hsp60 binds HLA-E and interferes with CD94/NKG2A recognition. *J. Exp. Med. 196:1403*.

41. Vukmanovic, S., and F. R. Santori. 2003. Cooperation or sabotage? Self-peptide-MHC complexes influence T-cell responses to antigens. *Trends Immunol. 24:472*.

42. Schwab, S. R., K. C. Li, C. Kang, and N. Shastri. 2003. Constitutive display of cryptic translation products by MHC class I molecules. *Science 301:1367*.