Human Immunology (2008) 69, 522-525

ASHI
AMERICAN SOCIETY FOR HISTOCOMPATIBILITY & IMMUNOGENETICS
www.ashi-hla.org

journal homepage: www.elsevier.com/locate/humimm

SKDM human leukocyte antigen (HLA) tool: A comprehensive HLA and disease associations analysis software

Stathis Kanterakis[a], Eleni Magira[a], Kenneth D. Rosenman[b], Milton Rossman[c], Keyur Talsania[a], Dimitri S. Monos[a,*]

[a] *Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, and Department of Pediatrics, University of Pennsylvania, School of Medicine, Philadelphia, PA, USA*
[b] *Department of Medicine, Michigan State University, East Lansing, MI, USA*
[c] *Department of Medicine, University of Pennsylvania, School of Medicine, Philadelphia, PA, USA*

Received 24 March 2008; received in revised form 23 May 2008; accepted 29 May 2008

KEYWORDS
HLA;
Immunogenetics;
Disease association;
MHC;
SKDM

Summary The immensely polymorphic and gene-rich landscape of the major histocompatibility complex on chromosome 6 necessitates a thorough and consistent investigation of its constituting elements. The human leukocyte antigens (HLAs) are an example of such polymorphic elements, implicated in many immune-based diseases. So far, analyses of HLA molecules in the context of diseases have been ad hoc, frequently incomplete, and extremely cumbersome. SKDM provides a comprehensive and automated workflow for detecting and dissecting HLA associations in diseases. We created a Java application to consistently perform our proposed method of analysis of HLAs in case–control datasets. The SKDM HLA tool can test for HLA allele differences between two populations and, by retrieving amino acid sequences, evaluates each polymorphic amino acid residue or a pocket of amino acids as an independent variant. Once primary associations are identified, the program examines zygosity and tests for strongest association, interaction, and linkage disequilibrium among amino acid epitopes of the same HLA molecule or between HLA isotypes. A summary of the analysis is output in plain language. The software and a user's manual are freely available at http://sourceforge.net/projects/skdm.

Introduction

The human leukocyte antigen (HLA) region of chromosome 6 contains some of the most polymorphic genes in the human genome. Studies on the immune response have focused on characterizing the sets of polymorphisms, or alleles, that are present in a population. Through the identification of disease-favored versus control-favored alleles, researchers have come to conclusions about the involvement of HLA molecules in the genetic susceptibility to immunological disorders [1–3].

* Corresponding author. Fax: 215-590-6361.
E-mail address: monosd@email.chop.edu (D.S. Monos).

0198-8859/$ -see front matter © 2008 American Society for Histocompatibility and Immunogenetics. Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.humimm.2008.05.011

ABBREVIATIONS

AA	amino acid
CBD	chronic beryllium disease
HLA	human leukocyte antigen
LD	linkage disequilibrium

The extreme polymorphism, intricate interrelations, high linkage disequilibrium (LD), and sharp recombination hotspots that characterize the HLA region [4,5] have prevented their systematic analysis, leaving researchers wondering whether there is more information to be mined from their data. Studies on HLAs initially focused on the differential distribution of alleles between a disease and control population. Eventually, investigators turned to the specific amino acid (AA) sequences of these alleles to assess whether a stronger association could be harbored from individual AA substitutions [2]. This type of thinking led to the identification of critical positions associated with diseases that go beyond a single residue. Multiple residues that form a "pocket" [6,7] and influence antigen binding and presentation on the HLA molecule were also associated with disease [3].

Researchers are commonly armed with allele frequency data, AA alignment information, and pocket positions for an array of HLA isotypes and disease populations. However, the HLA landscape is not getting any simpler and most analyses are still performed ad hoc. Also, with the identification of multiple associations with HLA loci, alleles, or residues, the need for tests for strongest association has been realized [8]. A recent report of a genome-wide single nucleotide polymorphism typing study in type 1 diabetes implicated the HLA class I, A, and B loci, in addition to the already known class II DR-DQ associations [9]. This underlines the necessity for thorough investigation of HLA components and their interactions in disease association studies.

Awareness of the above requirements led us to develop the SKDM HLA tool for the comprehensive and systematic analysis of case–control HLA typing data. To our knowledge, the only software similar in function is PyPop [10], which was developed for the analysis of data for the 13th and 14th International Histocompatibility Workshops [11]. However, its utility is geared toward population statistics and large sample sizes (1000–2000 individuals). PyPop yielded no output for datasets of smaller proportions. So far, SKDM is the only software specializing in case–control HLA analysis through the identification and subsequent dissection of AA associations. It is applicable to both large and small sample sizes ($\sim$50–300 individuals per group), a common feature of HLA typing data. Additionally, SKDM has a graphical interface, facilitates the user in visualizing the input data, uses straightforward statistics, and produces plain language output.

The program combines the unique feature of identifying AA associated with a particular disease with the elegant analysis proposed by Svejgaard and Ryder [8], evaluating their binary interactions and their comparative influence towards a disease phenotype.

Implementation

The goal of our proposed software solution is to provide abstraction from complex computation, thoroughly investigate HLA associations, and automate their analysis in the context of a case–control design. We acknowledge that most investigators are uncomfortable with working on a command line interface, with multiple input–output files, and on different operating systems. Therefore, we decided to develop our application in Java, a platform-independent language, and perform all calculations in memory, lending speed and power to our analysis. The simple interface of SKDM provides for an area to copy and paste the HLA typing for a case and a control population and, with a single command, seconds later, outputs the analysis in a readable format.

SKDM can analyze sample sets that have been molecularly typed—in high or low resolution—for the HLA class I and class II regions. The program accepts a set of HLA alleles for any of the A, B, C, DRB1, DQB1, DQA1, DPB1, and DPA1 HLA and MICA, MICB polymorphic loci for a list of individuals. The program performs a series of steps for the evaluation of case–control differences in HLA:

(1) The *allele* test, whereby the frequencies of individual alleles are evaluated for differential distribution;
(2) The *residue* test, where alleles are collectively evaluated for the differential distribution of their constituting polymorphic amino acids;
(3) A *pocket* test, where previously described HLA pockets are separately interrogated;
(4) The *zygosity* test, to investigate the significance of a homozygote or heterozygote condition. The variable that is evaluated is the AA(s) of a single locus, previously identified to be of differential frequencies in the compared populations;
(5) The *interactions* step, a series of tests for the strongest AA association, involving tests for independence, interaction, combined action, differential association, and LD.

For details as to how the zygosity and the interactions steps are calculated numerically, please see the User's Guide.

To assess the risk for disease given a particular HLA element, two-by-two contingency tables are produced. Odds ratios with Haldane's correction of Woolf's method [12,13] are used to reflect susceptibility to disease. Fisher's two-tailed exact test is used to calculate the significance of departure from unity. The *p* values obtained this way are multiplied by the number of tests performed by way of the Bonferroni correction.

In particular, the program first produces a summary of the frequency of each allele in the case and control populations and evaluates the difference (δ) of those frequencies. Although δ is not considered in the analysis (odds ratios and *p* values are used instead), they provide a good visual impression of which AAs are likely to be of significance in the population under study.

Using the collection of alleles in the input dataset, the program retrieves a list of aligned AA sequences from the IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla/align.html). Each polymorphic AA at each position in the alignment is interrogated for a differential distribution of fre-

Table 1 Excerpt of analysis with SKDM on a CBD population.

CBD disease					CBD controls		
HLA-DPB1 summary					HLA-DPB1 summary		
Allele	Pop fr		Allele fr		Allele	Pop fr	Allele fr
1001	8 (11.59%)		8 (5.80%)		1001	5 (1.74%)	5 (0.87%)
17 unique alleles total					25 unique alleles total		
69 samples total					288 samples total		
δ between for locue HLA-DPB1							
Allele	δ		*p* value	*p* corr	OR		
1001	9.85%		7.65E-4	0.0191	7.12		
HLA-DPB1 residues							
Pos	AA	Assoc	*p* value	*p* corr	OR		
9	H	+	8.27E-7	3.8E-5	5.8*		
55, 56	DE	+	1.85E-6	8.5E-5	5.2		
69	E	+	3.6E-12	1.65E-10	8.1**		
HLA-DPB1 pocket residues							
*in pocket 6 (pos: 9, 11, 28)					*p* corr = 2.15E-5		
**in pocket 4 (pos: 13, 69, 76, 68, 72, 24)					*p* corr = 9.35E-11		
Zygonity analysis							
DPB1_H-9: heterozygotes individually associated							
DPB1_DE-55, 56: homozygotes individually associated, heterozygotes individually associated							
DPB1_E-69: homozygotes individually associated, heterozygotes individually associated							
Interaction analysis							
DPB1_E69 independent of DPB1_DE55, 56; they have combined action; in LD (CASE); in LD (CTRL)							
DPB1_H9, DPB1_E69 have combined action; in LD (CASE); in LD (CTRL)							
DPB1_H9, DPB1_DE55, 56, have combined action; in LD (CASE); in LD (CTRL)							

From the above analysis, it can be deduced that the HLA-DPB1*1001 allele is associated with CBD. Further, glutamic acid at position 69 on the HLA-DPB1 molecule confers an even stronger risk to disease. Glu-69 is in LD with Asp and Glu at positions 55,56 and His at position 9, and their combined action is contributory to disease susceptibility. Glu-69. however, has the strongest association, being independently associated with DE-55, 56 and having the highest odds ratio.

quency between cases and controls. A list of significant AAs is thus produced.

Significant AAs are further examined for zygosity, that is, whether the homozygote and heterozygote conditions differentiate susceptibility to disease. The program also performs an inspection of the pairwise relationships, or interactions, between the significant AAs by making a number of stratifications: for each pair of AAs, *X* and *Y*, an assessment is made as to whether *X* is associated independently of *Y* and vice versa and whether they show interaction (four tests); whether their associations differ (Test 5), and a series of "less critical" tests: whether *X* and *Y* have combined action toward the disease (Test 6) and whether they are in LD in cases (Test 7) and controls (Test 8). These calculations are performed according to methods previously described [8]. Table 1 lists an exemplary analysis on a chronic beryllium disease (CBD) dataset.

For the zygosity and interactions tests, AA frequencies are calculated in the total of alleles (or single genes=2n) in the population, as opposed to a total of subjects (n). In this way, we interrogate the abundance of alleles/molecules that carry a particular AA(s). In the evaluation of zygosity, cases and controls are examined for the distribution of subjects carrying a particular AA in homozygosity or heterozygosity. It is, therefore, conceivable that two different alleles (two different molecules) that carry the same AA residue can be considered homozygous. In the evaluation of interactions, two AA previously identified to have significantly different distributions among cases and controls are examined for their combined or independent presence or absence a) within a single molecule if they belong to the same isotype or b) between molecules if they belong to different isotypes. This is accounted for in the output of the program, which specifies whether the comparisons are made within molecules or between molecules.

Using the analysis described, our program is capable of not only picking up the HLA alleles that are associated with a particular disease, but also identifying individual AA that might confer an even greater risk. The program is able to dissect the most strongly associated AA for each HLA molecule and examine interactions between different HLA isotypes.

Using Fisher's exact test, SKDM is particularly suited for small sample sizes. For such datasets a χ^2 test may be inadequate; effects of sampling and unbalanced contingency tables cause the test statistic not to be well approximated with the χ^2 distribution. Therefore, the two-tailed Fisher test with an appropriate Bonferroni correction is the most appropriate test statistic for our sample size and type of analysis.

Caution must be raised on the issue of correcting for multiple testing. The program suggests a correction coefficient for every test it performs to avoid erroneous claims of HLA associations. Researchers might choose to relax that correction given a priori knowledge about the data. To that end, raw *p* values are always provided and it is up to the investigator to accept the corrected *p* values and apply a different or no correction at all.

The SKDM software was tested for validation against two previously published studies [14,15]. The complete analysis for one of them [14] has been included as an electronic supplement to the current paper. This exercise revealed a few interesting points: (1) in several instances, minor discrepancies were noted in the total count of individuals who were counted using the computational and the manual methods. The differences were small (1–3) and did not affect the *p* values significantly, but nevertheless the manual analysis is prone to counting errors. This underscores the need for a software analysis program, because it is likely that mistakes will occur, particularly when the number of subjects are high (over 100) and the analysis is performed at the AA level where the complexity of the counts increases dramatically. (2) Table 6 in [14] indicates *p* values that are different from the *p* values presented in the output of the program (see supplement for comparisons between residues DRB1 E71 and DQA1 Y11). This is not a discrepancy. The *p* values of the program output have been multiplied by 5 as a correction factor because the analysis does not ask a specific question to be addressed by a particular comparison; it simply performs all possible comparisons. The *p* values of Table 6, however, have not been multiplied by any factor because we specifically performed the particular comparisons depicted in Table 6 when the analysis was performed manually; the logic of our analysis did not require doing all the others. We should be aware that the software automatically performs this correction and when the analysis is targeted to a particular question there is no need for unnecessary correction factors. (3) The significance of the trio of residues at DQβRPD55-57 (reference 15, Table IV) was missed by SKDM, even though the program evaluated each of these residues separately and in binary combinations amongst themselves. The program has been limited to evaluating combinations of two AA and is therefore likely to miss associations of combinations of three or more residues. An exception is the case where AA follow the exact same distribution among alleles (that is, they are in absolute LD), in which case they are grouped together and treated as a single entity. Limitations of memory, processing power and execution time considerations lead to the restriction of binary interactions. This allows the program to run sufficiently fast on any standalone computer. However, a sliding window approach, for testing the significance of neighboring, variable groups of AA will be considered in future upgrades. This limitation should prompt investigators to also examine the allele alignments visually in some cases. Our program enhances this task by sorting the AA alignments for each allele according to its differential distribution between cases and controls, with the most frequent alleles in each of these groups in opposite poles. (4) The computational analysis provides a significant body of information that is virtually impossible to obtain manually and forms the basis for identifying unsuspected relationships that, even if not significant in one study, can prove to be relevant with additional studies from different populations or larger sample sizes. In short, the program allows for a thorough evaluation of the data that is currently unavailable elsewhere.

Acknowledgments

Funding for this project was provided by the National Institute for Occupational Safety and Health Grant R01 OH007495. We also thank the Children's Hospital of Philadelphia for institutional support.

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.humimm.2008.05.011.

References

[1] Morel PA, Dorman JS, Todd JA, McDevitt HO, Trucco M. Aspartic acid at position 57 of the HLA-DQ beta chain protects against type I diabetes: a family study. Proc Natl Acad Sci USA 1988;85:8111-5.

[2] Todd JA, Bell JI, McDevitt HO. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature 1987;329:599-604.

[3] Zerva L, Cizman B, Mehra NK, Alahari SK, Murali R, Zmijewski CM, et al. Arginine at positions 13 or 70–71 in pocket 4 of HLA-DRB1 alleles is associated with susceptibility to tuberculoid leprosy. J Exp Med 1996;183:829-36.

[4] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science 2004;304:581-4.

[5] Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, et al. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 2005;76:634-46.

[6] Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, et al. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 1993;364:33-9.

[7] Stern LJ, Brown JH, Jardetzky TS, Gorga JC, Urban RG, Strominger JL, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 1994;368:215-21.

[8] Svejgaard A, Ryder LP. HLA and disease associations: detecting the strongest association. Tissue Antigens 1994;43:18-27.

[9] Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature 2007;450:887-92.

[10] Lancaster A, Nelson MP, Meyer D, Thomson G, Single RM. PyPop: a software framework for population genomics: analyzing large-scale multi-locus genotype data. Pac Symp Biocomput 2003;8:514-25.

[11] Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update—a software pipeline for large-scale multilocus population genomics. Tissue Antigens 2007;69(Suppl 1):192-7.

[12] Haldane JB. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet 1956;20:309-11.

[13] Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet 1955;19:251-3.

[14] Monos DS, Pappas J, Magira EE, Gaughan J, Sakkas L, Aplenc R, et al. Identification of HLA-DQα and -DRβ residues associated with susceptibility and protection to epithelial ovarian cancer. Hum Immunol 2005;66:554-62.

[15] Magira EE, Papaioakim M, Nachamkin I, Griffin JW, McKhann GM, Monos DS, et al. Differential distribution of HLA-DQβ/DRβ epitopes in the two forms of Guillain-Barré syndrome, acute motor axonal neuropathy and acute inflammatory demyelinating polyneuropathy (AIDP): identification of DQβ epitopes associated with susceptibility to and protection from AIDP. J Immunol 2003;170:3074-80.