BMC Bioinformatics

Software

Open Access

Predicting population coverage of T-cell epitope-based diagnostics and vaccines

Huynh-Hoa Bui[1], John Sidney[1], Kenny Dinh[1], Scott Southwood[2], Mark J Newman[2] and Alessandro Sette*[1]

Address: [1]La Jolla Institute for Allergy and Immunology, Division of Vaccine Discovery, 3030 Bunker Hill Street, Suite 326, San Diego, CA 92109, USA and [2]IDM Inc., 5820 Nancy Ridge Drive, Suite 100, San Diego, CA 92121, USA

Email: Huynh-Hoa Bui - hbui@liai.org; John Sidney - jsidney@liai.org; Kenny Dinh - kdinh@liai.org; Scott Southwood - ssouthwood@idm-biotech.com; Mark J Newman - mnewman@idm-biotech.com; Alessandro Sette* - alex@liai.org

* Corresponding author

Published: 17 March 2006

BMC Bioinformatics 2006, **7**:153 doi:10.1186/1471-2105-7-153

Received: 14 October 2005
Accepted: 17 March 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/153

Abstract

Background: T cells recognize a complex between a specific major histocompatibility complex (MHC) molecule and a particular pathogen-derived epitope. A given epitope will elicit a response only in individuals that express an MHC molecule capable of binding that particular epitope. MHC molecules are extremely polymorphic and over a thousand different human MHC (HLA) alleles are known. A disproportionate amount of MHC polymorphism occurs in positions constituting the peptide-binding region, and as a result, MHC molecules exhibit a widely varying binding specificity. In the design of peptide-based vaccines and diagnostics, the issue of population coverage in relation to MHC polymorphism is further complicated by the fact that different HLA types are expressed at dramatically different frequencies in different ethnicities. Thus, without careful consideration, a vaccine or diagnostic with ethnically biased population coverage could result.

Results: To address this issue, an algorithm was developed to calculate, on the basis of HLA genotypic frequencies, the fraction of individuals expected to respond to a given epitope set, diagnostic or vaccine. The population coverage estimates are based on MHC binding and/or T cell restriction data, although the tool can be utilized in a more general fashion. The algorithm was implemented as a web-application available at http://epitope.liai.org:8080/tools/population.

Conclusion: We have developed a web-based tool to predict population coverage of T-cell epitope-based diagnostics and vaccines based on MHC binding and/or T cell restriction data. Accordingly, epitope-based vaccines or diagnostics can be designed to maximize population coverage, while minimizing complexity (that is, the number of different epitopes included in the diagnostic or vaccine), and also minimizing the variability of coverage obtained or projected in different ethnic groups.

Background

T lymphocytes recognize a complex between a specific major histocompatibility complex (MHC) molecule and a particular pathogen-derived epitope. Thus, a given epitope will elicit a response only in individuals that express an MHC molecule capable of binding that particular epitope, explaining to a large extent the phenomenon known as "MHC restriction" [1]. In humans, MHC molecules are known as human leukocyte antigen (HLA) molecules and two different types exist: class I and class II. HLA class I molecules mostly bind peptides derived from the endogenous processing pathway, and their recognition is primarily associated with cytotoxic T lymphocytes (CTL), which are most important for antiviral and anticancer immunity responses. By contrast, HLA class II molecules bind peptides typically derived from the extracellular milieu, and they are important for helper T lymphocyte (HTL) responses, which regulate antibody and cytotoxic responses.

HLA molecules are extremely polymorphic. Over a thousand different HLA allelic variants have been defined to date [2]. Specific HLA alleles are expressed at dramatically different frequencies in different ethnicities [3,4]. Therefore, in the design and development of T-cell epitope-based diagnostics or vaccines, selecting multiple epitopes with different HLA binding specificities will afford increased coverage of the patient population. A pertinent goal, in this context, might be to identify optimal sets of HLA alleles with maximal coverages for different populations [5,6]. Extensive analyses by Longmate and coworkers [7] suggested that 90% population coverage of several ethnic groups can be achieved by targeting eleven different HLA molecules. However, 90% coverage of African and Asian ethnicities required four or more additional molecules. Dawson et al. also analyzed the problem [8] and concluded that to reach 80% coverage, 3 to 5 HLA molecules were required in a given ethnicity, but the actual HLA specificities required were different in different ethnic groups.

An important consideration in the process of epitope selection for a T-cell epitope-based diagnostic or vaccine is that the patient population coverage afforded by a given epitope set does not simply correspond to the sum of the coverage of the individual components. To calculate the coverage afforded by a given set of epitopes with multiple and/or overlapped HLA binding specificities, a more comprehensive approach, taking into account MHC binding and T cell recognition patterns, is required for this purpose. A suitable algorithm was previously utilized [9-11] but not described in detail. This method calculates the fraction of individuals predicted to respond to a given epitope or epitope set on the basis of HLA genotypic frequencies and on the basis of MHC binding and/or T cell restriction data. In this paper, we describe the algorithm and its implementation as a web application available to the public. We believe this is a useful tool to aid in the design and development of T-cell epitope-based diagnostics and vaccines intended to be effective across diverse populations.

Implementation

For a given HLA gene locus, let $\{m_1, m_2, ..., m_N\}$ denote a set of MHC alleles, with each allele associated with a genotypic frequency $G(m_i)$ for a population or ethnic group. To account for 100% of alleles of a given locus, the total genotypic frequency ($\sum G(m_i)$) should add up to 1. If $\sum G(m_i)$ is less than 1, an unidentified HLA allele with a genotypic frequency equal to the residual (1 - $\sum G(m_i)$) is added to the locus. If $\sum G(m_i)$ is greater than 1, the genotypic frequency of each m_i allele of the locus is scaled down proportionately by dividing the frequency by $\sum G(m_i)$. Next, let $\{e_1, e_2, ..., e_K\}$ denote a set of epitopes with known MHC binding or restriction data. For each epitope e_k, its restriction to an MHC allele m_i, $e_k(m_i)$, is defined as followed:

$$e_k(m_i) = \begin{cases} 0 & \text{if } e_k \text{ is not restricted to } m_i \\ 1 & \text{if } e_k \text{ is restricted to } m_i \end{cases} \qquad (1).$$

First, for each MHC allele (m_i), a total number of epitope "hits", $H(m_i)$, was tabulated by adding the number of epitopes that are restricted to (or bound by) m_i:

$$H(m_i) = \sum_{k=1}^{K} e_k(m_i) \ \ (i = 1, \cdots, N) \qquad (2).$$

Next, for each possible diploid MHC combination (m_i, m_j), a phenotypic frequency $F(m_i, m_j)$ was calculated based on individual allele genotypic frequency:

$$F(m_i, m_j) = G(m_i) \times G(m_j) \qquad (3)$$

For n MHC types, this corresponds to an $n \times n$ tabulation of the phenotypic frequency at which each specific pair of MHCs will be found in the population from which the MHC frequencies were derived. A similar table was also generated to contain the number of epitope hits per each of the MHC combinations $H(m_i, m_j)$. In the case of heterozygous combinations, $H(m_i, m_j)$ was calculated as the sum of the number of epitope hits associated with each of the two alleles, $H(m_i) + H(m_j)$. This is because m_i and m_j are two different alleles, and therefore the number of epitope hits recognized by each allele in the combination is independent of each other. However, in the case of homozygous combinations which contain two identical alleles, the number of epitope hits was the same as the number of epitope hits of the given allele:

$$H(m_i, m_j) = \begin{cases} H(m_i) + H(m_j) & \text{if } i \neq j \\ H(m_i) & \text{if } i = j \end{cases} \qquad (4).$$

Based on the calculated $F(m_i, m_j)$ and $H(m_i, m_j)$ tables, a frequency distribution was assembled by tabulating the phenotypic frequencies of all MHC combinations associated with a certain number of epitope/HLA combination hits (h):

$$F(h)=\sum_{i=1}^{N}\sum_{j=1}^{N}F(m_i,m_j)I_{\{H(m_i,m_j)=h\}} \qquad (5),$$

where $I_{\{H(m_i,m_j)=h\}}=\begin{cases}1 & \text{if } H(m_i,m_j)=h\\ 0 & \text{if } H(m_i,m_j)\neq h\end{cases}$ is an indicator function.

For calculation of coverage by epitope sets restricted to MHC alleles of multiple k different loci, a combined frequency distribution (P) as a function of epitope/HLA combination hits (n) was generated by merging k separate frequency distributions. This merging procedure is based on the assumption that linkages between MHC loci are in equilibrium, and was done as follows:

$$P(n)=\sum_{h_1\geq 1}\cdots\sum_{h_k\geq 1}\left(\prod_{i=1}^{k}F_i(h_i)I_{\left\{\sum_{i=1}^{k}h_i=n\right\}}\right) \qquad (6),$$

where $I_{\left\{\sum_{i=1}^{k}h_i=n\right\}}=\begin{cases}1 & \text{if } \sum_{i=1}^{k}h_i=n\\ 0 & \text{if } \sum_{i=1}^{k}h_i\neq n\end{cases}$ is an indicator function, and $F_i(h_i)$ is a phenotypic frequency associated with h_i epitope/HLA combination hits of locus i calculated from equation 5.

The population coverage (C) or fraction of individuals projected to respond to the epitope set was then calculated as the sum of the combined phenotypic frequencies associated with at least one epitope hit/HLA combination:

$$C=\sum_{n\geq 1}P(n) \qquad (7).$$

Based on equation 6, a histogram was generated to summarize the fraction of population coverage (P) as a function of the number of HLA/epitope combinations (n) recognized. A cumulative population coverage distribution frequency (Y) as a function of the number of HLA/epitope combinations (n) was also calculated:

$$Y(n)=\sum_{x\geq n}P(x) \qquad (8).$$

From this cumulative population coverage distribution of the whole epitope set, $PC90$, defined as the minimum number of epitope/HLA combination hits (n) recognized by 90% of the population, was determined as follow:

$$PC90=n+\frac{Y(n)-0.9}{Y(n)-Y(n+1)} \qquad (9),$$

where $Y(n) \geq 0.9 > Y(n+1)$. Because) $PC90$ was determined by data interpolation, it can be of any positive decimal value. Based on equation 9, if the population coverage is less than 90% or $C=\sum_{n\geq 1}P(n)\equiv Y(1)<0.9$, $PC90$ will be less than 1.

Additionally, the average number of epitope/HLA combination hits (A) recognized by the population is a weighted average and was calculated as follow:

$$A=\sum_{n\geq 1}n\times P(n) \qquad (10).$$

Results and discussions

The Population Coverage Calculation program was implemented as a Java servlet public web-application (see Availability and Requirements section). HLA allele (genotypic) frequencies were obtained from dbMHC database [12]. At present, dbMHC database provides allele frequencies for 78 populations grouped into 11 different geographical areas. In addition to the allele frequencies obtained from the dbMHC database, the Population Coverage Calculation program also accepts custom populations with allele frequencies defined by users. Multiple population coverages can be simultaneously calculated and an average population coverage is generated. Since MHC class I and MHC class II restricted T cell epitopes elicit immune responses from two different T cell populations (CTL and HTL, respectively), the program provides three calculation options to accommodate different coverage modes – (1) class I separate, (2) class II separate, and (3) class I and class II combined. For each population coverage, a histogram is generated to summarize the percentage distribution of individuals as a function of the number of epitope/HLA combinations recognized. A cumulative coverage distribution plot is also generated to determine the minimum number of epitope/HLA combinations recognized by 90% of the population (PC90). Finally, the average number of epitope/HLA combinations recognized by the population and coverages of individual epitope are also calculated.

It should be noted that when population coverages are projected from an epitope set restricted to alleles from multiple HLA loci, linkages between loci are taken into account. The overall population (phenotypic frequency),

(P_{total}), is mathematically derived as the sum of the individual locus' coverage corrected for the overlaps: $S = \sum_{k=1}^{n} \frac{n!}{k!(n-k)!}$, where P_{ij} is the frequency of the *ij* haplotype, P_{ijk} is the frequency of the *ijk* haplotype, etc... If gene linkage equilibrium is assumed, P_{ij} can be calculated as the product of the individual allele phenotypic frequencies ($P_i \times P_j$), and $P_{ijk} = P_i \times P_j \times P_k$, etc... This calculation is implicitly incorporated in our current algorithm (equation 6). However, if gene linkage is in disequilibrium, the frequency of a given haplotype is usually not equal to the product of their individual allele phenotypic frequencies, ($P_{ij} \neq P_i \times P_j$, $P_{ijk} \neq P_i \times P_j \times P_k$, ...). As a result, to account for linkage disequilibrium between HLA loci, complete data on haplotype frequencies must be known. Therefore, it would be difficult to factor in linkage disequilibrium at this time because linkage disequilibrium is known to be different in different ethnicities, and data regarding the specific disequilibrium in different ethnicities in general is not available or incomplete. As more comprehensive MHC linkage disequilibrium data becomes available, our method can be modified to incorporate this type of calculation.

Although the present program assumes linkage equilibrium between HLA loci, the impact of linkage disequilibrium, which is known to occur in the MHC region, on the calculated coverage is expected, in most contexts, to be minimal. For example, in the North American Caucasian population, the A1 and B8 antigens of HLA-A and -B loci, respectively, are known to be the strongest linked antigen pair with an observed haplotype frequency of 7.95% [13]. The genotypic frequencies of the A1 and B8 antigens are 15.18% and 9.41%, respectively [13]. Assuming the linkage between A1 and B8 antigens is in equilibrium, the overall population coverage calculated by the present program is 40.97%, and the individual population coverages by A1 and B8 antigens are 28.06% and 17.93%, respectively. The expected equilibrium frequency for the A1/B8 haplotype, in this case, is 5.03% (28.06% × 17.93%) which is 2.92% less than the observed frequency of 7.95%. Therefore, if linkage disequilibrium is considered, the overall population coverage will be 38.04% (28.06% + 17.93% - 7.95%). Thus, even for the most tightly linked A1/B8 haplotype in the Caucasian population, linkage disequilibrium, in this specific example, only accounted for less than 3% difference in the population coverage calculated by the present program. Furthermore, we have also investigated the deviations between the observed and expected equilibrium frequencies of 1012 HLA-A/-B haplotypes in the North American Caucasian population, based on available antigen- and haplotype-frequencies published by Mori *et al.* [14,15]. On average, the observed haplotype frequencies deviated from the expected equilibrium frequencies by approximately 0.58%. As a result, linkage disequilibrium is expected to impact the calculated population coverage, but the degree of the impact is expected to be negligible.

It should be pointed out that the calculations described herein can also be performed on data spreadsheets, but the process is laborious, error prone and also requires extensive immunological expertise. In our experience, a single calculation without the aid of this tool requires several hours to complete. To the best of our knowledge, at this time, there is no existing program that is publicly accessible as a web-resource that can offer the flexibility and range of utility similar to the Population Coverage Calculation program that we have developed. The present application represents a significant enhancement of the dbMHC database's utility by incorporating its compiled data of world-wide ethnic population frequencies to calculate HLA coverage for user-defined population subsets. The program is flexible by allowing the user to specify groups of related or unrelated ethnicities as well as specify the HLA alleles under consideration. Additional flexibility features include the implementation of separate calculations for both MHC Class I and Class II demarcated recognitions as they involve immune responses from two different populations of T cells – CTL and HTL, respectively. The output of the program was also specifically designed to be accessible to both specialists and neophytes in the field of MHC research. Therefore, having this tool publicly available is highly desirable. Additionally, in our future works, we plan to incorporate in the tool the ability to search for minimal epitope subset(s) within the given epitope set that will afford a specified population coverage level. This is not a trivial task due to a large number of possible epitope subsets (S) that has to be considered, $S = \sum_{k=1}^{n} \frac{n!}{k!(n-k)!}$ where n is the total number of epitopes and k is the number of epitopes in a subset. For example, for a set of 20 epitopes, there will be a total of 1,048,575 combinations of epitope subsets that needs to be evaluated. Therefore, a strategic searching approach must be devised to computationally accomplish this task. In summary, with the help of this Population Coverage Calculation program, epitope-based vaccines or diagnostics can be designed to maximize population coverage while minimizing complexity (that is, the number of different epitopes included in the diagnostic or vaccine), and also minimizing the variability of coverage obtained or projected in different ethnic groups.

Conclusion

Herein, we have implemented a method to calculate projected population coverage of a T-cell epitope-based diagnostic or vaccine using MHC binding or T cell restriction data and HLA gene frequencies. The Population Coverage Calculation program was designed to be user friendly and flexible. Besides the compiled HLA gene frequencies currently provided, users can also supply their own tabulated HLA gene frequencies for calculation. Therefore, researchers can use this tool to perform coverage analyses on their specific patient populations. We plan to continuously update the compiled HLA gene frequencies as more data are available, and thus to provide researchers with a useful tool to aid in the design and development of effective T-cell epitope-based diagnostics and vaccines.

Availability and requirements

Project name: Population Coverage Calculation

Project home page: http://epitope.liai.org:8080/tools/population

Programming language: Java

Operating system: Fedora Linux

Other requirements: Apache Tomcat 5.5.12, MySQL 4.1

Web browser: Population Coverage Calculation program has been tested and shown to work with the following browsers: Firefox version 1.5 (PC and Mac OS X), Netscape version 8.0.4 (PC), Netscape version 7.2 (Mac OS X), Internet Explorer version 6.0 (PC), Internet Explorer version 5.2 for Mac (Mac OS X). Default security settings were used.

Authors' contributions

HHB developed the computer algorithm and designed the web-resource. AS and JS contributed the calculation approaches. KD helped with programming and collecting HLA frequency data. SS and MN were involved in conceptualizing the calculation approaches. HHB wrote the manuscript, AS and JS edited the final version. All authors read and approved the manuscript.

Acknowledgements

The authors thank Howard Grey for helpful review and suggestions. This work was supported by the National Institutes of Health's contract HHSN26620040006C (Immune Epitope Database and Analysis Program).

References

1. Zinkernagel RM, Doherty PC: **Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system.** *Nature* 1974, **248(450):**701-702.
2. **Anthony Nolan Research Institute – HLA Informatics Group** [http://www.anthonynolan.org.uk/HIG/index.html]
3. Imanishi T, Akaza T, Kimura A, Tokunaga K, Gojoubori T: **Allele and haplotype frequencies for HLA and complement loci in various ethnic groups.** In *HLA 1991: Proceedings of the Eleventh International Histocompatibility Workshop and Conference* Edited by: Tsuji K MA, Sasazuki T. Oxford: Oxford University Press; 1992:1065-1220.
4. Gjertson DW, Lee S-H: **HLA-A/B and -DRB1/DQB1 allele-level haplotype frequencies.** In *HLA 1998* Edited by: Terasaki PI. Lenexa, KS, USA: American Society for Histocompatibility and Immunogenetics; 1998:365-450.
5. Schipper RF, van Els CA, D'Amaro J, Oudshoorn M: **Minimal phenotype panels. A method for achieving maximum population coverage with a minimum of HLA antigens.** *Hum Immunol* 1996, **51(2):**95-98.
6. Gulukota K, DeLisi C: **HLA allele selection for designing peptide vaccines.** *Genet Anal* 1996, **13(3):**81-86.
7. Longmate J, York J, La Rosa C, Krishnan R, Zhang M, Senitzer D, Diamond DJ: **Population coverage by HLA class-I restricted cytotoxic T-lymphocyte epitopes.** *Immunogenetics* 2001, **52(3–4):**165-173.
8. Dawson DV, Ozgur M, Sari K, Ghanayem M, Kostyu DD: **Ramifications of HLA class I polymorphism and population genetics for vaccine development.** *Genet Epidemiol* 2001, **20(1):**87-106.
9. Doolan DL, Southwood S, Chesnut R, Appella E, Gomez E, Richards A, Higashimoto YI, Maewal A, Sidney J, Gramzinski RA, *et al.*: **HLA-DR-promiscuous T cell epitopes from Plasmodium falciparum pre-erythrocytic-stage antigens restricted by multiple HLA class II alleles.** *J Immunol* 2000, **165(2):**1123-1137.
10. Wilson CC, Palmer B, Southwood S, Sidney J, Higashimoto Y, Appella E, Chesnut R, Sette A, Livingston BD: **Identification and antigenicity of broadly cross-reactive and conserved human immunodeficiency virus type 1-derived helper T-lymphocyte epitopes.** *J Virol* 2001, **75(9):**4195-4207.
11. Sette A, Chesnut R, Livingston B, Wilson C, Newman M: **HLA-binding peptides as a therapeutic approach for chronic HIV infection.** *I Drugs.* 2000, **3(6):**643-8.
12. **dbMHC** [http://www.ncbi.nlm.nih.gov/mhc/]
13. Mori M, Beatty PG, Graves M, Boucher KM, Milford EL: **HLA gene and haplotype frequencies in the North American population: the National Marrow Donor Program Donor Registry.** *Transplantation* 1997, **64(7):**1017-1027.
14. **Estimated gene frequencies of HLA-A antigens** [http://www.ashi-hla.org/publicationfiles/archives/prepr/mori_gf.htm]
15. **HLA-A,B Haplotype Frequencies** [http://www.ashi-hla.org/publicationfiles/archives/prepr/mori_ab.htm]