

The Curation Guidelines of the Immune Epitope Database and Analysis Resource

R. Vita,* B. Peters, A. Sette

La Jolla Institute for Allergy and Immunology, Division of Vaccine Discovery, 9420 Athena Circle, La Jolla, California 92037

Received 7 March 2008; Accepted 22 April 2008

Grant sponsor: National Institutes Health; Grant number: HHS266200400006C.

*Correspondence to: Randi Vita, La Jolla Institute for Allergy and Immunology, Division of Vaccine Discovery, 9420 Athena Circle, La Jolla, CA 92037, USA

Email: rvita@liai.org

Published online 7 August 2008 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20585

© 2008 International Society for Advancement of Cytometry

• Abstract

The IEDB houses antibody and T cell epitope data and makes them accessible and searchable. The curation of literature references requires explicit guidelines in order to capture the data in an objective and consistent manner. Description of these guidelines ensures transparency of the database and facilitates direct submissions to the database. © 2008 International Society for Advancement of Cytometry

• Key terms

immunology; epitope; database; antibody; MHC

THE goal of The Immune Epitope Database and Analysis Resource (IEDB) is to catalog epitope data and make them freely accessible to the scientific community. Epitope-related data is relevant across a number of research interests and is utilized for improved diagnostics as well as vaccines and therapeutics. At a recent MASIR conference focusing on antigenicity, the study and use of immune epitopes was highlighted. The utility of enhanced detection of epitope specific immune responses (1–3) and significant efforts to improve epitope mapping (4,5) in basic research and clinical settings were presented.

The IEDB houses antibody and T cell epitopes derived from infectious agents, allergens, and autoantigens was recognized in a diverse range of hosts. Available online since January 2005, the IEDB data is derived from over 4,000 literature references and imported from previously developed databases. Additionally, direct submissions from laboratories provides an avenue for research groups to widely disseminate, in parallel with publication in peer-reviewed journals, epitope-related data.

Previous reports have described various aspects of the project, including the blueprint of the database (6), its high-level data structure (7), and the processes put in place for the curation of the complex data within the scope of the database (8), and are not therefore discussed in detail here. Instead, we describe the actual guidelines utilized to curate epitope-related data.

Curation relies upon a team of doctoral level curators, a team of immunological experts who provide quality control and immunological insight, and the Curation Manual (8). The Curation Manual exists as a living document format (http://tools-int-01.liai.org/wiki/index.php/Main_Page) and is designed to assist in accurate and consistent data curation, while allowing continued growth and adaptation. These guidelines, derived from experience, allow us to capture immunological epitope data in a searchable and standardized format. Their description ensures transparency of the IEDB to the scientific community and facilitates direct submissions from laboratories involved in epitope-related research.

THE STRUCTURE OF THE IEDB

Relevant epitope data can be dispersed throughout the text and figures of an article or in laboratory notebooks and reports. This data must be curated or

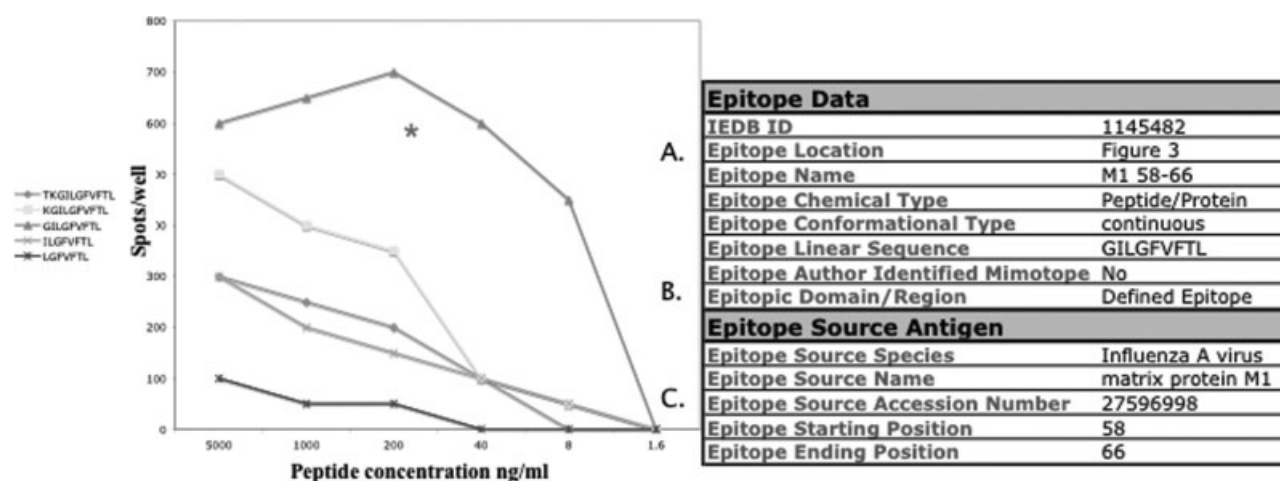


Figure 3. PBMC from healthy donors were incubated with the indicated concentrations of the matrix peptides and IFN production was measured by ELISPOT assay as described in Materials and Methods.

Figure 1. An example of how an epitope is curated from data presented in a reference. **A.** Based upon the results of the ELISPOT assay (in Figure 3 of the reference), the optimal peptide, GILGFVFTL, is curated as the epitope. **B.** Additionally, this sequence is described as a defined epitope due to its size ($7 < X < 11$). **C.** The source of this epitope is captured as matrix protein M1 from the influenza A virus based upon the information provided in the reference. The Source Accession Number represents an entry in GenBank to which this epitope's sequence, source protein and source species match.

“translated” into tables of information, corresponding to a relational database. To this end, curation criteria must be defined.

The general structure of the IEDB is comprised of three main concepts: the reference source, the epitope structure, and the assay context. A first group of fields describes the reference source such as its pubmed Id, article title, authors, and so on. Next, the sequence or molecular structure of the epitope is described. If the same epitope is described in multiple references, separate records capture the data from each reference source. Each epitope is linked to at least one assay context, which describes how the epitope was experimentally tested. If the same epitope was tested in multiple assays, separate records are created for each assay.

INCLUSION CRITERIA

To be included in the database, a reference must report original and experimental epitope-related data. Data related to MHC binding, elution from MHC, T cell responses (including NK T cells), and B cell/antibody responses are defined as relevant. Conversely, experimental data that involves immunological interactions unrelated to adaptive immunity, such as those of natural killer cells (non-T cell) or data related to MHC-super antigen, are excluded. Computer-derived predictions, sequence analysis without immunological data, reviews, and meta-analysis are also excluded.

EPITOPE STRUCTURE

For peptide and nucleotide epitopes, the structure is entered as standard one-letter codes and is designated as either linear (continuous) or conformational (discontinuous).

SMILES notations are used for all other epitope chemical types. When conformational epitopes are defined in a reference by only certain key residues, these residues are entered in the conformational sequence field.

In the event that the exact epitope structure is not provided in the reference, cited references are researched or the corresponding author(s) are contacted to obtain this information. Unfortunately, the lack of detailed structural information, such as the exact linear sequence, leads to exclusion of a great deal of data.

The resolution to which epitope structures are determined is highly variable. Accordingly, certain fields and criteria depict and classify the various degrees of epitope resolution. The database currently only includes epitopes of less than 50 residues or less than 5,000 Daltons. When the epitope is accurately mapped to its minimal or optimal size, the epitope structure is labeled as a “defined epitope.” When it is not clear if the identified epitope structure is the minimal/optimal epitope, B-cell and MHC Class I restricted epitopes are labeled as defined epitopes if they are less than 11 amino acids in size. Figure 1 depicts a hypothetical figure as it might appear in a manuscript. The IEDB field “Epitope Location” describes the location of the data within the curated manuscript (Figure 3) and the field “Epitopic Domain/Region” is used to define these class I restricted epitopes as “defined epitopes” due to their size of 11 residues or less [Fig. 1B]. MHC Class II restricted epitopes are deemed defined if they are seven to 15 residues in size. When the epitope is not mapped to its minimal or optimal size, the epitope structure is labeled as an “epitope containing region.” Finally, if only a portion of the epitope is mapped, as when specific residues are determined to be crucial components of the epitope, these are labeled as “residues involved in recognition.”

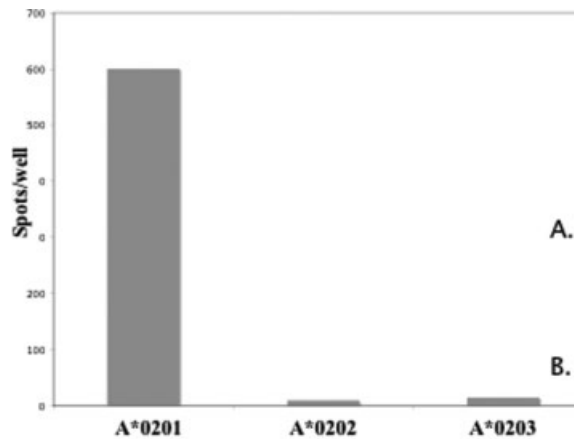


Figure 4. BLCL expressing select class I alleles were used in the ELISPOT assay as described in Materials and Methods.

Figure 2. A screen shot of a curated context as it appears in the IEDB. **A.** The ELISPOT assay in Figure 3 is curated under the optimal epitope. The other peptides that were studied are commented upon. **B.** Restriction studies, as shown in Figure 4 of the reference, are not curated as distinct contexts, but rather the information presented by them is captured in all relevant assays. **C.** The immunized species, which is the source organism of the effector cells, is captured along with the immunogen. In this case, natural exposure to the influenza virus is presumed for all humans.

A.	T Cell Assay	
	Assay ID	1404556
	T Cell Location of Data	Figures 3 and 4
	T Cell Qualitative Measurement	Positive
B.	T Cell Assay Comments	
	Truncated peptides were used to demonstrate this epitope as the optimal sequence for IFN γ production.	
	Assay Type	
	T Cell Assay Type	ELISPOT-Cytokine Release-IFN- γ
C.	Effector Cell	
	Effector Cells Tissue Type	Blood
	Effector Cells Cell Type	PBMC
	Effector Cells Origin	Directly assayed (in vitro)
	MHC Data	
	MHC Allele	HLA-A*0201
	MHC Evidence Code	
	T cell assay -Mismatched MHC molecules	
	Antigen Presenting Cell	
	Autologous or Syngeneic	Yes
	Antigen Presenting Cells Tissue Type	Blood
	Antigen Presenting Cells Cell Type	PBMC
	Antigen Presenting Cells Origin	
	Directly assayed (in vitro)	
	Antigen	
	Antigen Type	Epitope
	Antigen Name	M1 58-66
	Antigen Chemical Type	Peptide/Protein
	Antigen Source Species	
	Influenza A virus	
	Antigen Linear Sequence	
	GILGFVFTL	
	Antigen Source Name	
	matrix protein M1	
	Immunized Species	
	Immunization Category	Natural Exposure/Occurrence
	Immunized Species	Homo sapiens
	Immunogen	
	Immunogen Type	Source Species
	Immunogen Name	Influenza A virus
	Immunogen Chemical Type	Other
	Immunogen Source Species	
	Influenza A virus	

SPECIAL ISSUES

Epitope analogs, naturally occurring variants, modified epitopes, mimotopes, and cross-reactive epitopes all require special considerations. Analogs are synthetic peptides or chemical compounds that share some structural features with the epitope. They are usually not captured as separate epitope entries, but entered as an additional context of the wild type epitope. Exceptions to these rules are when analogs are studied without reference to the wild type structure, such as the use of analogs in MHC binding assays or in immunogenicity/antigenicity contexts where a wild type sequence is not present in either the antigen or the immunogen. Viral escape mutations are conceptually similar to nature-generated analogs. Typically, T cell recognition or antibody binding to the wild type is demonstrated along with a loss of recognition of the escape mutants. In these cases, the relevant residues of the wild type sequence are curated as residues involved in recognition. Mimotopes are defined within the IEDB as functional mimics of natural molecular structures that bear little or no sequence homology to their biological counterparts and are captured as separate epitope records. Cross-reactivity occurs when the epitope contained within the immunogen and the antigen are different. Should the immunogen or antigen be designated as the epitope? If both the immunogen and the antigen are natural or both are artificial, they are both captured as separate epi-

topes entries, thus allowing both to be treated with equal priority in the database. If one of them is artificial, the natural structure is designated as the epitope.

MINIMAL AND OPTIMAL EPITOPES

When a number of overlapping or truncated peptides and epitope analogs are tested for the purpose of defining the epitope structure, only the minimal or optimal structure is entered in the database. The optimal structure is defined as the one giving the highest response. If more than one structure is associated with optimal responses, the minimal epitope, the smallest structure that induces a response, is curated. It is important to emphasize that there are cases where the minimal epitope does not necessarily give the optimal response. This situation is common with MHC Class II and B cell epitopes. When the minimal epitope is not the optimal epitope, the optimal epitope is captured instead of the minimal epitope. Looking at the data in the example manuscript Figures 3 and 4, the epitope that is captured, GILGFVFTL, is the optimal peptide of those tested and subsequent data is curated under the optimal epitope [Fig. 1A, 2A].

Additionally, authors may further define fine specificity, distinctions in the detailed patterns of reactivity of different T cell clones or monoclonal antibodies when recognizing the same epitope. When multiple T cell clones or antibodies rec-

ognize a sequence of 15 amino acids or less, but differ in their individual reactivities within that stretch of amino acids, the longer sequence will be entered as the epitope. For conformational B cell epitopes, we assume that each monoclonal antibody defines its own epitope unless otherwise indicated.

Finally, a deduced epitope, one that is not directly tested as an isolated structure, but rather identified by the authors through methods such as the use overlapping peptide scans, can be entered as the epitope structure.

EPITOPE ORIGIN

Curation of the origins of each epitope, including the source antigen from which the epitope was derived and the species from which this antigen was derived, can be complicated. For example, the same exact structure is often found in many different species and in cases where the isolated epitope structure is used for experimentation, designation of the epitope's source antigen is therefore arbitrary.

The source antigen for the epitope is assigned utilizing a finder application linked to either Uni-Prot or GenBank IDs. If the authors provide a protein, the epitope is assigned to that source. However, when an ID is not provided, the NCBI's Protein BLAST is used to identify an exact source match in the Uni-Prot or GenBank databases. In the example figure, the epitope was assigned to the GenBank ID of 27596998 because this accession contains exactly the epitope sequence and matches the author identified source name of the epitope [Fig. 1C]. If the epitope sequence cannot be found within the source the authors describe, we contact authors and/or relevant cited references to verify the sequence. If, after verification, a sequence still cannot be assigned to an external source, an internal IEDB Source ID is created.

As mentioned, the same epitope structure may be derived from a number of sources; however, the database only allows reference to a single source antigen/source species per epitope entry. To enter multiple source organisms, multiple epitope entries are required. For example, when an epitope is analyzed for conservancy or cross-reactivity among different natural proteins or pathogens with experimental data presented in the reference for each of the different sources, each of the different natural sources are entered in separate epitope entries. In addition, when the epitopic sequence mutates over time, each new sequence represents a new epitope.

EXPERIMENTAL DATA

The experimental assay data are captured by four main categories, MHC binding, MHC elution, T cell and antibody assays. Each of these are associated with groups of fields such as immunization, antigen, and assay fields. In the following paragraphs, specific rules relating to each of these will be described.

An important issue is to determine to what level of granularity and detail each assay should be curated. To overcome potentially overwhelming the user with inordinately large number of records, we developed rules for the bulk curation of multiple data points as a single record. Common scenarios for bulk curation include contexts with multiple subjects and

dose-response curves. Likewise, whenever a series of assays varies by only one variable, such as the dose of a peptide or adjuvant, only the conditions giving the highest response or utilizing the most clear or sensitive technique is curated.

IMMUNOGEN AND ANTIGEN

The relationships between the epitope structure and the antigen or immunogen are clarified utilizing a menu of defined terms. When the exact epitope structure is used, the immunogen and/or the antigen are designated as such. The immunogen/antigen is described as the epitope source antigen (or source species) when the complete source antigen (or source species) of the epitope is used in the assay. The term fragment of source antigen is used for naturally occurring fragments larger than the epitope structure. This selection is also used for enzymatic or chemical degradation fragments of the source antigen. The immunogen may be left blank only when the immunization type is unknown, and with cancer, autoimmune diseases, and in cases of spontaneous responses.

IMMUNIZATION

The manner in which the host becomes exposed to the immunogen is captured in the immunization category field. When subjects are vaccinated or cells are infected or stimulated in vitro, the designation is administration. All other immunization categories imply that the initial exposure is by natural biological processes. Natural occurrence is used to identify naturally infected or exposed subjects including situations such as allergy, autoimmunity, and cancer. For certain ubiquitous pathogens such as influenza, CMV, EBV and Candida, all humans are presumed naturally exposed. Natural exposure to the ubiquitous pathogen, influenza A virus, is demonstrated in the "Immunization Category" and "Immunogen Name" fields in the example curated context [Fig. 2C]. Additionally, individuals living in particular endemic areas may be assumed exposed to certain pathogens.

To capture host information, a species finder, linked to NCBI taxonomy, is used. Ethnicity is described using NCBI (www.ncbi.nlm.nih.gov/projects/mhc/ihwg.cgi?cmd=PRJOV&ID=9). The disease state of the host is entered via a finder application utilizing the ICD10 codes. The IEDB defines the stage of the disease at the time the experimental data was generated as acute (short-term infection or disease characterized by a dramatic onset and rapid recovery), chronic (long-term infection or illness or partial remission), post (subject has recovered, latent, and complete remission), other, or unknown.

MHC BINDING AND ELUTION DATA

For all MHC binding assays a qualitative outcome is recorded and a quantitative value field is used to capture precise binding values when provided. MHC allele(s) that bind the epitope are recorded exactly as specified by the authors. Alleles that are mutated outside of the binding region are curated as the wild type allele.

Experimental data in which the authors elute peptides or ligands from cell expressed or purified MHC molecules are

also captured. The epitope is then detected in the eluate through sequencing or by a specific T cell line. In these contexts, the immunogen field is used to enter the organism, protein, or peptide fragment that is provided to the cells for processing. In some cases the origin of the eluted peptide is not specifically known. The antigen is always the epitope.

T CELL RESPONSES

T cell responses are captured using a series of fields such as the type of assay, the effector and antigen presenting cells, and the MHC restriction, if known. Responses from multiple T cell clones are curated as one context when the clones recognize the same epitope.

Often the exact MHC restriction of an epitope is determined in the reference as is done in the hypothetical manuscript Figure 4 [Fig. 2A]. This data is not captured as separate contexts, but the outcome is entered in all relevant contexts. Figure 2B shows how the data from the example Figures 3 and 4 would be entered into one curated context. The “Location of Data” field is used to inform end users of all manuscript data that is captured in each context. The “MHC Evidence Code” field specifies the experimental methods used such as binding assays, antibody blocking of either T cell subsets or MHC molecules, and population analysis.

B CELL RESPONSES

Details of antibodies such as the origin and purification status of are described. When a panel of monoclonal antibodies is used, antibodies demonstrating comparable reactivities are bulk curated by entering all antibody names (comma delimited) in the antibody name field. When multiple classes or isotypes are encountered, the most relevant or most common antibody class/isotype used is captured.

Additionally we describe the conformation of the antigen as used in the assay. The antigen is defined as native when the no alteration to the naturally existing tertiary structure of the tested antigen has been made. The antigen is defined as non-native/unknown with short synthetic peptides and proteins that are likely to be denatured in their preparation or by the assay itself.

DIRECTLY SUBMITTED DATA

Data can be directly submitted through the web interface in a machine-readable format (XML). The IEDB system and curators validate the data submitted by the authors in order to confirm that it is in accordance with the IEDB curation rules. Some data integrity rules cannot be sufficiently system-enforced; therefore a manual inspection is done in strict confidence. The IEDB will only expose this data to the public after author approval, preferably immediately following publication.

Epitopes entered through direct submission are not recurated; rather the publication regarding the directly submitted data will be curated as a new reference and linked to the submission. Submitted data will be clearly distinguishable from curated entries from peer-reviewed literature.

CONCLUSIONS

Through continuous reevaluation and expansion of both the database and the criteria used to capture the data, we strive to create a dynamic resource. Areas targeted for expansion in the near future include adoptive transfer, allergy, and autoimmunity. The concept of evidence codes will be evaluated for applicability to other fields of the database. For example, the criteria used to determine whether the epitope structure is minimal or optimal could be described through evidence codes. Evidence codes would also be able to provide additional insight into how definitively the disease state of the host was demonstrated. This need is illustrated by study of populations with presumed exposure to a pathogen due to living in an endemic area as compared to the study of patients diagnosed with a specific disease through definitive criteria. The database and its guidelines must also expand in response to the immunology that it houses. As new experimental assays are created, we must expand our rules and the database structure in order to co-evolve with scientific advances.

Through dissemination of the reasoning behind the IEDB guidelines, we hope to educate the end user as well as the scientists making submissions to the IEDB. This transparency may also facilitate discussion and feedback between the scientists generating the data and those curating the data.

LITERATURE CITED

1. van Baalen CA, Gruters RA, Berkhoff E, Osterhaus A, Rimmelzwaan GF. FATT-CTL assay for detection of antigen-specific cell-mediated cytotoxicity. *Cytometry Part A* 2008;73A:in press, DOI: 10.1002/cyto.a.20613 (this issue).
2. Cecconi V, del Mare S, Moro M, Dellabona P, Casorati G. Use of MHC class II tetramers to investigate CD4+ T cell responses: Problems and solutions. *Cytometry Part A* 2008;73A:in press, DOI: 10.1002/cyto.a.20603 (this issue).
3. Willemsen R, Chames P, Schooten E, Gratama JW, Debets R. Selection of human antibody fragments directed against tumor T-cell epitopes for adoptive T cell therapy. *Cytometry Part A* 2008;73A:in press, DOI: 10.1002/cyto.a.20644 (this issue).
4. Li Pira G, Ivaldi F, Dentone C, Righi E, Del Bono V, Viscoli C, Manca F. Miniaturization and automation for measuring antigen specific T-cell responses. 2008; Presented at the third MASIR Conference, La Plagne, France.
5. Precopio ML, Butterfield TR, Casazza JP, Koup RA, Roederer M. Optimization of T cell epitope identification. *Cytometry Part A* 2008;73A:in press, DOI: 10.1002/cyto.a.20646 (this issue).
6. Peters B, Sidney J, Bourne P, Huynh-Hoa B, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A. The immune epitope database and analysis resource: From vision to blueprint. *PLoS Biol* 2005;3:379–381.
7. Peters B, Sidney J, Bourne P, Huynh-Hoa B, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way S, Wilson S, Sette A. The design and implementation of the immune epitope database and analysis resource. *Immunogenetics* 2005;57:326–336.
8. Vita R, Vaughan K, Zarebski L, Salimi N, Fleri W, Grey H, Sathiamurthy M, Mokili J, Bui HH, Bourne PE, Ponomarenko J, de Castro R, Chan RK, Sidney J, Wilson SS, Stewart S, Way S, Peters B, Sette A. Curation of complex, context-dependent immunological data. *BMC Bioinformatics* 2006;12:341.