Pacific Symposium on Biocomputing 15:359-370(2010)

SEQUENCE FEATURE VARIANT TYPE (SFVT) ANALYSIS OF THE HLA GENETIC ASSOCIATION IN JUVENILE IDIOPATHIC ARTHRITIS

GLENYS THOMSON[†]

Department of Integrative Biology, University of California
Berkeley, CA, USA

NISHANTH MARTHANDAN

Department of Pathology, University of Texas Southwestern Medical Center
Dallas, TX, USA

JILL A. HOLLENBACH[1], STEVEN J. MACK[1], HENRY A. ERLICH[1,2]

[1]Children's Hospital Oakland Research Institute, Oakland, CA, USA
[2]Roche Molecular Systems, Pleasanton, CA, USA

RICHARD M. SINGLE

Department of Mathematics and Statistics, University of Vermont
Burlington, VA, USA

MATTHEW J. WALLER[3], STEVEN G. E. MARSH[3,4]

[3]Anthony Nolan Research Institute, Royal Free Hospital, London, UK, and
[4]UK and UCL Cancer Institute, Royal Free Campus London, UK

PAULA A. GUIDRY[5], DAVID R. KARP[6], RICHARD H. SCHEUERMANN[3]

[5]Departments of Pathology and [6]Internal Medicine, University of Texas Southwestern Medical Center
Dallas, TX, USA

SUSAN D. THOMPSON, DAVID N. GLASS

Cincinnati Children's Hospital Medical Center
Cincinnati, OH, USA

WOLFGANG HELMBERG

Department of Blood Group Serology and Transfusion Medicine, Medical University of Graz
Graz, Austria

The immune response HLA class II DRB1 gene provides the major genetic contribution to Juvenile Idiopathic Arthritis (JIA), with a hierarchy of predisposing through intermediate to protective effects. With JIA, and the many other HLA associated diseases, it is difficult to identify the combinations of biologically relevant amino acid (AA) residues directly involved in disease due to the high level of HLA polymorphism, the pattern of AA variability, including varying degrees of linkage disequilibrium (LD), and the fact that most HLA variation occurs at functionally important sites. In a subset of JIA patients with the clinical phenotype oligoarticular-persistent (OP), we have applied a recently developed novel approach to genetic association analyses with genes/proteins sub-divided into biologically relevant smaller sequence features (SFs), and their "alleles" which are called variant types (VTs). With SFVT analysis, association tests are performed on variation at biologically relevant SFs based on structural (e.g., beta-strand 1) and functional (e.g., peptide binding site) features of the protein. We have extended the SFVT analysis pipeline to additionally include pairwise comparisons of DRB1 alleles within serogroup classes, our extension of the Salamon Unique Combinations algorithm, and LD patterns of AA variability to evaluate the SFVT results; all of which contributed additional complementary information. With JIA-OP, we identified a set of single AA SFs, and SFs in which they occur, particularly pockets of the peptide binding site, that account for the major disease risk attributable to HLA DRB1. These are (in numeric order): AAs 13 (pockets 4 and 6), 37 and 57 (both pocket 9), 67 (pocket 7), 74 (pocket 4), and 86 (pocket 1), and to a lesser extent 30 (pockets 6 and 7) and 71 (pockets 4, 5, and 7).

[†] corresponding author: glenys@berkeley.edu

1. Introduction

Genes in the major histocompatibility complex of humans, termed HLA, are known to influence susceptibility to over 300 diseases. These include: complex autoimmune and inflammatory diseases such as type 1 diabetes, rheumatoid arthritis, ankylosing spondilitis, psoriasis, multiple sclerosis, and narcolepsy; nasopharyngeal cancer, Hodgkin disease and other cancers; infectious diseases including malaria, tuberculosis, and AIDS, and diseases of unknown etiology [1]. HLA allele or haplotype and genotype associations with specific diseases are well established; the most complex pattern is seen with type 1 diabetes and HLA DRB1-DQB1 haplotypes, with a hierarchy from very predisposing, through intermediate ("neutral"), to very protective effects, with consistent patterns in associations seen across ethnic groups [2]. Specific amino acid (AA) residues, as well as combinations of AAs, have been implicated in type 1 diabetes risk, see e.g., refs. [2-4] and references contained therein, as well as many other HLA associated diseases. The strong association of the HLA DRB1 "shared epitope" set of AAs 70-74 and rheumatoid arthritis is well established; recently autoimmunity to citrullinated protein antigens has been shown to define a clinically and genetically distinct subset of rheumatoid arthritis that is specifically associated with the "shared epitope" alleles (reviewed in [5]). The recent development of a novel approach to genetic association analyses with genes/proteins sub-divided into biologically relevant smaller sequence features (SFs), and their variant types (VTs) [6], allows a systematic search focusing on the most likely actual causative genetic variants in HLA associated diseases.

HLA molecules are cell-surface proteins that present peptide fragments to T-cells to activate the recognition, and response to, foreign antigens. The classical class I (HLA-A, -B, and -C) and class II (HLA-DR, -DQ, and -DP) genes encode structurally homologous heterodimers. The processed self and foreign peptide fragments presented by the classical HLA molecules usually consist of 8–10 AAs (class I) or 12–20 AAs (class II), and are of either intracellular or extracellular origin respectively. An HLA molecule binds only to peptides conforming to certain structural requirements. A particular HLA allele is thus only able to present a subset of the available peptides to T-cells. The polymorphism of many of the HLA genes is extraordinary, with over 3,000 alleles at the class I and II genes identified to date (www.ebi.ac.uk/imgt/hla/) (and see e.g., ref. [7]), with much of the variation present at the protein level and occurring at functionally important sites. The class II HLA DRB1 gene, the subject of this study, has 623 alleles defined at the AA level; for any specific population or disease study only a fraction of these alleles will be observed. There are multiple lines of evidence for the role of balancing selection (at the allele and AA levels) in maintaining this most polymorphic set of genes in the human genome, including relatively even allele frequency distributions, see e.g., refs. [7, 8] and references therein. The causal argument presented is that individuals heterozygous for HLA genes can more effectively defend themselves from infection by successfully responding to a broader range of pathogens.

Peptide motifs important for binding to HLA molecules, including critical residues, have been defined by sequence analysis of naturally processed peptides eluted from HLA molecules, analysis with synthetic peptides, phage display libraries, and predictive inference of binding preference based on similarity of peptide-binding environments, see e.g., [9-11]. Specific AAs and combinations thereof have been identified as potentially causal in a number of HLA associated diseases. These usually rely on differential risk effects within serogroups of alleles, which involve a more restricted set of AAs compared to overall allele level comparisons, or identification of patterns from the sequence alignments of polymorphic sites at all alleles, again stratified by risk categories, e.g., DRB1 and the "shared epitope" set of AAs and rheumatoid arthritis discussed above. Determining the critical HLA AA residues involved in a specific disease can facilitate predictions about peptide epitopes, which are important for the design of novel vaccines and the understanding of autoimmunity.

Other analysis methods have been applied to type 1 diabetes and other diseases, and have successfully identified AAs important in disease risk heterogeneity, e.g., with type 1 diabetes, the Unique Combinations algorithm of Salamon et al. [3], and the Conditional Haplotype Method in Valdes et al. [4]. The aim with Sequence Feature Variant Type (SFVT) analysis [6] is to systematically perform association tests focusing on variation (VTs) at biologically relevant SFs, which are based on structural and functional features of the protein. The SFs include classical HLA allele level, and single AA, polymorphisms. The second round of SFVT analysis tests disease associations of temporary SFs (tSFs). These tSFs are constructed based on potentially informative combinations of

AAs that are identified from the first round of SFVT analysis as occurring frequently in SFs with the strongest associations with disease, including SFs composed of a single AA. With systemic sclerosis, specific AAs in pockets 4 and 7 of the peptide binding site explain much of the molecular determinant of disease risk [6].

We have applied SFVT analysis to HLA DRB1 variation and Juvenile Idiopathic Arthritis (Oligoarticular-Persistent) (JIA-OP) data [12]. We have extended SFVT analysis by creating an automated pipeline to include additional complementary and informative analyses (see Figure 1 and the Methods and Results Sections): including pairwise comparisons of HLA DRB1 alleles, and our extension of the Salamon Unique Combinations algorithm [3], to detect single AAs and combinations thereof that uniquely define different risk categories of alleles. AAs implicated from these analyses are now additionally combined with those identified by the first round of SFVT analysis in the construction of tSFs. The calculation of linkage disequilibrium (LD) patterns of AA variability in controls is another addition to the SFVT pipeline, guiding our understanding of effects that may be due to high correlation of AA variation. The final step in the analysis pipeline, which is not yet automated, is to apply a series of Conditional Haplotype Method analyses to differentiate AA effects which may be directly causative in disease versus those whose associations can be explained by LD with a causative AA or set of AAs.

2. Data and Methods

2.1. *Juvenile Idiopathic Arthritis (oligoarticular-persistent) (JIA-OP) HLA DRB1 data*

HLA DRB1 high resolution (4 digit AA level variation) data on 354 JIA-OP patients and 273 controls were analyzed (see Table 1 later). See Hollenbach et al. [12] for details on the data set, HLA typing, standard HLA association studies and results, and the general background on HLA associations with JIA and its subtypes.

2.2. *Sequence feature variant type (SFVTs) analysis*

The 181 SFs for the HLA DRB1 gene, as well as SFs for the other classical HLA genes, are defined in Karp *et al.* [6]. These range in size from the entire polypeptide sequence to single amino acids and involve:

(a) structural features: e.g., allele level variation (SF1), beta-strand 2 (SF13);
(b) biological function: e.g., peptide antigen binding site (SF127), T-cell peptide antigen binding pocket 6 (SF136);
(c) sequence alteration: all single AA positions with sequence variation, e.g., AA position 57 (SF90); and
(d) combinational structural_functional: e.g., beta-strand 2_peptide antigen binding pocket 7 (SF152).

Each HLA DRB1 allele is defined as a vector with these 181 SFs and their respective "allelic variations", which are referred to as variant types (VTs). The VTs for each SF defined in [6] use DRB1*0101 as the base; for example, for SF135 which is defined by AA positions 70 and 71, the VT1 (70Q_71R) is found in DRB1*0101, DRB1*0102, DRB1*0403, and a number of other alleles, while VT2 (70D_71E) is found in DRB1*0103, *0402, *1102 and a number of other alleles, etc. For the analysis of each specific SF, the VT frequencies are obtained by adding over the respective DRB1 alleles that carry that VT in the patient and control groups.

The DRB1 typing system used for the JIA-OP data set is based on identification of polymorphisms within exon 2, and hence we focus our attention on polymorphic AA variation within exon 2 (AAs 9-86). After initial analysis of the SFVT results, and combining information from the other analyses described below (Figure 1), so-called temporary SFs (tSFs), and their respective VTs, were defined based on sets of AAs with evidence of a role in differential disease risk.

2.3. *Chi-square heterogeneity testing*

A standard chi-square test was performed for heterogeneity testing of patient versus control DRB1 allele counts at the overall level, in pairwise comparisons, in the relative predispositional effects (RPE) analysis described below, and in SFVT analyses. Counts were combined in a "binned" category if the expected patient or control counts in the heterogeneity test were < 5. If the "binned" class had an expected control count < 5 it was not considered. The overall allelic effects in terms of disease risk are ranked by the standard Odds Ratio (OR) from most predisposing to most protective (see Table 1 later). The *p*-values are not corrected for multiple comparisons, since we are using these

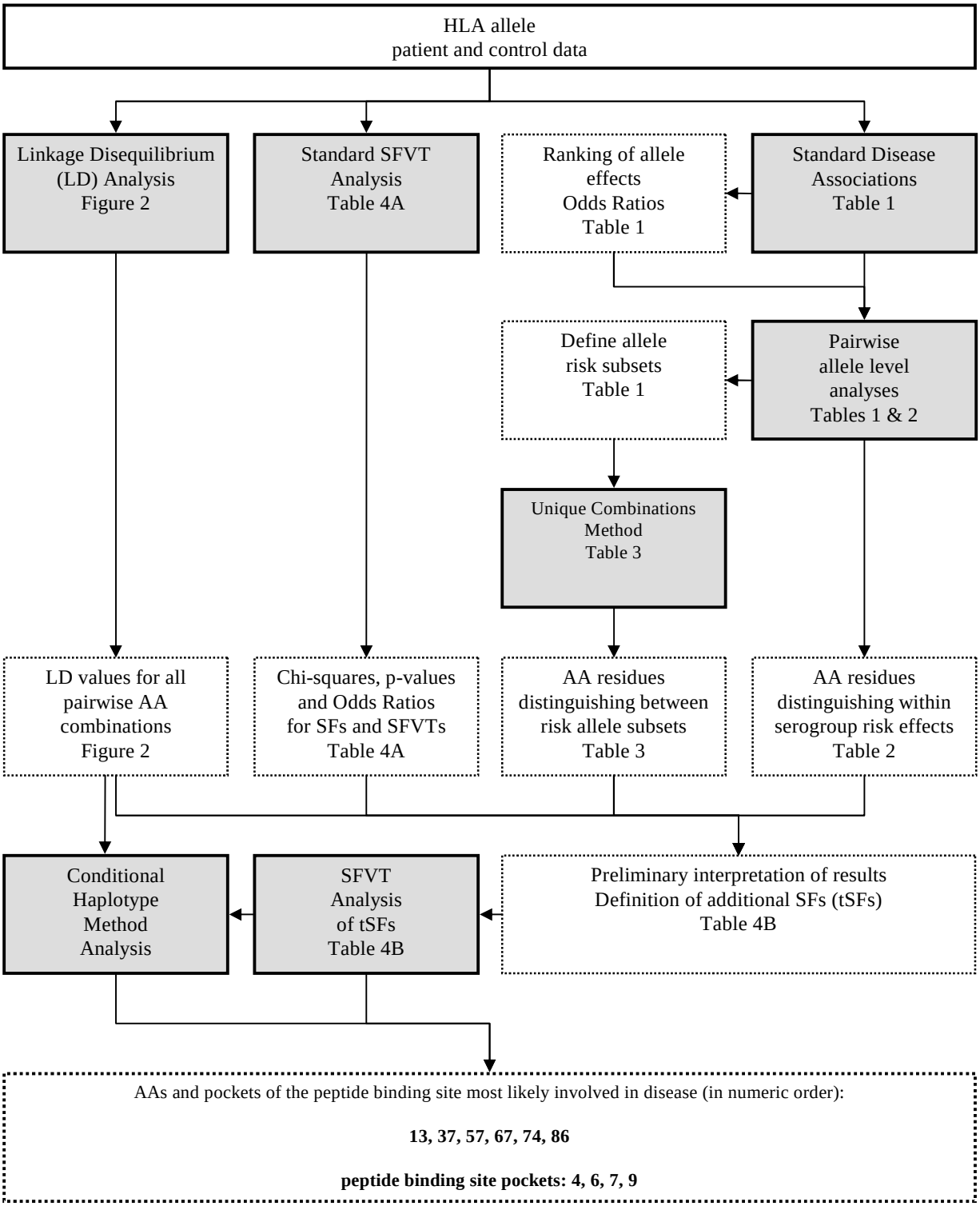

Figure 1. Analysis pipeline of SFVT analysis.

Methods are symbolised by solid boxes with grey shading, results are depicted as dotted boxes.

primarily for the purpose of relative rankings in this exploratory analysis; note that the AAs and SFs listed in this analysis as important in disease risk would still be very significant after correction if it were applied.

2.4. *Relative Predispositional Effects (RPE) analysis*

In Relative Predispositional Effects (RPE) analysis [13], alleles are sequentially removed from the standard chi-square heterogeneity test, starting with the strongest effects (based on the individual contribution of the allele to the overall chi-square), until there is no further significant heterogeneity detected either overall or for an individual allele. Note that this weights targeting and removal of the more common alleles; there is no intrinsic bias except for this factor. When there are relatively similar predisposing and protective effects—in terms of the strengths of their effects—in a particular round of analysis, these were both removed from the next round.

2.5. *Linkage disequilibrium (LD)*

The correlation measure W_n, also known as Cramer's V statistic, is used to quantify the overall (global) strength of LD among SFs; each SF is considered as a separate locus ("gene"), with the VTs the "alleles" thereof as in standard LD estimation [14]. For two bi-allelic loci, W_n is equivalent to the correlation coefficient ($r = \sqrt{D_{11}^2 / p_1 p_2 q_1 q_2}$), where p_1 and p_2 are the allele frequencies at the first locus ($p_1 + p_2 = 1$), and similarly q_1 and q_2 for the second locus, and $D_{11} = x_{11} - p_1 q_1$, where x_{11} is the frequency of the haplotype composed of these two alleles ($D_{11} = D_{12} = D_{21} = D_{22}$). The multi-allelic extension, W_n, is defined as $W_n = \frac{\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} D_{ij}^2 / p_i q_j}}{\sqrt{\min(I-1, J-1)}}$, where p_i and q_j are the allele frequencies at the first and second loci with I and J alleles respectively, and the individual LD coefficients are $D_{ij} = x_{ij} - p_i q_j$. Both r and W_n have a range from 0 to 1. When all $D_{ij} = 0$, $r = 0$ for bi-allelic loci, and $W_n = 0$ for multi-allelic loci, and the alleles at the two loci are randomly associated. In contract, for high LD the alleles at the two loci are very highly correlated, with complete correlation when r or $W_n = 1$.

2.6. *Salamon's Unique Combinations algorithm*

The Unique Combinations algorithm developed in [3] efficiently identifies combinations of AAs that distinguish a sequence or set of sequences from a set of other sequences. See Section 3.4 in Results for more details.

3. Results

3.1. *Summary*

Given the complexity of the integration of results and incorporation of information from within and between various levels of analyses, e.g., pairwise allele level comparisons, Unique Combinations analyses, LD patterns, and SFVT analysis (Figure 1), we summarize the findings now to aid navigation through the series of results detailed below. The AAs we identify as major or important in JIA-OP disease risk, always indicated in bold below are (listed in numeric order): **AAs 13** (pockets 4 and 6), **37** and **57** (both in pocket 9), **67** (pocket 7), **74** (pocket 4), **86** (pocket 1), and those potentially involved in disease risk are (underlined): <u>30</u> (pockets 6 and 7), and <u>71</u> (pockets 4, 5, and 7).

3.2. *HLA DRB1 allele level analyses*

A total of 38 DRB1 alleles were observed: 17 were included in the "binned" category, leaving 21 frequent alleles (Table 1). (Note that for the SFVT analysis, the rare DRB1 alleles are always included.) As previously described [12] there is significant heterogeneity in allele counts between patients and controls (overall test: $p < 1.1E\text{-}27$). Two predisposing alleles: DRB1*0801 and *1104, and three protective alleles: DRB1*1501, *0701, and *0401, show very strong individual effects (based on their *p*-values), with weaker effects of DRB1*1301, *1103, *0404, and *0103. Note that for the rarer alleles, e.g., DRB1*1103 which has the highest OR, the 95% OR confidence interval (CI) spans a large range; hence conclusions from analyses that use this information are subject to this uncertainty.

The first column of Table 1 lists the alleles (labeled in categories 1-3) which were sequentially removed in successive rounds in the Relative Predispositional Effects (RPE) analysis until no significant differential effects were seen. Note that (as mentioned in Section 2.4) the higher frequency alleles are targeted with this analysis. In the second column, the more common alleles are divided into a set (A) of three differential risk categories containing high frequency alleles: I (predisposing), II (neutral), and III (protective). The boundaries, and inclusion of alleles in each category, were predicated on evidence of disease risk heterogeneity between categories, and homogeneity within categories. In the third column, the list of alleles (set B) in each risk category is expanded to include some rarer alleles, but with the boundaries still delineated by more common alleles; inclusion in each respective risk category is now indicated by Ix, IIx, and IIIx. The rare allele DRB1*0403 is not included as it cannot be classified as predisposing versus neutral, similarly the four rare alleles between sets IIx and IIIx (the neutral/protective boundary). These sets A and B of differential risk categories are used below in the Unique Combinations analyses.

Table 1. JIA-OP HLA DRB1 allele data ranked by Odds Ratio (OR)

RPE[a]	A[b]	B[c]	DRB1	JIA-OP	Controls	Chi-sq.	p-value[d]	OR	CI[e]	CI[e]
		Ix	*1103	12	1	6.80	0.01	9.40	1.22	72.49
1	I	Ix	*0801	102	13	48.61	3.1E-12	6.90	3.83	12.43
2	I	Ix	*1104	57	11	20.71	5.3E-06	4.26	2.21	8.20
			*0403	9	3	1.68	0.20	2.33	0.63	8.65
3	II	IIx	*1301	90	38	9.99	0.002	1.95	1.31	2.90
		IIx	*0102	9	5	0.35	0.55	1.39	0.46	4.18
	II	IIx	*1101	60	36	1.42	0.23	1.31	0.85	2.02
		IIx	*0901	9	6	0.08	0.78	1.16	0.41	3.28
	II	IIx	*0101	74	50	0.52	0.47	1.16	0.79	1.69
	II	IIx	*0301	89	61	0.50	0.48	1.14	0.81	1.62
		IIx	*1201	10	8	0.006	0.94	0.96	0.38	2.46
	II	IIx	*1302	28	23	0.05	0.82	0.94	0.53	1.64
			*1303	10	9	0.11	0.74	0.86	0.34	2.12
			binned[f]	27	27	0.92	0.34	0.76	0.44	1.31
			*1601	6	8	1.05	0.30	0.58	0.20	1.67
			*1401	11	18	4.05	0.04	0.46	0.22	0.99
			*1502	5	10	3.26	0.07	0.38	0.13	1.12
	III	IIIx	*0404	7	16	6.34	0.01	0.33	0.14	0.81
1	III	IIIx	*1501	38	80	28.24	1.1E-07	0.33	0.22	0.49
1	III	IIIx	*0701	30	65	23.92	1.0E-06	0.33	0.21	0.51
2	III	IIIx	*0401	21	47	18.10	2.1E-05	0.33	0.19	0.55
3		IIIx	*0103	4	11	5.42	0.02	0.28	0.09	0.87
			TOTAL	708	546	182.1	1.1E-27			

[a] Numbers denote the order of removal due to largest effect(s) in the RPE analysis
[b] Set A: The common alleles are divided into mutually exclusive, and significantly different, predisposing (I), intermediate (II), and protective (III) categories for use in the Unique Combinations comparisons
[c] Set B: The sets I, II, and III above are expanded to include rare alleles, while excluding those alleles which do not clearly fall into one of the 3 risk categories
[d] The individual *p*-values are biased (conservative with respect to finding significant effects) as the assumption of a 1 df chi-square is incorrect; the *p*-values can be used however for a relative ranking of the allelic effects
[e] The upper and lower 95% confidence intervals (CIs) for the Odds Ratio (OR) are given
[f] The binned category consists of all alleles with an expected value < 5 under the chi-square test of heterogeneity of patient and control allele counts

3.3. *HLA DRB1 allele level pairwise within serogroup analyses*

HLA nomenclature (except for DP) is such that alleles sharing the same 2 first digits generally belong within the same serogroup. Variation in the 2 last digits indicates AA differences within the serogroup. For example, DRB1*0101, *0102, and *0103 belong to the serogroup denoted *01XX. Alleles within the same serogroup are more closely related at the AA level, hence significant differences in risk within serogroups, and between specific pairs of alleles, may identify specific AAs, or a few AAs, involved in disease. Pairwise comparisons within serogroups of alleles with sufficient sample size—DRB1*01XX, *04XX, *11XX, and *13XX—were performed; this was followed by manual inspection of their respective sequences for comparisons where significant risk heterogeneity was detected. Significant results (ordered by *p*-value) are given in Table 2. The strong evidence for the role

Table 2. JIA-OP HLA DRB1 allele pairwise comparisons

Alleles compared[a]	p-value[b]	AAs[c,d]
*0403 vs *0401 + *0404	0.002	**74**
*1104 vs *1103	0.003	**86**
*0403 vs *0401	0.004	71, **74**, or **86**
*0101 vs *0103	0.02	**67**, 70, or 71
*1103 vs *1101	0.04	71 or **86**
*1301 vs *1302	0.05	**86**

[a] Within serogroup allele comaprisons
[b] Uncorrected p-value from the chi-square test of heterogeneity
[c] Amino acid residues that uniquely define these specific alleles
[d] Amino acids indicated in **bold** are those identified as playing a major or important role in disease risk, those underlined as potentially having an effect, albeit weaker

of **AA 86** in differential disease risk is of particular interest, since with SFVT analysis this AA shows *no* significant effect (see Section 3.6 and Table 4A below). There is also evidence for a direct role of **AA 74**.

3.4. *Unique Combinations comparisons*

In the original Unique Combinations algorithm of Salamon et al. [3], two categories of sequences are defined by the user: those in the "check" category are compared against those in the "group" category in order to identify combinations of sites that are unique between these two sets of sequences. However, when there are two or more sequences in the "check" category, sites that are polymorphic between these "check" sequences are excluded from consideration. We have extended the algorithm to allow inclusion of all sites that are polymorphic in the "check" category, thus expanding the utility of the method. Also, this means that the "group" and "check" categories are now interchangeable, whereas before there was an asymmetry.

This extension of the Unique Combinations algorithm provides an ordered list of a minimal number of polymorphic positions, which as a haplotype (combination of AAs on a chromosome) can differentiate between any set of sequences of alleles in the "check" versus "group" categories. Deriving the vectors of AAs that correspond to the resulting minimal unique combination generates unique sequences that either belong to the "check" category or the "group" category.

Using the subdivisions of common DRB1 alleles into the three categories defined above as sets A and B (Table 1)—I and Ix (predisposing), II and IIx (intermediate), and III and IIIx (protective)—we performed various Unique Combinations comparisons of each risk category versus the other two risk categories, e.g., I versus II + III (Table 3). The AAs identified as important in the Unique Combinations analyses are **AA 86** (as in the pairwise allele within serogroup analyses in Section 3.3 above) combined with **AAs 13** and **37**, or **13** and **67**.

Table 3: JIA-OP HLA DRB1 Unique Combinations (UC) analyses

group[a]	I	Ix	III	IIIx	II	IIx	
check[a]	II + III	IIx + IIIx	I + II	Ix+IIx	I + III	Ix+IIIx	AAs[b]
			X				**13**
				X			**13, 67**
			X	X			**37, 67**
	X	X			X		**13, 37, 86**
	X	X			X	X	**13, 67, 86**

[a] The sets of predisposing (I and Ix), intermediate (II and IIx), and protective (III and IIIx) alleles are defined in Table 1, the group and check categories define the two groups of alleles compared

[b] Only AAs in exon 2 that are consistently seen in all comparisons are listed in this column, other AAs which appear in some comparisons are: 47, 57, 70, 71, 74

3.5. *HLA DRB1 amino acid LD patterns*

The AA LD values in the control data for exon 2 of the HLA DRB1 variation (AAs 9-86) are given in Figure 2. This information is used in evaluation of the SFVT data below. The LD values show a complex pattern with (using examples from the six AAs we have identified as most strongly implicated in disease risk): (1) "blocks" of AAs where adjacent sites all have high LD with each other (**13** and 9-12); (2) individual AAs, each with high and moderate levels of LD with quite a few other AAs (**13** and **37**), and similarly but with lower levels of LD (**57** and **74**); and (3) individual AAs with very low levels of LD with most or all other AAs (**67** and **86**).

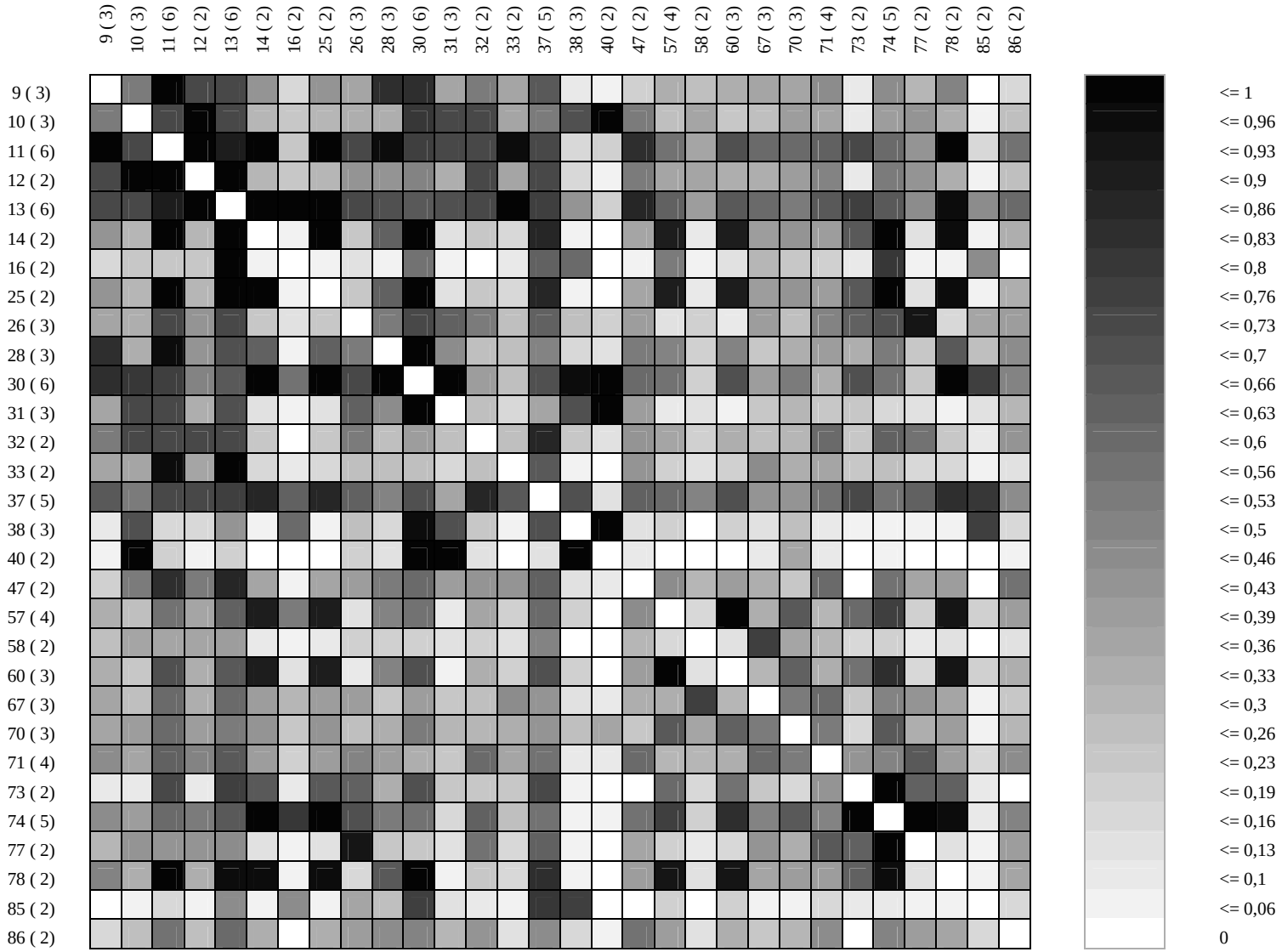

Figure 2. Linkage disequilibrium (LD) plot of polymorphic amino acids 9 - 86 of HLA DRB1 [a]

[a] The data are from the control samples; the axes give the amino acid positions and the number of residues (in parentheses) at each position

3.6. *SFVT analyses (Table 4, part A)*

The SFVT data in Table 4A are ranked by the *p*-value of the overall chi-square heterogeneity patient versus control analysis of the VTs for each SF listed. Note that any attempt to draw conclusions from minor differences in *p*-values of this magnitude is an over-interpretation of the data. Also listed are the maximum (max) and minimum (min) ORs seen for *individual* VTs for each SF, for example, for SF1 (allele) (rank 5 in Table 4A), these are 9.40 and 0.28 (see Table 1). These values are used to determine the ability of a SF to differentiate between risk categories of VTs, both with consideration of the max and min ORs and the range of the ORs for a particular SF. However, keep in mind, as mentioned above, that the highest OR of 9.40 for allele DRB1*1103 is based on a very rare allele, the next highest allele level OR is 6.90 for DRB1*0801 (see Table 1).

AA 13, pockets 4 and 6: The single **AA 13** (SF57) is rank 1 in the SFVT analysis (Table 4A), with a reasonable range in OR values (4.91 to 0.33). From Table 5, we see how **AA 13** by itself partitions the disease risk (although certainly not perfectly): the residues G and S are only seen in the predisposing and neutral allele level disease risk categories, while the other 4 residues: F, R, H and Y are only seen in the neutral and protective categories. **AA13** is the major contributor to the **pocket 6** and **pocket 4** associations (ranks 2 and 3). AA 30 (**pocket 6**) may play some role in disease risk, but the effect may be explained by LD.

The effects of the single AAs 9, 10, 11, 12 and 16, which occur in the top 20 ranked SFs in Table 4A, can be explained by LD with **AA 13** (Figure 2). These AAs are indicated in *italics* in Table 4A, and are not discussed individually below. **AA 13** is chosen over these AAs since it individually has a stronger effect, based on *p*-values and OR values and range, it occurs in more top ranked SFs than do any of these AAs, and it was identified individually over these other AAs, in combination with **AAs 37** and **86**, or **67** and **86**, in the Unique Combinations analyses.

AA 67, pocket 7: AA 67 (SF98, rank 14) has the second highest rank of the single AAs, and is the major contributor to the SFs ranked above it and below SF1 (allele) (excluding the 3 single AA SFs AAs *10*, *11*, and *12*, see above), and was identified in the Unique Combinations analyses. While AA 71 (SF102, rank 22) contributes an additional effect to **AA 67** by itself, due to greatly increasing the OR max value of **pocket 7** (rank 6), this in fact is a minor effect (in this data set) reflecting unique identification of the very *rare* predisposing DRB1*1103 allele.

AAs 13, 67, 74, and 86, and pockets 6, 4, and 7: AAs 13 and **67** together distinguish all the *significant* effects of the alleles listed in Table 1, with two exceptions which are covered by **AAs 74** and **86**. **AA 74** (the third ranked highest single AA effect (SF104, rank 16)) was identified in the allele level pairwise comparisons (Table 2) as necessary to distinguish between the *rare* neutral/predisposing DRB1*0403 allele and the protective *0401 and *0404 alleles ($p < 0.002$). Note that this effect of **AA 74** was not picked up in the Unique Combinations analyses, since our inability to definitively place DRB1*0403 in either the predisposing or neutral categories precluded its consideration in the Unique Combinations analyses. **AA 86** was identified in the allele level pairwise comparisons as necessary to distinguish between the predisposing DRB1*1104 and neutral *1101 effects ($p < 0.003$) (Table 2); it was also picked up in the Unique Combinations analyses. However, in the SFVT analysis (Table 4A), note that the individual effect of **AA 86** is *not* significant, nor is it identified as potentially involved in disease from its presence in other SFs: it is the 8th *lowest* ranked SF (SF110, rank 42), and pocket 1 (SF132) in which it occurs ranks two below this. However, as indicated in Table 2, **AA 86** is *necessary* to explain significant disease risk effects. The four **AAs 13, 67, 74,** and **86** uniquely define all the alleles listed in Table 1 (see Table 5), not considering the very rare alleles in the binned category in Table 1.

AAs 37 and 57 (pocket 9): AAs 37 (SF74, rank 17) and **57** (SF90, rank 19) are the next to enter into our consideration. The Unique Combinations results (Table 2) implicate **13**, **37**, and **86** as a potential AA combination that can explain disease risk, along with **13**, **67**, and **86** as an alternate combination. Further, to jump ahead, the AA combination **13** and **37** together have rank 1 in the analysis of tSFs discussed below (Table 4B). The combination of **AAs 13, 37, 74**, and **86** also explains all known disease risk at the allele level, except the marginally significant allele pairwise comparison of DRB1*0101 (neutral) to the *rare* *0103 (protective) ($p < 0.02$); **AA 67** distinguishes this risk, as does AA 71; as above AA 71 also further distinguishes DRB1*1103 and *1104, which in this data set are not significantly different in their effects.

Table 4. SFVT results

Rank	SF #	Description	amino acid[a]	*p*-value	max OR	min OR
A: JIA-OP HLA DRB1 SFVT analysis with SFs ranked by overall *p*-values						
1	57	**position 13**	**13**	1.8E-28	4.91	0.33
2	136	**pocket 6**	*11*, **13**, 30	3.9E-28	7.07	0.31
3	134	**pocket 4**	**13**, 26, 28, 70, 71, **74**, 78	5.7E-28	6.84	0.28
4	151	*beta1_pep ant-TCR*	*11*, **13**	9.4E-28	4.89	0.33
5	1	**allele**[b]		1.1E-27	9.40	0.28
6	137	**pocket 7**	28, 30, 47, 61, **67**, 71	8.9E-27	9.40	0.28
7	142	pep ant position 12	*9*, 56, **57**, 60, 61, **67**	4.2E-26	7.14	0.31
8	55	*position 11*	*11*	8.8E-25	3.15	0.33
9	130	pep ant & T cell rec	60, **67**, 70, 71, 77, 78, 85	7.0E-24	9.40	0.32
10	56	*position 12*	*12*	1.1E-22	3.15	0.32
11	54	*position 10*	*10*	1.9E-22	3.14	0.32
12	162	alpha2_pocket 7	**67**, 71	2.9E-21	9.40	0.33
13	19	alpha-helix 1	52..62	3.5E-18	3.92	0.44
14	98	**position 67**	**67**	3.0E-17	3.39	0.54
15	138	**pocket 9**	9, **37, 57**	3.6E-16	3.92	0.33
16	104	**position 74**	**74**	3.8E-16	6.84	0.33
17	74	**position 37**	**37**	3.7E-13	1.80	0.34
18	59	*position 16*	*16*	5.4E-13	4.91	0.20
19	90	**position 57**	**57**	5.5E-13	3.92	0.44
20	53	*position 9*	*9*	2.3E-11	2.30	0.42
21	135	pocket 5	70, 71	1.4E-10	1.79	0.33
22	102	position 71	71	1.2E-09	1.48	0.33
23	101	position 70	70	4.5E-09	2.03	0.54
24	71	position 33	33	3.1E-07	2.62	0.38
25	58	*position 14*	*14*	3.6E-07	3.05	0.33
26	63	*position 25*	*25*	3.6E-07	3.05	0.33
27	91	position 58	58	1.0E-06	2.35	0.43
28	24	position 78	78	3.3E-06	2.56	0.39
29	93	position 60	60	3.5E-06	2.36	0.44
30	68	position 30	30	2.3E-05	1.55	0.33
31	152	beta2_pocket 7	28, 30	4.3E-05	1.55	0.33
32	155	beta2_pocket 4	26, 28	4.4E-05	1.31	0.34
33	141	pep ant position 4	77, 78, 81, 82, 85	7.4E-05	1.37	0.39
34	153	beta2_pep ant bind	26, 28, 30	0.0001	1.32	0.33
35	13	beta-strand 2	23..32	0.0002	1.29	0.33
36	66	position 28	28	0.004	1.54	0.63
37	22	position 73	73	0.007	1.47	0.68
38	81	position 47	47	0.03	1.28	0.78
39	178	beta1_CD4 bind	41..56	0.03	1.28	0.78
40	70	position 32	32	0.07	1.25	0.80
41	154	beta2_alpha chain	29, 31, 32	0.13	1.27	0.80
42	110	**position 86**	**86**	0.34	1.12	0.90
43	106	position 77	77	0.41	1.16	0.86
44	132	pocket 1	82, 85, **86**, (89, 90)	0.58	1.14	0.89
45	173	alpha4_pocket1	82, 85, **86**	0.63	1.13	0.90
46	64	position 26	26	0.67	1.16	0.93
47	109	position 85	85	0.74	1.13	0.88
48	69	position 31	31	0.86	1.03	0.97
49	75	position 38	38	0.91	1.04	0.96

Rank	SF #	Description	amino acid[a]	p-value	max OR	min OR
B: JIA-OP HLA DRB1 SFVT analysis of tSFs ranked by overall *p*-values						
1	t201		**13**, **37**	5.6E-30	7.04	0.37
2	t205		**13**, **37**, **74**, **86**	2.7E-29	6.69	0.37
3	t211		**13**, 30, **37**, **74**, **86**	3.4E-29	6.69	0.37
4	t210		**13**, 30, **67**, **74**, **86**	1.2E-28	6.47	0.28
5	t206		**13**, **67**, **74**, **86**	1.2E-28	6.47	0.28
6	t224		**13**, **37**, **67**	1.3E-28	6.84	0.28
7	57	position **13**	**13**	1.8E-28	4.91	0.33
8	t225		**37**, **67**	2.2E-28	3.90	0.31
9	t203		**13**, **67**	2.3E-28	6.84	0.28
10	t204		**13**, **67**, **86**	2.5E-28	6.69	0.28
11	t218		**13**, **37**, 57, **67**, **74**, **86**	3.6E-28	6.90	0.28
12	t207		13, **37**, **67**, **74**, **86**	3.3E-28	6.47	0.28
13	136	pocket 6	11, **13**, 30	3.9E-28	7.07	0.31
14	t212		**13**, 30, **37**, **67**, **74**, **86**	4.4E-28	6.47	0.28
15	t214		**13**, 57, **67**, **74**, **86**	4.5E-28	6.90	0.28
16	151	beta1_pep ant-	*11*, **13**	9.4E-28	4.89	0.33
17	t216		**13**, 30, **37**, 57, **67**, 71, **74**, **86**	1.1E-27	9.40	0.28
18	t215		**13**, 57, **67**, 71, **74**, **86**	1.1E-27	9.40	0.28
19	1	allele		1.1E-27	9.40	0.28
20	137	pocket 7	28, 30, 47, 61, **67**, 71	8.9E-27	9.40	0.28

[a] Amino acids indicated in **bold** are those identified as playing a major or important role in disease risk, those underlined as potentially having an effect, albeit weaker, and for those in *italics* their effect may be explained by LD with **AA 13**.

[b] SF127 (peptide antigen binding site) AAs: 9, 11, 13, 26, 28, 30, 37, 47, 56, 57, 60, 61, 67, 70, 71, 74, 77, 78, 81, 82, 85, 86, 89, 90 (not listed), has identical VT counts as SF1 (allele)

Other AAs: Single AAs and other SFs listed below rank 20 in Table 4A either do not themselves have a wide range of OR values, and/or their effects may be explained by LD with other AAs or SFs, or they do not have a significant overall effect. However, we note that AAs 9, 10, 11, 12, 26, 28, 47, 56, 58, 60, and 85 also show up frequently in the top SFVT results, including comparisons based on *p*-values for individual VTs of each SF, and these should be considered in additional analyses.

3.7. *SFVT Analysis of Temporary SFs (tSFs) (Table 4, part B)*

The SFVT data in Table 4B are the most relevant data from analyses of tSFs defined by specific potentially informative combinations of the **AAs 13, 37, 57, 67, 74, 86,** 30, and 71 (based on the analyses above). Again, these are ordered by *p*-value, but note that the range of *p*-values in Table 4B shows only minor differences, and the effects we will concentrate on hence focus more on the max and min OR and the range of OR values. The SFs from Table 4A which include the span of *p*-values of the tSFs are also included in Table 4B for comparison.

Table 5: JIA-OP HLA DRB1 Amino Acid Residue Variation

	AA position	13	67	74	86	37	57	30	71
	# alleles	6	3	4	2	5	5	5	4
OR[a]	DRB1 alleles								
9.4	DRB1*1103	S	F	A	V	Y	D	Y	E
6.9	DRB1*0801	G	F	L	G	Y	S	Y	R
4.26	DRB1*1104	S	F	A	V	Y	D	Y	R
2.33	DRB1*0403	H	L	E	V	Y	D	Y	R
1.95	DRB1*1301	S	I	A	V	N	D	Y	E
1.39	DRB1*0102	F	L	A	V	S	D	C	R
1.31	DRB1*1101	S	F	A	G	Y	D	Y	R
1.16	DRB1*0901	F	F	E	G	N	V	G	R
1.16	DRB1*0101	F	L	A	G	S	D	C	R
1.14	DRB1*0301	S	L	R	V	N	D	Y	K
0.96	DRB1*1201	G	I	A	V	L	V	H	R
0.94	DRB1*1302	S	I	A	G	N	D	Y	E
0.86	DRB1*1303	S	I	A	G	Y	S	Y	K
0.58	DRB1*1601	R	F	A	G	S	D	Y	R
0.46	DRB1*1401	S	L	E	V	F	A	Y	R
0.38	DRB1*1502	R	I	A	G	S	D	Y	A
0.33	DRB1*0404	H	L	A	V	Y	D	Y	R
0.33	DRB1*1501	R	I	A	V	S	D	Y	A
0.33	DRB1*0701	Y	I	Q	G	F	V	L	R
0.33	DRB1*0401	H	L	A	G	Y	D	Y	K
0.28	DRB1*0103	F	I	A	G	S	D	C	E

[a] Odds Ratio (OR), see Table 1

The pair of **AAs 13** and **37** (SFt201, rank 1 in Table 4B) captures the JIA-OP disease risk with 11 VTs, including a good range of OR values, as does the pair **13 and 67** (SFt203, rank 9) with 13 VTs. Both increase the max OR, but barely change the min OR, compared to **AA 13** by itself (SF57, rank 7, 6 VTs). The addition of **AAs 74 and 86** to both these combinations (SFt205, rank 2: and SFt206, rank 5) adds no further discrimination based on the SFVT analysis, but as noted above are required to account for all disease risk heterogeneity.

Only addition of AA 71 gives the full range of ORs seen with DRB1 allele level variation (SF1, rank 19) (see SFt216, rank 17 and SFt215, rank 18); as noted above, AA 71 distinguishes the rare highest predisposing risk DRB1*1103 allele from the predisposing (common) DRB1*1104 allele.

4. Discussion

While many HLA associations at the allelic level are well documented and consistently found, identification of the AAs directly involved in disease risk is difficult, due to the nature of HLA polymorphism (discussed in the Introduction). SFVT analysis was designed to facilitate the analysis of HLA disease associations by focusing on structurally and functionally important subsets of HLA variation [6]. SFVT analyses can also be informative in study of other highly polymorphic genes such as HIV gp160 and influenza hemagglutinin.

Our SFVT analysis of HLA DRB1 and JIA-OP, with an expanded pipeline (Figure 1) to integrate results from within serogroup comparisons of alleles, Unique Combinations analysis of alleles in distinct risk categories, and study of LD patterns of individual AAs, identified the following HLA DRB1 AAs as major or important JIA-OP risk factors

(listed in numeric order): **13**, **37**, **57, 67**, **74**, and **86**; with lesser effects of 30 and 71. These AAs are all contained within the peptide binding site, and explain the increased disease risk associated with **pockets 4, 6, 7, and 9** of the peptide binding site. When there is high LD between AAs, it is difficult to distinguish the causative agent, e.g., **AA 13** is in high LD with a number of other AAs, and further study of these AAs is required. Note however, that when the LD is very high it may be impossible to distinguish between highly correlated AA sites; studies in other ethnic groups where the LD pattern may be different are then useful. The AAs we have focused on certainly do not represent all of the HLA DRB1 contribution to JIA-OP. Note that defining a minimal set of AAs that describe all AA allele level variation does not mean that additional AAs do not have effects on disease risk.

Further study of the various combinations of the AAs we have identified and their contributions to JIA-OP risk is also required. Preliminary analyses, using the Conditional Haplotype Method [4, 15 and references therein] show significant heterogeneity, albeit with somewhat weak effects, of **13 + 37** versus **67**, and **13 + 67** versus **37**, i.e., neither of the pairwise combinations alone can explain all of the disease risk. However, the two combinations of **AAs: 13, 37, 74, 86** and **13, 67, 74, 86** (Table 5); both uniquely define all JIA-OP disease risk as well as all the alleles listed in Table 1, except for the non-significant effects of the predisposing DRB1*1103 (*rare*) and *1104 alleles, which is distinguished by AA 71.

We have shown the power of inclusion of complementary analyses in the SFVT pipeline (Figure 1), some AAs may be detected via one analysis and not another, e.g., **AA 86** which was not identified in the initial SFVT results. There is consistency in the results from the analyses when all information is integrated.

We have fully automated all the analysis methods outlined in the pipeline in Figure 1, except for the Conditional Haplotype Method, which will be added in the future. At this stage, the different methods are run separately, with the researcher providing input when results from one method are required as input for another method. While many of these steps may be automated in the future, nonetheless it may not be feasible, or wise, to completely do away with oversight of some of these steps by the researcher.

Future studies of this JIA data set [12] will include SFVT analysis of: other common clinical subsets of JIA which show heterogeneity in their HLA DRB1 associated alleles; HLA class II DRB1-DQA1-DQB1 data (no effect of DQB1 is currently seen with allele and haplotype level analyses); and other classical HLA genes which via conditional haplotype method analysis have shown significant effects on disease risk, e.g., the class II DPB1 gene, as well as class I genes.

Acknowledgements

This research was supported by NIH contracts and grants: AI40076 (GT, NM, RMS, PAG, DRK, RHS), AI67068 (JAH, SJM, HAE), AI67150, AR42272, and AR47363 (SDT, DNG).

References

1. E. Thorsby, *Hum Immunol.* **53**, 1 (1997)
2. G. Thomson, A. M. Valdes, J. A. Noble JA, *et al.*, *Tissue Antigens* **70**, 110 (2007).
3. H. Salamon, J. Tarhio, K. Rønningen and G. Thomson, *J. Comp Biol.*.**3**, 497 (1996).
4. A. M. Valdes, S. McWeeney and G. Thomson, *Am. J. Hum. Genet.* **60**, 717 (1997)
5. J. B. Imboden, *Annu. Rev. Pathol. Mech. Dis.* **4**, 417 (2009).
6. D. R. Karp, N. Marthandan, S. G. E. Marsh, *et al.*, *Hum. Mol. Genet.* under revision (2009).
7. O. D. Solberg, S. J. Mack, A. K. Lancaster, *et al.*, *Hum. Immunol.* **69**, 443 (2008).
8. D. Meyer and G. Thomson, *Ann. Hum. Genet.* **65**, 1 (2001).
9. C. Leisner, N. Loeth, K. Lamberth, *et al.*, *PLoS One* **27**, e1678 (2008)
10. M. Nielsen, C. Lundegaard, P. Worning, *et al.*, *Bioinformatics* **20**, 1388 (2004)
11. N. Frahm, B. Baker, C. Brander, *In HIV Molecular Immunology200, Los Alamos, NM,* 3 (2008)
12. J. A. Hollenbach, S. D. Thompson, T. L. Bugawan, *et al.*, *Rheu. Arth.* under revision (2009)
13. H. Payami, S. Joe, N. R. Farid, *et al.*, *Am. J. Hum. Genet.* **45**, 541 (1989)
14. P. W. Hedrick, *Genetics* **117**, 331 (1987)
15. G. Thomson, L. F. Barcellos, A. M. Valdes. *Advances in Genetics* **60**, 255 (2008)