[Frontiers in Bioscience 10, 879-886, January 1, 2005]

A FRAMEWORK TO SUB-TYPE HLA SUPERTYPES

Pandjassarame Kangueane [1], Meena Kishore Sakharkar [1], Ganapathy Rajaseger [2], Subhashini Bolisetty [3], Balasubramanian Sivasekari [3], Bing Zhao [1], Manickam Ravichandran [4], Paul Shapshak [5], Subramanian Subbiah [6]

[1] School of Mechanical and Production Engineering, Nanyang Technological University, Nanyang Centre for Supercomputing and Visualization, Singapore 639798, [2]DMERI, Defense Science Organization, Singapore, [3] Centre for Biotechnology, Anna University, India 600 025, [4]Department of Microbiology and Parasitology, School of Medical Sciences, University Science Malaysia, Malaysia, [5]Department of Psychiatry and Behavior Science, University of Miami Medical School, USA, [6] Department of Applied Physics, Stanford University, USA

TABLE OF CONTENTS

1. ABSTRACT

The human leukocyte antigen (HLA) alleles are extremely polymorphic among ethnic population and the peptide binding specificity varies for different alleles in a combinatorial manner. However, it has been suggested that majority of alleles can be covered within few HLA supertypes, where different members of a supertype bind similar peptides, yet exhibiting distinct repertoires. Since the overlap between different members of a supertype appears to be extensive, it is crucial to develop a framework for grouping alleles into supertypes just from sequence information.

In this report, we define sub supertypes, where members show functional overlap with identical repertoire, and describe a strategy to group HLA-A, B and C alleles into different categories of sub supertypes. The strategy grouped 47% of 295 A alleles, 44% of 540 B alleles and 35% of 156 C alleles to just 36, 71 and 18 groups, respectively. The grouping is moderately validated using available binding data. However, the validation is limited due to lack of binding data. Hence, the data presented in this article serve as a framework to test specific functional overlap between alleles.

The grouping of HLA alleles into different categories of sub supertypes has profound use in the understanding of antigenic peptide selection, degeneration and discrimination during T-cell mediated immune response. A complete knowledge of this phenomenon finds utility in epitope design for the development of HLA based vaccines and immuno-therapeutics.

2. INTRODUCTION

The human leukocyte antigen (HLA) alleles are highly polymorphic among ethnic population. Today, more than 1,800 HLA alleles are known and about a 1,000 of them refer to the class I loci (1). Class I alleles bind peptides of length 8-10 residues during T-cell mediated immune response (2). Therefore, the huge combination of specific HLA-peptide binding is clearly beyond our realization. However, it has been suggested that a majority of alleles can be grouped within few "HLA supertypes", where the members of a supertype bind similar peptides, yet distinct binding repertoires (3). The functional overlap between different alleles within defined supertypes will significantly reduce peptide binding diversity. A catalogue of functional overlap is critical for grouping alleles into supertypes from sequence information. In recent years, a number of supertypes have been defined by comparing peptide binding data. Thus, HLA- A1 (4), A2 (3, 5), A3 (5), A24 (4), B7 (5), B27 (4), B44 (6), B58 (4), and B62 (4) supertypes have been defined. Classification of alleles into supertypes using binding data is seldom comprehensive and conclusive. Moreover, a complete grouping of all known alleles using binding data is laborious and expensive. It is also practically impossible to cluster all known alleles using binding data at multiple levels of functional overlaps. Therefore, it is critical to develop theoretical procedures for grouping alleles into supertypes. However, such grouping procedures require rigorous validation prior to routine application. Chelvanayagam *et al.* (7), Zhang *et al.* (8), Zhao *et al.* (9) and Doytchinova *et al.* (10) grouped HLA alleles into functionally overlapping clusters from sequence data. Chelvanayagam *et al.* (7) identified interaction pockets from MHC-peptide (MHCp) crystal structures, Zhang *et al.* (8) defined A-F structural binding pockets, Zhao *et al.* (9) defined functional pockets made of critical polymorphic functional residue positions (CPFRP), Doytchinova *et al.* (10) used molecular interaction fields (MIF), hierarchical clustering (HC) and principal component analysis (PCA) and Lund *et al.* (11) used clustering procedures for grouping HLA alleles into putative supertypes. In this report, we utilize the procedure described by Zhao et al. (9) to define HLA 'sub-supertypes' with overlapping function of identical binding repertoire. Using this approach, we grouped 991 alleles (class I) into several groups of 'sub-supertypes'. The grouping is further validated using binding data extracted from MHCBN (12). The importance of HLA sub-supertypes in establishing a conclusive framework for functional overlaps between alleles is discussed.

3. MATERIALS AND METHOD

3.1. Dataset

The protein sequences of HLA-A (295 alleles), HLA-B (540 alleles) and HLA-C (156 alleles) were obtained from IMGT/HLA (release 2.5) for this analysis (1).

3.2. Structural basis for HLA supertypes

HLA allele sequences show high degree of homology among themselves. Therefore, their structures have similar fold and peptide binding groove, where the 3-dimensional spatial orientations of residue backbone atoms are similar for various alleles at different residue positions. However, HLA alleles exhibit extreme polymorphism among themselves and their peptide binding specificity varies between them. Nonetheless, the peptide binding residues show similarity among certain alleles and these alleles binding identical peptides through the concept of HLA supertypes (Table 1).

3.3. Definition of HLA sub-supertypes

In this approach, we define critical polymorphic functional residue positions (CPFRP) for each allele using a methodology described elsewhere (9). Here, CPFRPs are described as those positions that are predominantly involved in peptide binding in known MHCp crystal structures and show polymorphism at least once among known A, B and C allele sequences. Thus, 21 CPFRPs were defined for each allele. The physical and chemical properties of residues at the CPFRPs characterize the binding difference between alleles. It has been shown that it is possible to sub-group members of a supertype using just 21 residues polymorphic in pockets A-F of the peptide binding groove (9). Hence, the 21 CPFRP residues were extracted for HLA A, B and C alleles. The extracted discontinuous residue segment patterns formed by the CPFRPs were compared and alleles having identical CPFR segments were grouped together. Thus, alleles clustered within a group share identical CPFR segments and are proposed to form functional pockets capable of binding similar peptides with identical repertoire (Tables 2 - 4). Therefore, members of a group bind similar peptides and the groups represent "sub-supertypes". It is proposed that the "sub-supertypes" would have the predictive power previously promised for the supertype itself.

3.4. Validation of predictive sub-supertypes

The grouping is reasonably validated using binding data (Tables 5 - 6) extracted from MHCBN (12). Here, we show several peptides, found in the literature, that are bound by two members of the same sub-supertype. In this article, a total of five pairs of alleles in five sub-supertypes (three of which are HLA-A, and two of which are HLA-B) are found to have similar peptide binding. It should be noted that the utility of this approach awaits more rigorous experimental validation.

4. RESULTS AND DISCUSSION

4.1. HLA Supertypes

More than 1,800 HLA alleles have been defined (1). Therefore, the number of theoretically possible combinations of HLA-peptide complexes is extremely large. However, the immune system maintains a homogenous balance by specific selection, degeneration and discrimination (self/non-self) of short peptides using HLA molecules. Although, HLA molecules are polymorphic in ethnic population, they exhibit a substantial amount of functional overlap through the phenomenon of 'HLA supertypes', where members bind similar peptides and yet display distinct repertoires. A number of 'HLA supertypes' have already been defined using binding data (Table 1). Table 1 shows six peptides binding to all members of the A2 supertypes (A*0201, A*0202, A*0203, A*0206 and A*6802). The functional overlap between different members of the supertype is intriguing. However, it also shows several peptides binding to some members but not all members of the A2 supertypes (Table 1).

4.2. Perplexing issues with HLA supertypes

The concept of HLA supertypes is that alleles belonging to supertypes bind a highly shared set of peptides; in principle it should be possible to predict peptide binding of other members of a supertype using experimental results based on just one member of the type. However, as illustrated in Table I, this promise does not hold in the major supertypes A and B. Hence, the binding of peptides to different members of the A2 supertype is combinatorial in selection and degeneration. Moreover, this grouping is inconclusive given the known number of HLA alleles. Therefore, we devised a theoretical procedure to group HLA alleles into clusters of overlapping function from sequence information (9).

Table 1. HLA supertype definition

Peptide	Supertype	A*0201	A*0202	A*0203	A*0206	A*6802	Reference
LLFNILGGWV	A2	b	b	b	b	b	13
YLVAYQATV	A2	b	b	b	b	b	13
KVAELVHFL	A2	b	b	b	b	b	14
FLWGPRALV	A2	b	b	b	b	b	14
FLLLADARV	A2	b	b	b	b	b	13
IMIGVLVGV	A2	b	b	b	b	b	14
KIFGSLAFL	A2	b	b	b	b	**nb**	14
CLTSTVQLV	A2	b	b	b	b	**nb**	14
RLIVFPDLGV	A2	b	b	b	b	**nb**	13
YLQLVFGIEV	A2	b	b	b	b	**nb**	14
LLTFWNPPV	A2	b	b	b	b	**nb**	14
VLVGGVLAA	A2	b	b	b	b	**nb**	13
WMNRLIAFA	A2	b	b	b	**nb**	b	13
DLMGYIPLV	A2	b	**nb**	b	b	b	13
ILHNGAYSL	A2	b	b	b	**nb**	**nb**	14
YLSGANLNL	A2	b	b	b	**nb**	**nb**	14
VMAGVGSPYV	A2	b	b	b	**nb**	**nb**	14
ILAGYGAGV	A2	b	b	b	**nb**	**nb**	13
LMTFWNPPV	A2	b	**nb**	b	b	**nb**	14
YLVTRHADV	A2	b	**nb**	b	b	**nb**	13
HMWNFISGI	A2	b	**nb**	b	b	**nb**	14
YLLPRRGPRL	A2	b	**nb**	b	b	**nb**	13
LLFLLLADA	A2	b	b	**nb**	**nb**	**nb**	13
LLTFWNPPT	A2	b	**nb**	b	**nb**	**nb**	14
ALCRWGLLL	A2	b	**nb**	b	**nb**	**nb**	13
		A*0301	**A*1101**	**A*3101**	**A*3301**	**A*6801**	
KTSERQPR	A3	b	b	b	**nb**	b	13
RMYVGGVEHR	A3	b	b	b	**nb**	**nb**	13
QLFTFSPRR	A3	b	b	**nb**	**nb**	b	13
LGFGAYMSK	A3	b	b	**nb**	**nb**	b	13
LIFCHSKKK	A3	b	b	**nb**	**nb**	b	13
GVAGALVAFK	A3	b	b	**nb**	**nb**	b	13
VAGALVAFK	A3	b	b	**nb**	**nb**	b	13
RLGVRATRK	A3	b	b	b	?	?	15
		B*0702	**B*3501**	**B*5101**	**B*5301**	**B*5401**	
LPGCSFSIF	B7	b	b	b	b	**nb**	15

b = binder; nb = non-binder; ? = undetermined

4.3. Reduction of HLA sequence diversity

Our previous analysis grouped 101 HLA-A alleles into 29 clusters such that each group contains at least two alleles with identical functional binding CPFR segments (9). In this report, we extended the grouping to HLA-A, -B and -C alleles. Here, we used 991 alleles (class I) of which, 295 are 'A alleles', 540 are 'B alleles' and 156 are 'C alleles'. The discontinuous CPFR residue segment patterns were extracted for each allele (9). Comparison of these segment patterns among HLA-A, B and C alleles grouped A alleles into 192 unique clusters, B alleles into 373 unique clusters and C alleles into 119 unique clusters. Thus, 991 alleles were grouped into 684 clusters. The total diversity among the 991 alleles is reduced to 69% with a 'reduction in variability' of 31%. The diversity in A, B and C is reduced to 65%, 69% and 76%, respectively with a 'reduction in variability' of 35%, 31% and 24%, respectively. An understanding of such reduction among the functionally important binding residues is critical to decipher the molecular basis for HLA-supertypes.

4.4. HLA sub-supertypes

In this study, 139 A alleles (47% of all known A alleles) were clustered into 36 groups such that each group contains at least two alleles (Table 2). Members in each of these groups have identical CPFR segment patterns. Thus, alleles within a group have identical functional pockets. The proposed hypothesis is that alleles in a group bind to similar peptides and members in the group form 'sub-upertypes' with identical binding repertoire. If binding of a peptide is known for one representative allele in a group, its binding to other members in the group can be deduced. However, this clustering has to be validated using experimentally determined binding data for a representative set of alleles. This requires large scale generation of binding data and it is difficult to generate binding data at such scale due to limitations in the synthesis and purification of HLA alleles for binding assay. If this hypothesis is reasonably validated using binding data, it is possible to infer the binding property of a peptide to 47% of A alleles using binding values for 36 representative A alleles. The 36 groups clustered A alleles into A1, A2, A3, A11, A23, A24, A26, A29, A30, A31, A32, A33, A34, A36, A66, A68 and A74 'predictive sub-supertypes' where members of a group is proposed to bind similar peptides (Table 2). We extended our analysis to B alleles and showed that 238 B alleles (44% of all known B alleles) were clustered into 71 groups such that each group contains at least two alleles (Table 3). We also show that 55 C alleles (35% of all known C alleles) were grouped into 18 groups such that each group contains at least two alleles (Table 4). Thus, we demonstrate that the strategy (9) used to group alleles into sub-supertypes is extremely powerful in reducing functional diversity.

4.5. Validation of predictive sub-supertypes

In order to validate the proposed hypothesis, we extracted peptides from MHCBN (12). This experiment

Table 2. HLA-A alleles (139) grouped into putative supertypes

Group	CPFR	HLA-A alleles
1	YFARENHTDANTLIYRDARRG	A*0101 A*0104 A*0109
2	YFAREKHTHVDTLRYHYVLTW	A*0201 A*0209 A*0224 A*0225 A*0231 A*0240 A*0243 A*0245 A*0246 A*0258 A*0259 A*0266 A*0267 A*0268
3	YFAREKHTHVDTLRYHYVWTW	A*0202 A*0222 A*0250 A*0263
4	YYAREKHTHVDTLRYHYVLTW	A*0206 A*0214 A*0221 A*0228 A*0251 A*0261
5	YFAREKHTHVDTLRCHYVLTW	A*0207 A*0215 A*0218
6	YFAREKHTHVDTLRYHYVQTG	A*0219 A*0237
7	YFARENQTHVDTLRYHYVLTW	A*0256 A*0262
8	YFARENQTDVDTLIYRDELTW	A*0301 A*0304 A*0305 A*0306 A*0313 A*0314
9	YFARENQTDVDTLIYRDVQTW	A*0302 A*0310
10	YYARENQTDVDTLIYRDAQRW	A*1101 A*1102 A*1105 A*1107 A*1109 A*1112 A*1113 A*1115 A*1116
11	YSAREKHTDENIAMFHYVLTG	A*2301 A*2303 A*2306 A*2307 A*2413
12	YSAREKHTDENIAMFHYVWTG	A*2302 A*2406
13	YSAREKHTDENIAMFHYVQTG	A*2402 A*2405 A*2409 A*2411 A*2421 A*2426 A*2427 A*2429 A*2435 A*2437 A*2439 A*2440
14	YSAREKHTDENIAMFHYVQTW	A*2403 A*2423 A*2433
15	YSAREKHTDENIAMFRDVQTG	A*2417 A*2441
16	YSAREKHTHENIAMFHYVQTG	A*2430 A*2442
17	YYARNNHTDANTLRYQDEWRW	A*2601 A*2610 A*2614 A*2615 A*2617
18	YYARNNHTHVDTLRYQDEWRW	A*2603 A*2606
19	YYARNNHTDANTLRYQDVWRW	A*2612 A*2618
20	YTARQNQTDANTLMYRDVLTW	A*2901 A*2902 A*2906 A*2909 A*2910 A*2911
21	YSARENQTDVDTLIYEHWLTW	A*3001 A*3011
22	YSARENHTDENTLIYEHRLTW	A*3002 A*3003 A*3012
23	YSARENHTDENTLIYEHVWTW	A*3004 A*3006
24	YTARENHIDVDTLMYQDVLTW	A*3101 A*3109
25	YTARENHIDVDTLIYRDVLTW	A*3103 A*3104
26	YTAREKHTDENIAMYQDVLTW	A*3107 A*3108
27	YFARENHTDESIAMYQDVLTW	A*3201 A*3206
28	YTARNNHIDVDTLMYQDVLTW	A*3301 A*3303 A*3304 A*3305 A*3306 A*3307
29	YYARNKQTDVDTLRYQDEWTW	A*3401 A*3405
30	YYARNNQTDVDTLIYRDELTW	A*3402 A*3403 A*3404
31	YFARENHTDANTLIYRDARTW	A*3601 A*3602
32	YYARNNQTDVDTLRYQDEWRW	A*6601 A*6604
33	YYARNNQTDVDTLMYRDVWTW	A*6801 A*6816 A*6819 A*6821 A*6822 A*6824 A*6825
34	YYARNNHTHVDTLMYRDVWTW	A*6805 A*6820
35	YYARENQTDVDTLMYRDVWTW	A*6810 A*6813 A*6814
36	YFARENHTDVDTLMYQDVLTW	A*7401 A*7402 A*7403 A*7408 A*7409

CPFR = critical polymorphic functional residue

identified five peptides: (1) FLWGPRALV, (2) AAGIGILTV, (3) GILGFVFTL, (4) VLYRYGSFSV and (5) YLEPGPVTA binding A*0201 and A*0209 (Table 5). In Table 2, A*0201 and A*0209 are categorized together (Group #2). Comparison of Table 5 with Table 2 suggest that peptides binding to A*0201, also binds to A*0209. This observation is very interesting. However, it is important to establish the binding of these peptides to other members of the group such as A*0224, A*0225, A*0231, A*0240, A*0243, A*0245, A*0246, A*0258, A*0259, A*0266, A*0267 and A*026. It should be noted that it is labor intensive to clone and express all of these alleles to determine binding to the above peptides. Nonetheless, if experimentally determined binding data show similar binding to all these alleles, then extrapolation of the proposed hypothesis to other predictive sub-supertype is trivial.

We found yet another peptide GILGFVFTL which binds to A*0206 and A*0214 (Table 5). These two alleles are categorized together in 'Group #4' (Table 2). This observation is also interesting. However, if similar binding of this peptide to other members of 'Group 4' such as A*0221, A*0228, A*0251 and A*0261 is established, extrapolation of this strategy to other predictive sub-supertypes will be trivial (Table 2). Interestingly, GILGFVFTL binds A*0201, A*0209, A*0206 and A*0214 (Table 5). Although, GILGFVFTL binds all the above four alleles, our strategy grouped these alleles into two distinct groups (Group #2 (A*0201 & A*0209) and Group #4 (A*0206 & A*0214)). We further probed into their functional overlap but examining their CPFR segments. Examination of CPFR segments in Group #2 and Group #4 (Table 2) shows a single residue mutation (9F to 9Y) between these two groups. F (phenylalanine) and Y (tyrosine) are aromatic and the mutation is synonymous. This explains binding overlap between members of Group #2 and Group #4. By definition, our strategy groups members with similar peptide binding and members across groups (for example, Group #2 and Group #4) do not always bind similar peptides. We also found that a number of peptides (LLFNILGGWV, YLVAYQATV, KVAELVHFL, FLWGPRALV, FLLLADARV, IMIGVLVGV) bind A*0201 (Group #2) and A*0206 (Group #4). A*0201 and A*0206 are grouped as members of A2 supertype using binding data (Table 1). The data implies that some peptides binding to A*0201 are also found to bind A*0206. However, this is not always true. Peptides (WMNRLIAFA, ILHNGAYSL, YLSGANLNL, VMAGVGSPYV, ILAGYGAGV) that bind to A*0201 do not always bind to A*0206 (Table 1). This suggests that A*0201 and A*0206 are not strict members of the A2 'sub-supertype'. On the other hand, we propose that peptides that bind to A*0201 should strictly bind to A*0209 as they share identical CPFR segments (Table 2). Similarly, A*0206 and A*0214 should always bind similar peptides as they share identical CPFR segments (Table 2).

The CPFR segments in A*0201 (Group 2) and A*0202 (Group 3) shows single residue mutation (156L to

Table 3. HLA-B alleles (238) grouped into putative supertypes

Group	CPFR	HLA-B alleles
1	YYSRNIQTDESNLSYDYEREW	B*0702 B*0710 B*0721 B*0722 B*0723 B*0730 B*0733 B*0735
2	YYSRNINTDESNLSYDYEREW	B*0703 B*0708
3	YYSRNIQTDESNLSYNYEREW	B*0705 B*0706
4	YYSRNIQTDESNLRYDYEREW	B*0707 B*0712 B*0714 B*0718
5	YYSRNIQTDESNLSYDSEREW	B*0709 B*0717
6	YYSRNINTYESNLSYDYEREW	B*0716 B*0737
7	YDSRNINTDESNLSYNYVDTW	B*0801 B*0804 B*0818 B*0819
8	YYTREINTYENTARYNLVLEW	B*1301 B*1311
9	YYTREINTYENTATYNLVLEW	B*1302 B*1308
10	YYSRNINTDESNLWYNFELTW	B*1401 B*1402
11	YYAREINTYESNLRYDSEWLW	B*1501 B*1528 B*1533 B*1534 B*1538 B*1546 B*1556 B*1560 B*1566 B*1578 B*1581 B*1582
12	YYARNINTYESNLRYDSELLW	B*1502 B*1521 B*3511 B*3521
13	YYSREINTYESNLRYDSELLW	B*1503 B*1562 B*1574
14	YYAREINTYESNLTYDSEWLW	B*1504 B*1535
15	YYAREINTYESNLRYDSVLLW	B*1505 B*1520 B*3528
16	YYARNINTYESNLRYDSEWLW	B*1508 B*1511 B*1515 B*3514 B*3543
17	YYSRNINTYESNLRYDYELLW	B*1510 B*1537
18	YYAREINTYESNLRYDSEWLG	B*1512 B*1519
19	YYARNINTYENIARYDSELLW	B*1513 B*5306 B*5308
20	YYSRNINTYESNLRYDSELLW	B*1518 B*1529 B*1564 B*1572 B*1580
21	YYAREINTYESNLRYDSELLW	B*1525 B*1539
22	YYSREINTYESNLRYDSEREW	B*1547 B*1549
23	YYARNINTYESNLSYDSVLLW	B*1555 B*3505 B*3517 B*3530
24	YYARNIQTDESNLRYDSEWLW	B*1576 B*5603
25	YHSRNINTYESNLRYDSVLTW	B*1801 B*1805 B*1807 B*1811
26	YYARNINTYESNLRYDSVLTW	B*1804 B*3535
27	YHTREIKTDEDTLNYHDVLEW	B*2703 B*2705 B*2713 B*2717
28	YYARNINTYESNLRYDSVLLW	B*3501 B*3507 B*3519 B*3520 B*3524 B*3526 B*3532 B*3540 B*3541 B*3542 B*3546 B*3547 B*3549
29	YYARNINTYESNLRYNYVLLW	B*3502 B*3504 B*3509 B*3512
30	YYARNINTYESNLRYDFVLLW	B*3503 B*3536
31	YYARNINTYESNLRYDYVLLW	B*3534 B*3539
32	YHSREINTYEDTLRSNFVDTW	B*3701 B*3704
33	YYSRNINTYENIARYNFVLTW	B*3801 B*3805 B*3806 B*3807 B*3809
34	YYSRNINTDESNLRYNFVLTW	B*3901 B*3904 B*3910 B*3916 B*3919 B*3926
35	YYSREINTDESNLRYNFVLTW	B*3902 B*3922 B*3923
36	YYSRNINTDESNLSYNFVLTW	B*3903 B*3924
37	YYSRNINTDESNLTYNFVLTW	B*3906 B*3928
38	YHTREINTYESNLRYNYVLEW	B*4001 B*4004 B*4007 B*4011 B*4014 B*4046
39	YHTREINTYESNLSYNYVLEW	B*4002 B*4029 B*4035 B*4045
40	YHTREINTYESNLSYDYVLEW	B*4009 B*4018 B*4024 B*4031
41	YYAREINTYESNLRYNYVLEW	B*4010 B*4021
42	YYSREINTYESNLRYNYVLEW	B*4012 B*4803
43	YHTREINTYENIASYNYVLEW	B*4013 B*4019
44	YHTREINTYESNLSYNYEREW	B*4015 B*4016
45	YHTREINTYESNLRYNYELLW	B*4026 B*4028
46	YHTREINTYESNLVYNYVLEW	B*4030 B*4034
47	YHTREINTYESNLRYDYVLEW	B*4033 B*4042
48	YHTREINTYESNLRYNYVDTW	B*4101 B*4103 B*4106
49	YYSRNIQTDESNLSYNYVDTW	B*4201 B*4205
50	YYTREINTYENTARYDDVDLS	B*4402 B*4408 B*4422 B*4424 B*4427 B*4433
51	YYTREINTYENTARYDDVLLS	B*4403 B*4407 B*4413 B*4426 B*4429 B*4430 B*4432 B*4436 B*4438
52	YHTREINTYESNLRYNLVDLS	B*4501 B*4503 B*4505
53	YYTREINTYESNLRFHDVLEW	B*4702 B*4703
54	YYSREINTYESNLSYNYVLEW	B*4801 B*4804
55	YHTREINTYESNLRYNLELLW	B*5001 B*5004
56	YYARNINTYENIATYNYELLW	B*5101 B*5102 B*5107 B*5111 B*5112 B*5117 B*5118 B*5122 B*5124 B*5126 B*5128 B*5130 B*5132
57	YYARNINTYENIARYNYELLW	B*5104 B*5106
58	YYARNINTYENIATYNYVLLW	B*5109 B*5119
59	YYARNINTYENIATYNYELEW	B*5116 B*5134
60	YYAREINTYENIATYNYELLW	B*5201 B*5202 B*5204 B*5205
61	YYARNINTYENIARYDSVLLW	B*5301 B*5302
62	YYARNIQTDESNLTYNLVLTW	B*5401 B*5502 B*5507
63	YHARNIQTDESNLTYNLVLTW	B*5402 B*5516
64	YYARNIQTDESNLTYNLELTW	B*5501 B*5505 B*5515
65	YYARNIQTDESNLRYNLVLLW	B*5602 B*5604
66	YYARNIQTDESNLTYNYELLW	B*5605 B*5606
67	YYARENSTYENIAVYDSVLLW	B*5701 B*5706 B*5708
68	YYARENSTYENIARYDSVLLW	B*5801 B*5804 B*5809
69	YYARNINTDESNLTYNYELLW	B*7801 B*7803
70	YYSRNIQTDESNLSYNYVLEW	B*8101 B*8102
71	YYSRNIQTDESNLRFNLVDLS	B*8201 B*8202

CPFR = critical polymorphic functional residue

Table 4. HLA-C alleles (55) grouped into putative supertypes

Group	CPFR	HLA-Cw alleles
1	YYAREKQTDVNKLRYDSEWEW	Cw*0202 Cw*0208 Cw*0209
2	YYAREKQTDVSNLRYDYELLW	Cw*0303 Cw*0304 Cw*0309 Cw*0313
3	YSAREKQADVNKLRFNFERTW	Cw*0401 Cw*0405 Cw*0407 Cw*0409 Cw*0412
4	YYAQEKQTDVNKLRYNFERTW	Cw*0501 Cw*0503 Cw*0505 Cw*0506 Cw*0509
5	YDSREKQADVNKLWYDSEWTW	Cw*0602 Cw*0607
6	YDSRENQADVSNLRYDSALTW	Cw*0701 Cw*0706 Cw*0716 Cw*0718 Cw*0721 Cw*0724
7	YDSREKQADVSNLRSDSALTW	Cw*0702 Cw*0710 Cw*0717 Cw*0723 Cw*0725
8	YDSREKQADVSNLRYDFADTW	Cw*0704 Cw*0711 Cw*0712
9	YDSRENQADVNKLRYDSALTW	Cw*0707 Cw*0709
10	YYAQEKQTDVSNLRYNFTLTW	Cw*0801 Cw*0803
11	YYAQEKQTDVSNLRYNFERTW	Cw*0802 Cw*0807
12	YYAQEKQTDVSNLRYDSTLTW	Cw*0809 Cw*0811
13	YYAREKQADVSNLWYDSEWTW	Cw*1203 Cw*1206
14	YSAREKQTDVSNLWFDSERTW	Cw*1402 Cw*1403
15	YYARENQTDVNKLRYDLELTW	Cw*1502 Cw*1510 Cw*1512
16	YYARENQTDVNKLRYDSELTW	Cw*1504 Cw*1509
17	YYAREKQADVNKLRYNFELEW	Cw*1701 Cw*1702 Cw*1703
18	YDSREKQADVNKLRFNFERTW	Cw*1801 Cw*1802

CPFR = critical polymorphic functional residue

Table 5. Validation of HLA-A supertypes

Group	Peptide	Allele pair		Binding affinity	T-cell activity
		Category #1			
2	FLWGPRALV	A*0201	A*0209	YES (HIGH/?)	YES
	AAGIGILTV	A*0201	A*0209	YES (MOD/HIGH)	YES
	GILGFVFTL	A*0201	A*0209	YES (HIGH)	YES
	VLYRYGSFSV	A*0201	A*0209	YES (HIGH)	YES
	YLEPGPVTA	A*0201	A*0209	YES	YES
		Category #2			
4	GILGFVFTL	A*0206	A*0214	YES (?/LOW)	MOD
		Category #3			
22	RISGVDRYY	A*3002	A*3003	YES	?

MOD = moderate; ? = undetermined, The group numbers indicated in this Table refers to the group numbers designated in Table 2.

156W). The residues leucine (L) and tryptophan (W) are hydrophobic and the mutation is synonymous. Therefore, some peptides that bind A*0201 are also found to bind A*0202 (Table 1). This trend is not always true (Table 1). Therefore, A*0201 and A*0202 are not strict members of the A2 'sub-supertype'. In yet another case, the peptide RASGVDRYY binds A*3002 and A*3003 (Table 5). Table 2 shows that A*3002 and A*3003 are grouped together with another allele A*3012 (Group #22) and members of this group are proposed to bind similar peptides. Therefore, the binding of this peptide to A*3012 will be of great significance to validate this hypothesis. We further validated the grouping for B alleles using binding data for 2 pairs of alleles (B*2703/B*2705 and B*5101/B*5102) and the data is given in Table 6. The pairs B*2703/B*2705 (Group #27) and B*5101/B*5102 (Group #56) strictly bind similar peptides (Tables 3 and 6). More than 20 peptides were shown to bind each of the allele pairs clustered in Group #27 and Group #56. However, the binding of other members of the group is required for further extrapolation.

Here, we demonstrate using binding data for 3 pairs of HLA-A alleles (A*0201/A*0209 (Group #2), A*0206/A*0214 (Group #4) and A*3002/A*3003 (Group #22) that alleles within a group (Table 2) bind similar peptides (Table 5). The validation is extended to two pairs of B alleles B*2703/B*2705 and B*5101/B*5102 (Table 3 and Table 6). Thus, our methodology groups alleles into strict supertypes, where members always bind similar peptides (Tables 2 and 5). We also note that there may be some degree of functional overlap (Table 6) across members of different groups (A*0201 (Group 2) / A*0206 (Group #4)) due to synonymous residue substitutions at the CPFRP (Table 2). Several peptide binders of A*0201 are non-binders to A*0206 (Table 1). Hence, the functional overlap across groups (Table 1 and Table 2) is not always true. This warrants that grouping of HLA alleles into supertypes based on binding data is seldom conclusive and comprehensive (Table 1). The validation of the grouping strategy is limited in the current report. Further validation is required to apply this methodology to group alleles into 'sub-supertypes' from sequence information on a large scale. It is our hope that the clustering provided here will serve as a theoretical framework for investigating the phenomenon of HLA supertypes using binding data.

5. CONCLUSIONS

Knowledge on all possible combinations of MHCp binding is useful in the design of peptide vaccine candidates, immuno-therapeutic targets and diagnostics agents. The theoretically possible combinations are overwhelmingly large. However, the functional overlap between alleles and the grouping of alleles into 'sub-supertypes' is extremely powerful in understanding peptide selection and degeneration. Grouping of alleles into supertypes using binding data is seldom conclusive and comprehensive. The strategy described in this report, grouped 47% of known A alleles (295), 44% of known B alleles (540) and 35% of known C alleles (156) to just 36, 71 and 18 groups,

Table 6. Validation of HLA-B supertypes

Group	Peptide	Allele pairs		Binding affinity	T-cell activity
Category #1					
27	ARHGFLPRH	B*2703	B*2705	YES (MOD)	?
	ARTAHYGSL	B*2703	B*2705	YES	?
	ARYQKSTEL	B*2703	B*2705	YES	?
	FQYNGLIHR	B*2703	B*2705	YES	?
	FRYNGLIHR	B*2703	B*2705	YES	?
	GRAFVTIGA	B*2703	B*2705	YES	?
	GRAFVTIGK	B*2703	B*2705	YES	?
	GRERFEMER	B*2703	B*2705	YES (MOD)	?
	GRFFGGDRG	B*2703	B*2705	YES	?
	GRGLSLSRF	B*2703	B*2705	YES	?
	GRIDKPILA	B*2703	B*2705	YES	?
	GRIDKPILK	B*2703	B*2705	YES	?
	SRAHSSHLK	B*2703	B*2705	YES	?
	SRFSWGAEG	B*2703	B*2705	YES (LOW)	?
	SRHKKLMFK	B*2703	B*2705	YES (HIGH)	?
	SRSGSPMAR	B*2703	B*2705	YES (MOD)	?
	SRYWAITR	B*2703	B*2705	YES	?/YES
	VRRCPHHER	B*2703	B*2705	YES (LOW/MOD)	?
	VRVCACPGR	B*2703	B*2705	YES (MOD)	?
	RRYQKSTEL	B*2703	B*2705	YES	?
	QRHGSKYLA	B*2703	B*2705	YES (LOW/MOD)	?
	RRIKEIVKK	B*2703	B*2705	YES (MOD)	?
	RRTEEENLR	B*2703	B*2705	YES (MOD)	?
Category #2					
56	FPISPIETV	B*5101	B*5102	YES (HIGH/?)	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	FPVRPQVPL	B*5101	B*5102	YES	?
	CPKVSFEPI	B*5101	B*5102	YES	?
	CPSGHAVGI	B*5101	B*5102	YES	?
	DARAYDTEV	B*5101	B*5102	YES (HIGH/?)	NO/?
	EPLDLPQIIB	B*5101	B*5102	YES	?
	YPFKPPKVB	B*5101	B*5102	YES	?
	IPLGDAKLV	B*5101	B*5102	YES	?
	IPTSGDVVI	B*5101	B*5102	YES	?
	LPALSTGLI	B*5101	B*5102	YES	?
	LPCRIKQIIB	B*5101	B*5102	YES (LOW/?)	?
	LPEKDSWTV	B*5101	B*5102	YES	?
	LPPLERLTL	B*5101	B*5102	YES	?
	LPPTTGPPIB	B*5101	B*5102	YES	?
	LPPVVAKEI	B*5101	B*5102	YES (HIGH)	?
	NALFRNLDV	B*5101	B*5102	YES	?
	NANPDCKTI	B*5101	B*5102	YES	?
	NPPIPVGEIB	B*5101	B*5102	YES	?
	QGWKGSPAI	B*5101	B*5102	YES	?
	TAVQMAVFI	B*5101	B*5102	YES	?
	TGYLNTVTV	B*5101	B*5102	YES	?
	VAQRAYRAI	B*5101	B*5102	YES	?
	VGCLVGLRI	B*5101	B*5102	YES	?
	VPVKLKPGM	B*5101	B*5102	YES	?
	YAPPIGGQI	B*5101	B*5102	YES	?
	YPCTVNFTI	B*5101	B*5102	YES	?
	YPLASLKSL	B*5101	B*5102	YES	?
	YPLTSLRSL	B*5101	B*5102	YES	?
	APTLWARMI	B*5101	B*5102	YES	?

MOD = moderate; ? = undetermined, The group numbers indicated in this Table refers to the group numbers designated in Table 3.

respectively. This grouping procedure is useful because the binding of a peptide to ~50% of all known alleles can be inferred using a handful of binding data representing all predictive sub-supertypes. However, a comprehensive validation is required for large scale extrapolation. Some members across groups show overlapping function. However, the overlap across group is not always true. It should be noted that the methodology described here (9) is different from the procedure described by Doytchinova *et al.* (10) and Lund *et al.* (11) and they differ among themselves. We hope to establish consensus among these procedures in future investigations.

7. ACKNOWLEDGEMENTS

This research is supported by Nanyang Technological University, Singapore. P.K and M.K.S acknowledge the support from A*STAR-BMRC, Grant # 03/1/22/19/242. PS is supported by NIH grants DA014533 and GM056529.

8. REFERENCES

1. Robinson J., M.J. Waller, P. Parham, N. de Groot, R. Bontrop, L.J. Kennedy, P. Stoehr & S.G.E. Marsh IMGT/HLA and IMGT/MHC - sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* 31(1) 311-314 (2003)

2. Yewdell J.W., E. Reits & J. Neefjes: Making sense of mass destruction – quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.* 3(12) 952-961 (2003)

3. Del Guercio M.F., J. Sidney, G. Hermanson, C. Perez, H.M. Grey, R.T. Kubo & A. Sette: Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol.* 154(2) 685-693 (1995)

4. Sette A. & J. Sidney: Nine Major HLA class I supertypes account for the vast preponderance of HLA-A and –B polymorphism. *Immunogenetics* 50(3-4) 201-212 (1999)

5. Sette A. & J. Sidney: HLA supertypes and supermotifs - a functional perspective on HLA polymorphism. *Curr. Opin. Immunol.* 10(4) 478-482 (1998)

6. Sidney J., S. Southwood, V Pasquetto & A. Sette: Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. *J. Immunol.* 171(11) 5964- 5974 (2003)

7. Chelvanayagam G: A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics* 45(1) 15-26 (1996)

8. Zhang C, A. Anderson & C. DeLisi: Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol.* 281(5) 929-947 (1998)

9. Zhao B, A.E.H. Png, E.C Ren, P.R. Kolatkar, V.S Mathura, M.K Sakharkar & P. Kangueane: Compression of functional space in HLA-A sequence diversity. *Hum. Immunol.* 64(7) 718-728 (2003)

10. Doytchinova IA, P. Guan & D.R. Flower: Identifying human MHC supertypes using bioinformatics methods. *J. Immunol.* 172(7), 4314-4323 (2004)

11. Lund O., M, Nielsen, C. Kesmir, A.G. Petersen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, G. Roder, S. Justesen, S. Buus & S. Brunak: Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 55(12), 797-810 (2004)

12. Bhasin M, H. Singh & G.P. Raghava: MHCBN - a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19(5), 665-666 (2003)

13. Scognamiglio P., D. Accapezzato, M.A. Casciaro, A. Cacciani, M. Artini, G. Bruno, ML Chircu, J Sidney, S Southwood, S Abrignani, A Sette & V Barnaba. Presence of effector CD8+ T cells in hepatitis C virus-exposed healthy seronegative donors. *J. Immunol.* 162(11), 6681-6689 (1999)

14. Kawashima I, S.J. Hudson, V. Tsai, S. Southwood, K. Takesako, E. Appella, A. Sette & E. Celis: The multi-epitope approach for immunotherapy for cancer: identification of several CTL epitopes from various tumor-associated antigens expressed on solid epithelial tumors. *Hum. Immunol.* 59(1), 1-14 (1998)

15. Chang K, M., N.H. Gruener, S. Southwood, J. Sidney, G.R. Pape, F.V. Chisari & A. Sette: Identification of HLA-A3 and -B7-restricted CTL response to hepatitis C virus in patients with acute and chronic hepatitis C. *J. Immunol.* 162(2), 1156-1164 (1999)

Abbreviations: HLA: human leukocyte antigen, IC_{50}: inhibitory concentration 50, TCR: T-cell receptor, CPFR: critical polymorphic functional residue, CPFRP: critical polymorphic functional residue position, IMGT: ImMunoGeneTics,MHC: major histocompatibility complex, MHCBN: major histocompatibility complex binding and non-binding peptides, MHCp: MHC-peptide complex, MIF: molecular interaction fields, PCA: principal component analysis,HC: hierarchical clustering

Key Words: HLA, Supertypes, Sub-Supertypes, Functional Overlap, Vaccine Design, Review

Send correspondence to: Dr. Pandjassarame Kangueane, School of Mechanical and Production Engineering, 50 Nanyang Avenue, Nanyang Technological University, Singapore 639 798; Tel: +65 (6) 790-5836; Fax: +65 (6) 774-4340, E-mail: mpandjassarame@ntu.edu.sg

http://www.bioscience.org/current/vol10.htm