Definition of MHC Supertypes Through Clustering of MHC Peptide Binding Repertoires

Pedro A. Reche[1,2*] and Ellis L. Reinherz[1,2]

[1] Laboratory of Immunobiology and Department of Medical Oncology, Dana-Farber Cancer Institute
[2] Department of Medicine, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA.
TEL: +1-617-632-3412, FAX: +1-617-632-3351
reche@research.dfci.havard.edu

Abstract. MHC molecules, also known in the human as human leukocyte antigens (HLA), display peptides on antigen presenting cell surfaces for subsequent T cell recognition. Identification of these antigenic peptides is especially important for developing peptide-based vaccines. Consequently experimental and computational approaches have been developed for their identification. A major impediment to such an approach is the extreme polymorphism of HLA, which is in fact the basis for differential peptide binding. This problem can be mitigated by the observation that despite such polymorphisms, HLA molecules bind overlapping set of peptides, and therefore, may be grouped accordingly into supertypes. Here we describe a method of grouping HLA alleles into supertypes based on analysis and subsequent clustering of their peptide binding repertoires. Combining this method with the known allele and haplotype gene frequencies of HLA I molecules for five major American ethnic groups (Black, Caucasian, Hispanic, Native American, and Asian), it is now feasible to identify supertypic combinations for prediction of antigenic peptide, offering the potential to generate peptide-vaccines with a population coverage ≥95%, regardless of ethnicity. One combination including five distinct supertypes is available online at our PEPVAC web server (http://immunax.dfci.harvard.edu/PEPVAC/). Promiscuous peptides predicted to bind to these five supertypes represent around 5% of all possible peptide binders from a given genome.

1 Introduction

T cell immune responses are responsible for fighting viruses, pathogenic bacteria and the elimination of cancer cells, and are triggered by the recognition of foreign peptide antigens bound to cell membrane expressed MHC molecules via their T cell receptors (TCR)(reviewed in [1-3]. In the human, MHC molecules are also termed human leukocyte antigens or HLA. Traditionally, identification of T cell epitopes required the synthesis of overlapping peptides (15-20 mers overlapping 10 amino acids) spanning the entire length of a protein, followed by experimental assays on each peptide such as in vitro intracellular cytokine staining [4]. This method is economically viable only

* To whom correspondence should be addressed.

G. Nicosia et al. (Eds.): ICARIS 2004, LNCS 3239, pp. 189–196, 2004.

for single proteins or pathogens consisting of a few proteins. As a result, computational approaches are used for the anticipation of antigenic peptides. Since T cells recognize antigenic peptides only in the context of MHC molecules [5], methods for the anticipation of antigenic peptides rely on the prediction of peptide-MHC binding. Peptides bound to the same MHC are related by sequence similarity [6, 7], and thus we have recently developed a method for the prediction of peptide-MHC binding based on the the use of position specific scoring matrix (PSSMs) or profiles derived from aligned peptides known to bind to MHC [8] [9, 10].

A major complication to the development of T cell based immunotherapies (vaccines) using peptide antigens lies in the polymorphism of HLA molecules, which is the basis for their differential peptide binding specificity [10]. Because of the required HLA restriction and ethnic variation in HLA distribution [11], such epitope vaccines might not be effective across populations. Conversely, developing a broadly protective multi-epitope vaccine will require the targeting of a large number of HLA molecules for peptide-binding predictions, yielding an impractical large number of peptides to work with. Interestingly, groups of several HLA molecules (supertypes) can bind largely overlapping sets of peptides [12, 13]. A systematic selection of HLA supertypic peptide binders would allow the immune response to be stimulated in individual of different genetic backgrounds. Thus, identification of HLA supertypes would facilitate the practical development of epitope-based vaccines. Here we describe a method to define HLA supertypes based on the clustering of the predicted peptide binding repertoire of HLA molecules.

In this paper, we have applied the method to class I HLA molecules (HLA I), unraveling new peptide binding relationships, and defining new supertypes. Furthermore, using the HLA I allele and haplotype gene frequencies for five major American ethnicities (Black, Caucasian, Hispanic, Native American, and Asian), we have identified combinations of supertypes that when targeted for peptide predictions are able to give a population coverage 95%, regardless of ethnicity. One of these combinations comprises 5 supertypes and is available online at our PEPVAC web server (http://immunax.dfci.harvard.edu/PEPVAC/). The selected supertypes with the included alleles (in parenthesis) are the following: A2 (A*0201-07, A*0209, A*6802), A3 (A*0301, A*1101, A*3101, A*3301, A*6801, A*6601), A24 (A*2402, B*3801), B7 (B*0702, B*3501, B*5101-02, B*5301, B*5401), B15 (A*0101, B*1501_B62, B1502). The PEPVAC resource also allows the prediction of proteasomal cleavage using language models [14]. Identification of promiscuous peptide binders to these supertypes using PEPVAC reduces the total number of predicted epitopes without compromising population coverage, thus being useful for the design of multi-epitope vaccines.

2 Matherial and Methods

2.1 MHCI-Peptide Binding Repertoire

Peptide-binding repertoires of the 55 HLA I molecules considered in this study consisted of sets of peptides that were predicted to bind to the relevant HLA I molecules. Peptide binding predictions were obtained from a random protein of 1000 amino acids in length (swiss-prot amino acid distribution), using position specific scoring matrices

(PSSMs) obtained from aligned peptides known to bind to that HLA I molecule [9, 10]. The peptide binding repertoire for each HLA I molecule consisted of the 2% top scoring peptides, and thus was composed of 20 peptides. PSSMs used in this study were all obtained from peptides of 9 residues in length (9mers), and thereby predicted peptide binders were all 9mers.

2.2 Supertype Construction

HLA I supertypes were derived by clustering the peptide-binding repertoire overlap of HLA I as it is illustrated in Fig. 1. First, the overlap between the peptidebinding repertoire of any two pairs of HLA I molecules was computed as the number of peptides binders shared by the two HLA I molecules. Let that number be n_{ij}, where i and j, represent the peptide binding repertoire of the HLA I molecules i and j, respectively. Subsequently, a distance coefficient (d_{ij}) was obtained as follows:

$$d_{ij} = N - n_{ij} \quad (1)$$

Where N is the total number of peptides considered in the predicted peptide binding repertoire of any HLA I molecule, and is equal to 20 (peptide binding repertoire consisted 2% of top scoring peptides from a random protein of 1000 amino acids; see above). Thus, if the peptide binding repertoire between two HLA I molecules is identical, then $d_{ij} = 0$. Alternatively, if they share no peptides in common, then $d_{ij} = 20$. Consequently, a quadratic distance matrix was derived containing the d_{ij} coefficient for all distinct pair of HLA I molecules. Finally, clustering of the peptide-binding overlap was carried from this distance matrix using the clustering algorithms in the Phylogeny Inference Package (PHYLIP) [15], and the relationship were visualized in the form of a phylogenic tree.

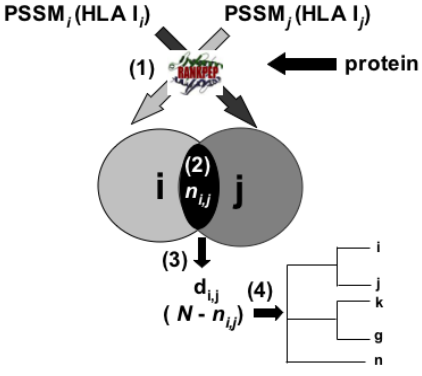


Fig. 1. Overview of the method followed for the identification HLA I supertypes. HLA I supertypes are identified by clustering their peptide binding repertoire (Materials and Methods).The method consists of 4 basis step. (1) Prediction of the peptide biding repertoire (*i,j* sets in figure) of each HLA I molecule from the same random protein using the relevant PSSMs in combination with the RANKPEP scoring algorithm [9]. (2) Compute the number of common peptides between the binding repertoire of any two HLA I molecules. (3) Build a distance matrix whose coefficients are inversely proportional to the peptide binding overlap between any pair of HLA I molecules. (4) Use a phylogenic clustering algorithm to compute and visualize HLA I supertypes (clusters of HLA I molecules with overlapping peptide binding repertoires).

3 Results and Discussion

3.1 Identification of HLA I Supertypes

Sidney, Sette and co-workers [12, 13](hereafter Sidney-Sette et al) described the first HLA I supertypes by carefully inspection of the reported peptide binding motifs of individual HLA alleles. While we acknowledge the pioneering work of these authors, the relationships between peptide binding specificities of HLA molecules may be too subtle to be defined by visual inspection of these peptide binding motifs. Furthermore, such sequence patterns have proven to be too simple to describe the binding ability of a peptide to a given MHC molecule [16, 17]. In view of these limitations, we have developed an alternative method to define HLA supertypes (outlined in Fig. 1 and described in Material and Methods). The core of the method consists of the generation of a distance matrix whose coefficients are inversely proportional to the peptide binders shared by any two HLA molecules (Fig. 1). Subsequently, this distance matrix is fed to a phylogenic clustering algorithm to establish the kinship among the distinct HLA peptide binding repertoires. Fig. 2 shows a phylogenic tree built upon the peptide binding repertoire of 55 HLA I molecules using a Fitch and Margoliash clustering algorithm [18]. In this representation, HLA I alleles with overlapping peptide binding repertoires (similar peptide binding specificities) branch together in groups or clusters, and we have identified as supertypes those clusters including alleles with at least a 20% overlap in their peptide-binding repertoire (pairwise)(highlighted in Fig. 2). It is important to indicate, that relationships between the HLA I peptide binding specificities noted in this study are restricted by the available HLA I binding repertoires obtained using our profiles, and relationships might shift in the future as the result of increasing the number of HLA I binding repertoires considered. As expected, our analysis indicates that the overlap between the peptide binding specificities of HLA I molecules is mostly confined to alleles belonging to the same gene. Nevertheless, it has also become apparent that peptidebinding overlap exists between alleles belonging to the HLA-A and HLA-B genes (Fig. 2, B15 cluster; B*4402 and A*2902; and A*2402 and B*3801). Relationships between the peptide binding specificities of HLA-A and HLA-B alleles (as well as new defined supertypes) would need experimental confirmation, but nevertheless it suggest that peptide-vaccine development would benefit from this inter gene cross-presentation. Cross-presentation of peptide in the context of two different HLA I genes might increase the frequency at which a peptide is recognized, thereby increasing its potential immunogenicity.

3.2 Supertypes and Population Coverage Studies

Supertypes defined in this study include the A2, A3, B7, B27 and B44 supertypes previously identified by Sidney-Sette et al. In our analysis, the alleles included in these supertypes match very well to those described earlier. However, there are discrepancies that extent beyond the limits imposed by the availability of profiles to obtain the peptide binding repertoire to specific HLA I alleles. The B27 supertype we defined here is more restricted than that proposed by Sidney-Sette et al, that also included the B*1509, B*1510, and B*3801 alleles, among others. The B7 supertype defined by Sidney-Sette et al. included the B*1508 allele, whereas in our analysis,

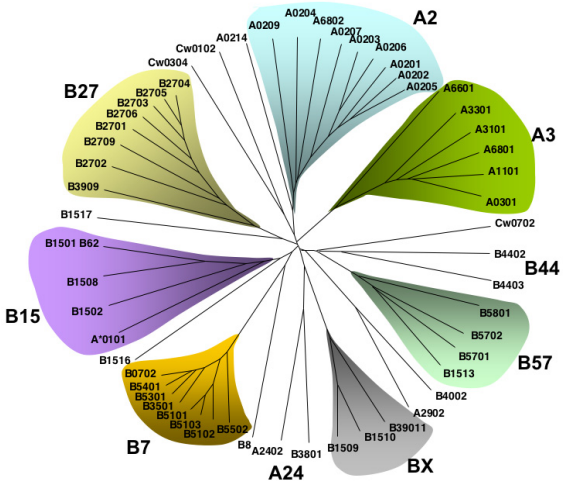


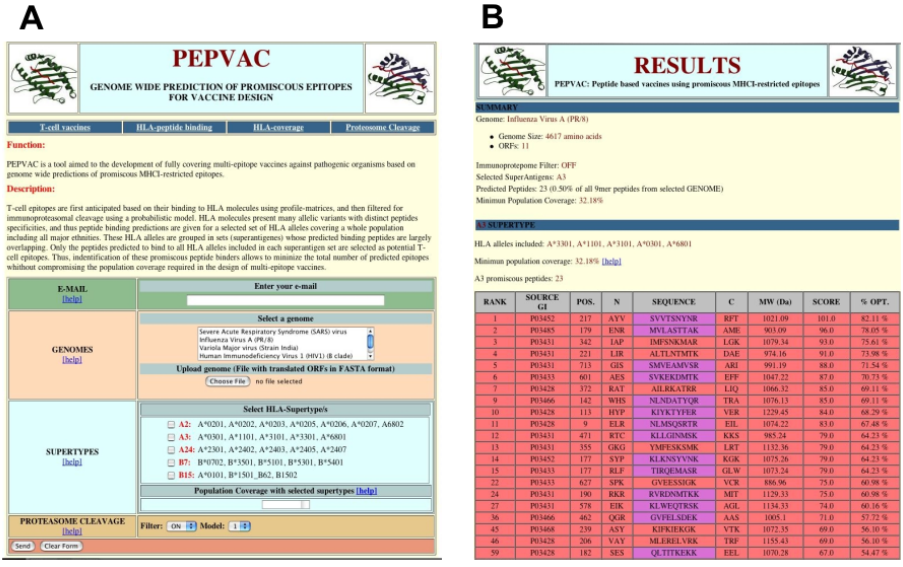
Fig. 2. *HLA I peptide binding overlap and supertypes.* Figures shows an unroot dendrogram built after clustering the overlap between the peptide-binding repertoire of the indicated HLA I molecules. The dendrogram reflects the relationship between the peptide-binding specificities of HLA I molecules. HLA I alleles with similar peptide binding specificities branch together in groups or cluster. The closer HLA I alleles branch, the larger is the overlap between their peptide-binding repertoires. Supertypes (shadowed with different colors) consist of groups HLA I alleles with at least a 20% peptide binding overlap (pairwise between any pair of alleles).

B*1508 cluster with the A*0101, and B*1501-02 alleles (Fig. 1, B15 supertype). Sidney-Sette et al also described a potential A24 supertype including the alleles A*2301, A*2402 and A*3001. We do not have profiles for the prediction of peptide binding to A*2301 or A*3001, but nevertheless with the available profiles the A*2402 binding specificity seems to be closer to that of B*3801. In addition, a new interesting peptide binding relationship includes that between the A*2902 and B*4002 alleles. Also, following our study we have defined two new supertypes, BX and B57. The BX supertype would include the alleles B*1509, B*1510 and B*39011, and B57 includes the alleles B*5801, B*5701-02, and B*1513.

Table 1. Cumulative phenotype frequency of defined supertypes

Supertype	Alleles	Blacks	Caucasians	Hispanics	*N.A.Natives	Asians
A2	A*0201-7, A*6802	43.7%	49.9%	51.8%	52.4%	44.7%
A3	A*0301, A*1101, A*3101, A*3301, A*6801, A*6601	35.4%	46.9%	41.5%	40.7%	47.9%
B7	B*0702, B*3501, B*5101-02, B*5301, B*5401	45.9%	42.2%	40.5%	52.0%	31.3%
B15	A*0101, B*1501_B62, B1502	13.06%	37.80%	16.75%	27.26%	21.04%
A24	A*2402, B*3801	15.5%	17.28%	25.85%	41.94%	35.0%
B44	B*4402, B*4403	10.4%	27.7%	17.15%	14.4%	10.1%
B57	B5701-02, B5801, B*1503	19.2%	10.3%	5.9%	5.8%	16.5%
ABX	A*2902, B*4002	7.4%	11.3%	19.1%	16.3%	16.3%
B27	B*2701-06, B*2709, B*3909	2.3%	4.8%	5.1%	16.9%	4.7%
BX	B*1509, B*1510, B*39011	3.1%	0.7%	4.2%	7.8%	4.1%

Cumulative phenotype frequency was obtained using the HLA I gene and haplotype frequencies published by published by Cao et al [19] corresponding to the indicated 5 American ethnic groups. Method for computing the cumulative phenotype frequency considered the disequilibrium linkage between the HLA-A and -B gene and was based on that reported by Dawson et al [20] *North American Natives.



RANK	SOURCE GI	POS.	N	SEQUENCE	C	MW (Da)	SCORE	% OPT.
1	P03452	217	AYV	SVVTSNYNR	RPT	1021.09	101.0	82.11 %
2	P03485	179	ENR	MVLASTTAK	AME	903.09	96.0	78.05 %
3	P03431	342	IAP	IMFSNKMAR	LGK	1079.34	93.0	75.61 %
4	P03451	221	LIR	ALTLNTMTK	DAE	974.16	91.0	73.98 %
5	P03431	713	GIS	SMVEAMVSR	ARI	991.19	88.0	71.54 %
6	P03433	601	AES	SVKEKDMTK	EFF	1047.22	87.0	70.73 %
7	P03428	372	RAT	AILRKATRR	LIQ	1066.32	85.0	69.11 %
9	P03466	142	WHS	NLNDATYQR	TRA	1076.13	85.0	69.11 %
10	P03428	113	HYP	KIYKTYPER	VER	1229.45	84.0	68.29 %
11	P03428	9	ELR	NLMSQSRTR	EIL	1074.22	83.0	67.48 %
12	P03451	471	RTC	KLLGINMSK	KKS	985.24	79.0	64.23 %
13	P03431	355	GKG	YMPESKSMK	LRT	1132.36	79.0	64.23 %
14	P03452	177	SYP	KLKNSYVNK	KGK	1075.26	79.0	64.23 %
15	P03433	177	RLF	TIRQEMASR	GLW	1073.24	79.0	64.23 %
22	P03433	627	SPK	GVEESSIGK	VCR	886.96	75.0	60.98 %
24	P03451	190	RKR	RVRDNMTKK	MIT	1129.33	75.0	60.98 %
27	P03431	578	EIK	KLWEQTRSK	AGL	1134.33	74.0	60.16 %
36	P03466	462	QGR	GVFELSDEK	AAS	1005.1	71.0	57.72 %
45	P03468	239	ASY	KIPIJEKGK	VTK	1072.35	69.0	56.10 %
46	P03428	206	VAY	MLERELVRK	TRF	1155.43	69.0	56.10 %
59	P03428	182	SES	QLTITKEKK	EEL	1070.28	67.0	54.47 %

Fig. 3. *The PEPVAC web server.* A) PEPVAC input page. The page is divided into the following sections: E-MAIL, GENOMES, SUPERTYPES, AND PROTEASOMAL CLEAVAGE. Basically, the server allows the targetting of pathogenic organisms (GENOMES section) for the prediction of promiscous peptide binders to any combination of the supertypes A2, A3, B7, A24, and B15. The CPF of the selected supertypes is calculated on-the-fly and shown on the relevant window. Prediction of proteasomal cleavage using three optimal language models are carried out in parallel to the peptide binding predictions (PROTEASOMAL CLEAVAGE section). B) PEPVAC result page. In the example shown, the A3 supertype was selected for peptide binding predictions from the genome of *Influenza A virus*. The result page first displays a summary of the predictions which reports the chosen selections, the number of predicted peptides and the minimum population coverage provided by the supertypic selection, followed by the predicted peptide binders to each of the selected supertypes. Peptides are ranked by score, and are predicted to bind to all alleles included in the supertype. Relevant information about each sorted peptide includes its protein source as well as its molecular weight. Peptides shown in violet contain a C-terminal residue that is predicted to be the result of proteasomal cleavage. If the proteasomal cleavage filter is checked ON in the input page, only violet peptides will be shown.

The cumulative phenotypic frequency (CPF) of the supertypes defined in this study is shown in Table 1. CPF was calculated by 5 distinct American ethic groups (Black, Caucasian, Hispanic, North America Natives and Asian), and it represents the population coverage that would be provided by a vaccine consisting of epitopes restricted by the alleles included in the supertype. The A2, A3 and B7 supertypes have the largest CPF in the 5 studied ethnic groups, and in fact, taken together, they provide a CPF close to 90%, irrespective of ethnicity. To increase the population coverage to 95% regardless of ethnicity it is necessary to include at least two more supertypes. Specifically, the supertypes A2, A3, B7, B15 and A24 or B44 represent the minimal supertypic combination with the largest population coverage. Note that the supertype B57, despite its quite large CPF, is not included in this minimal supertypic combination since alleles in the B57 supertype are in linkage disequilibrium with other alleles included in more prevalent supertypes.

3.3 Tool for the Design of Epitope-Based Vaccines

Prediction of promiscuous peptide-binders of supertypes that include highly prevalent alleles in the human population provides a head start to the development of broadly covering epitope-based vaccines. With that objective, we have implemented a tool that allow the prediction of promiscuous peptide binders to the supertypes A2, A3, B7, B15 and A24 (Fig. 1 and Table 1), which have a CPF greater than 95%, irrespective of ethnicity. We named this tool PEPVAC (Promiscous Epitopes based VACcines), and it is online at the site http://immunax.dfci.harvard.edu/PEPVAC/ hosted by the Molecular Immunology Foundation/Dana-Farber Cancer Institute. The web interface to PEPVAC (shown in Fig. 3A) allows targetting of any combination of the mentioned selected supertypes, displaying CPF of the selected supertypes.

MHCI-restricted epitopes derive from protein fragments generated by the protease activity of the proteasome, and it is thought that the C-terminus of any MHCIrestricted epitope is the result of the original proteasomal cleavage [21]. Consequently, in PEPVAC we have combined the predictions of promiscuous supertypic peptide binders using profiles, with probabilistic models that indicate whether a C-terminus of a given peptide is likely be the result of proteosomal cleavage. Probabilistic models for proteasomal cleavage were generated from a set of known epitopes restricted by human HLA I molecules, as indicated elsewhere [14]. Promiscuous peptide binders containing a C-terminal end predicted to be the result of proteasomal cleavage are shown in violet in the result page (Fig. 3. B). Threshold for the prediction of promiscuous peptide binders in PEPVAC have been fixed to provide a reduced and manageable set of promiscuous peptide-binders to each supertype. As an example, predicted promiscuous peptides to the above 5 supertypes from a genome such as that of *Influenza virus A* (4160 amino acids distributed in 10 distinct open reading frames (ORF)), represents only 5.51% (254 9mer peptides) of all possible peptides (4617 9mer peptides). Furthermore, this figure further contracts to 170 peptides (3.7% of all 9mer peptides from *Influenza virus A* genome) if only those peptides that are predicted to be cleaved by the proteasome are considered. Thus, PEPVAC is well fit to provide genome-wide predictions of promiscuous HLA I restricted epitopes at a practical scale for the development multi-epitope vaccines against pathogenic organisms and offering a broad population coverage.

Acknowledgments. This manuscript was supported by NIH grant AI50900 and the Molecular Immunology Foundation. We wish to acknowledge John-Paul Glutting for programming assistance.

References

1. Margulies, D.H., *Interactions of TCRs with MHC-peptide complexes: a quantitative basis for mechanistic models.* Curr Opin Immunol, Vol. **9** (1997) 390-5.
2. Garcia, K.C., L. Teyton, and I.A. Wilson, *Structural basis of T cell recognition.* Annu Rev Immunol, Vol. **17** (1999) 369-397.
3. Wang, J.-H. and E. Reinherz, *Structural basis of T cell recognition of peptides bound to MHC molecules.* Molecular Immunology, Vol. **38** (2001) 1039-1049.

4. Draenert, R., et al., *Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T cell responses.* J Immunol Methods, Vol. **275** (2003) 19-29.
5. Zinkernagel, R.M. and P.C. Doherty, *Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system.* Nature, Vol. **248** (1974) 701-702.
6. Falk, K., et al., *Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.* Nature, Vol. **351** (1991) 290-296.
7. Rammensee, H.G., T. Friede, and S. Stevanoviic, *MHC ligands and peptide motifs: first listing.* Immunogenetics, Vol. **41** (1995) 178-228.
8. Gribskov, M., A.D. McLachlan, and D. Eisenberg, *Profile analysis: detection of distantly related proteins.* Proc Natl Acad Sci USA, Vol. **84** (1987) 4355-4358.
9. Reche, P.A., J.P. Glutting, and E.L. Reinherz, *Prediction of MHC class I binding peptides using profile motifs.* Hum Immunol, Vol. **63** (2002) 701-9.
10. Reche, P.A. and E.L. Reinherz, *Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms.* J Mol Biol, Vol. **331** (2003) 623-41.
11. David W. Gjertson and Paul I. Terasaki, E., *HLA 1998.* (1998).
12. Sette, A. and J. Sidney, *Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism.* Immunogenetics, Vol. **50** (1999) 201-12.
13. Sette, A. and J. Sidney, *HLA supertypes and supermotifs: a functional perspective on HLA polymorphism.* Curr Opin Immunol, Vol. **10** (1998) 478-82.
14. Reche, P.A., J.-P. Glutting, and E.L. Reinherz, *Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.* Immunogenetics, Vol. **Submitted** (2004).
15. Retief, J.D., *Phylogenetic analysis using PHYLIP.* **132** (2000) 243-58.
16. Bouvier, M. and D.C. Wiley, *Importance of peptide amino acid and carboxyl termini to the stability of MHC class I molecules.* Science, Vol. **265** (1994) 398-402.
17. Ruppert, J., et al., *Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules.* Cell, Vol. **74** (1993) 929-937.
18. Fitch, W.M. and E. Margoliash, *Construction of phylogenetic trees.* Science, Vol. **155** (1967) 279-84.
19. Cao, K., et al., *Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations.* Hum Immunol, Vol. **62** (2001) 1009-30.
20. Dawson, D.V., et al., *Ramifications of HLA class I polymorphism and population genetics for vaccine development.* Genet Epidemiol, Vol. **20** (2001) 87-106.
21. Craiu, A., et al., *Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide.* Proc Natl Acad Sci U S A, Vol. **94** (1997) 10850-5.