Human Immunology (2008) 69, 522-525

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 (~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-