Ab and T cell epitopes of influenza A virus, knowledge and opportunities

Huynh-Hoa Bui*, Bjoern Peters*, Erika Assarsson*, Innocent Mbawuike†, and Alessandro Sette*‡

*Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037; and †Department of Molecular Virology, Baylor College of Medicine, Houston, TX 77030

Communicated by Howard M. Grey, La Jolla Institute for Allergy and Immunology, San Diego, CA, November 1, 2006 (received for review October 1, 2006)

The Immune Epitope Database and Analysis Resources (IEDB) (www.immuneepitope.org) was recently developed to capture epitope related data. IEDB also hosts various bioinformatics tools that can be used to identify novel epitopes as well as to analyze and visualize existing epitope data. Herein, a comprehensive analysis was undertaken (*i*) to compile and inventory existing knowledge regarding influenza A epitopes and (*ii*) to determine possible cross-reactivities of identified epitopes among avian H5N1 and human influenza strains. At present, IEDB contains >600 different epitopes derived from 58 different strains and 10 influenza A proteins. By using the IEDB analysis resources, conservancy analyses were performed, and several conserved and possibly cross-reactive epitopes were identified. Significant gaps in the current knowledge were also revealed, including paucity of Ab epitopes in comparison with T cell epitopes, limited number of epitopes reported for avian influenza strains/subtypes, and limited number of epitopes reported from proteins other than hemagglutinin and nucleoprotein. This analysis provides a resource for researchers to access existing influenza epitope data. At the same time, the analysis illustrates gaps in our collective knowledge that should inspire directions for further study of immunity against the influenza A virus.

B lymphocytes | T lymphocytes | conservancy | pandemic cross-reactivity

Influenza A viruses are widely distributed in nature and can infect a variety of birds and mammals. Their genomes consists of eight single-stranded RNA segments that code for 10 different proteins, one nucleoprotein (NP), three polymerase proteins (PA, PB1, and PB2), two matrix proteins (M1 and M2), two nonstructural proteins (NS1 and NS2), and two external glycoproteins [hemagglutinin (HA) and neuraminidase (NA)]. The viruses are classified on the basis of differences in the antigenic structure of HA and NA proteins, with their different combinations representing unique virus subtypes that are further classified into specific strains. Although all known subtypes can be found in birds, currently circulating human influenza A subtypes are H1N1 and H3N2, with intermittent circulation of H1N2 reassortants. Seasonal outbreaks are caused by subtypes already circulating among people, whereas pandemics are caused by either an emerging novel subtype derived by reassortment with avian viruses (1957 A/H2N2 pandemic and 1968 A/H3N2 pandemic), or all-avian (1918 A/H1N1 pandemic). It is possible that some of these pandemics are "recycled" subtypes that had not circulated in human populations for many years. This was, for example, the case for the 1977 "pseudopandemic" of A/H1N1 viruses that resurfaced after 20 years of absence after the 1957 A/H2N2 pandemic.

"Avian" influenza refers to subtypes found chiefly in birds, but infections with these viruses can also occur in humans. Confirmed cases of human disease caused by several subtypes of avian influenza, including H7N7, H9N2, and other emerging avian viruses such as low-pathogenic H5N1 and H5N2, have been reported since 1997 (1–5). However, of the few avian influenza viruses that have crossed the species barrier to infect humans, the emerging high-pathogenic H5N1 virus in Asia has caused the largest number of detected cases of severe disease and death in humans (6). Because these viruses do not commonly infect humans, little or no immunity may be present in the general human population (with the exception of potential cross-reactive immunity originating from exposure to the other strains commonly infecting humans). Therefore, if the high-pathogenic H5N1 virus were to gain the capacity to spread easily from person to person, an influenza pandemic could ensue (7–10).

Results and Discussion

Why Analyze Influenza A-Derived Epitopes? Because of recent events, there has been resurgent interest in the study of influenza A virus in general and avian influenza H5N1 in particular. Further studies must be completed, ranging from basic studies of immune responses and interactions of influenza virus with its hosts, to the evaluation of new vaccine candidates (11–14). Epitopes can be used to accurately monitor immune responses as well as to tease out which influenza responses are specific for a given virus strain or subtype or are cross-reactive with several or most strains.

Immune responses to influenza A virus have been studied for decades, not only as a model system, but also because of their medical importance. However, the vast amount of resulting epitope information available in the literature has not been globally analyzed and made accessible to the scientific community. Herein, we perform such an analysis (*i*) to compile and inventory existing knowledge regarding influenza A epitopes and (*ii*) to determine possible cross-reactivities of identified epitopes among avian H5N1 and human influenza strains. The data source and results of our analysis are available in the Immune Epitope Database and Analysis Resources (IEDB), which was recently developed to capture epitope related data and is publicly available at www.immuneepitope.org (15, 16). Besides the efforts of compiling and making comprehensive epitope information available to the public domain (17), the IEDB also hosts various bioinformatics tools to analyze epitope data (including, for example, population coverage (18) and epitope conservancy) as well as tools to predict epitope cellular processing (19), binding to MHC (20–22), and recognition by T cell receptors and Ab molecules. In the context of this analysis, the conservancy tool provided by the IEDB was used to identify conserved epitopes that might be cross-reactive among avian H5N1 and human influenza strains. Finally, as an outcome of this analysis, important gaps in the global knowledge relating to

Author contributions: A.S. designed research; H.-H.B. and B.P. contributed new reagents/analytic tools; H.-H.B., B.P., E.A., and I.M. analyzed data; and H.-H.B. and E.A. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: HA, hemagglutinin; IEDB, Immune Epitope Database and Analysis Resource; NA, neuraminidase; NP, nucleoprotein.

‡To whom correspondence should be addressed. E-mail: alex@liai.org.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609330104/DC1.

immunity directed against the influenza A virus were also identified, pointing a way forward in immune epitope research.

Ab and T cell epitopes are defined as the molecular structures interacting with Abs and T cell receptor (TCR) molecules, respectively (23). In our analysis of the existing scientific literature relating to influenza A derived epitopes, we considered only epitopes shown to be recognized by Abs or TCR in the context of the whole influenza virus or proteins. We excluded epitopes that were defined solely by their use as immunogens (to induce the responses) and as antigens (to measure the response), because it is not possible to evaluate the relevance of such data with respect to antiviral immune responses.

Historically, a variety of different assays (ranging from T cell proliferation, cytokine production, and ELISAs, to neutralization and protection from live virus challenge) have been used to evaluate the recognition of influenza epitopes. Challenge with live virus and neutralization assays are used to define protective Ab and T cell epitopes. We make no attempt to enforce a common set of criteria for defining immunogenicity and protective efficacy, because widely divergent methodologies were used by different laboratories to measure immune responses. Rather, we record, for each epitope the specific assay category and conditions used, and conform to the criteria for defining positive and negative measurements as reported by the authors themselves in each published article.

We believe that the definition of the structural and functional determinants of influenza-derived epitopes could be useful in detecting and monitoring infections as well as being crucial to project potential cross-reactive immunity and efficacy against new strains by existing vaccines and diagnostics (24, 25), because once the structure of an epitope is known, databases of influenza genomic information such as Influenza Sequence Database (26), Influenza Virus Resources (27) and BioHealthBase (28) can be searched to project whether the same structure is also conserved in all or most influenza strains, or is specific to a particular influenza strain or subtype. The Influenza Sequence Database (www.flu.lanl.gov) contains all published influenza viral sequences that have been curated by domain experts to ensure high standards of accuracy and completeness (26). The Influenza Virus Resource (www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) presents data obtained from the NIAID Influenza Genome Sequencing Project as well as from GenBank, combined with tools for flu sequence analysis and annotation (27). Finally, the BioHealthBase system (www.biohealthbase.org) focuses on six priority pathogens, including influenza, to help fill in gaps in genomic and other data critical to scientific researchers (28).

In a diagnostic and disease-monitoring setting, epitopes that are specific to a given strain or subtype can be used to monitor responses to that particular strain or subset, removing the confounding influence of immune responses derived from previous exposures to partially cross-reactive strains or subtypes (11, 29). One of the shortcomings of the currently available influenza vaccines is the induction of a strain-specific immunity, which requires a new vaccine to be produced each year and for each different strain. In this context, if conserved epitopes can be defined, different immunization regimens and vaccine candidates could be evaluated for their capacity to induce immune responses to those specific conserved determinants.

Conversely, samples from individuals vaccinated and/or naturally infected with viral strains commonly infectious for humans, such as H1N1 and H3N2, could be screened for the presence of cross-reactive immunity. Such cross-reactive immune recognition may represent a minor component of the total response, but its precise mapping would nevertheless be of significant interest. Several groups have analyzed the potential for cross-reactive epitopes, both at the Ab level (between different types of N1) and in the highly conserved internal gene segments. This work constitutes the basis for the recent suggestion that one of the potential strategies to develop universal influenza vaccines relies on the identification of protective and cross-reactive antibodies, followed by the mapping of the epitopes recognized by such antibodies (30).

How Many Influenza A Epitopes Have Been Reported in the Literature?

As mentioned above, immune responses against influenza A virus have been intensely characterized over the course of several decades. However, this knowledge is dispersed over a large number of scientific references, and a simple search in PubMed using the keywords "epitope" and "influenza" reveals >2,000 different scientific reports. It is unclear how many of these reports contain data relating to new epitopes or new information relating to old ones. Furthermore there is no simple way to extract from these references answers to simple questions. For example, how many epitopes are known from strain "X"?; In which host have they been characterized?; Which epitopes are unique, and which are conserved in other strains, and so on. To address these issues, we perform a comprehensive analysis of all epitope data relating to influenza A virus. The analysis consists of two separate tasks: (*i*) data-compilation efforts that involve identification and curation of influenza A epitope literature into IEDB and (*ii*) data-analysis efforts that involve the use of the IEDB-provided conservancy tool to analyze and identify epitopes that are conserved among various avian H5N1 and human influenza strains.

As a first task of the analysis, the current state of knowledge of influenza A-derived Ab and T cell epitopes was determined (Fig. 1). To accomplish this task, a query [see supporting information (SI) Fig. 2] was constructed to identify potentially relevant influenza epitope-related articles from the entirety of published literature available in PubMed. As of May 22, 2006, the PubMed contained >16 million references, of which 2,063 were identified as influenza epitope-related. Running a similar query without any specific constraints on the source of the epitope yielded ≈100,000 references. Thus, a significant fraction (≈2%) of the worldwide epitope literature is related to the flu virus, likely reflecting the extended period that this pathogen has been studied, its biomedical importance, and its use as a model for basic studies in virology, immunology, and vaccinology. By comparison, a similar search in the case of HIV (AIDS) yielded 4,442 references (4.4% of the total). In the case of lymphocytic choriomeningitis virus (LCMV), *Mycobacterium tuberculosis* (tuberculosis) and *Plasmodium* (malaria), the corresponding figures were 472, 856, and 1,397 (0.5%, 0.9%, and 1.4%), respectively. After manual inspection of all abstracts and full-text review of potentially relevant influenza A epitope articles, a total of 429 references were curated in detail (17). Of these references, 103 contained Ab epitope information. In addition, a total of 114, 13, and 291 references, respectively, contained data relating to MHC binding, elution of MHC ligands, and T cell assays.

To determine how many influenza A epitopes have been described in the literature, a query was performed to search data contained in the IEDB. A total of 412 T cell epitopes (175 CD4, 148 CD8, and 89 undefined) and 190 Ab epitopes (75 linear and 115 conformational) were retrieved. These data provide an indication of the wealth of information already available in the scientific literature relating to influenza A epitopes and should constitute a useful resource for researchers worldwide. Given the well-established importance of Ab responses in vaccine efficacy and in prevention of influenza infection, the relatively small number of published Ab epitopes is unexpected. Although the structure and technological means for identifying Ab and T cell epitopes are radically different, given the fact that Ab titers are the only accepted correlate of protection from influenza and of vaccine efficacy, the paucity of Ab epitopes in comparison with T cell epitopes is indeed surprising. The >2:1 ratio of T cell vs. Ab influenza epitopes is likely because of the fact that Ab

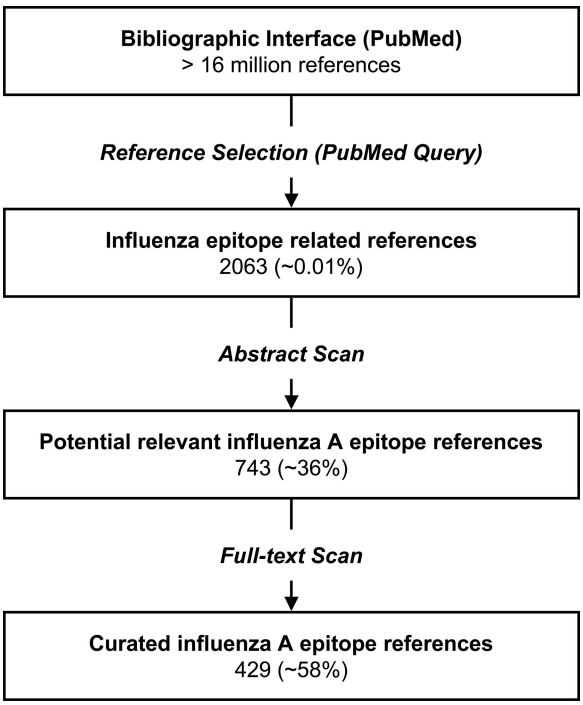

Fig. 1. Process for selection and curation of relevant influenza A epitope literature references.

epitopes are inherently more difficult to characterize than T cell epitopes.

Of 190 identified Ab epitopes, ≈40% are linear sequences. The knowledge of epitope 3D structure can offer important insights into understanding virus neutralization, predicting epitope conservancy across different strains, and rationally designing new vaccine candidates. However, we note that the 3D structures of only 22 epitope/receptor complexes, which represent an average of 4% of all reported epitopes, were determined (an additional 12 epitope/MHC structures have also been described).

The issue of which strain of influenza A was used to define the various epitopes is of obvious importance, in light of the potential use of the epitopes to monitor immune responses to influenza vaccination and infection. Knowledge relating to a diverse set of strains is also desirable to ensure a general biological and immunological relevancy of the results. A lesson learned from HIV research is that excessive reliance on long-term maintained laboratory strains can lead to difficulties in extrapolating results to fresh patient isolates. Our influenza A analysis identifies epitopes from 13 different subtypes and 58 different strains (SI Table 4). The vast majority are from the human influenza H1N1 and H3N2 subtypes, and a relatively large proportion of these epitopes are derived from prototype strains used for model studies, such as A/Puerto Rico/8/34(H1N1) (≈24%) and A/X-31(H3N2) (≈32%), with fewer epitopes having been characterized from fresh isolates of human pathogenic strains (≈1.2%, on average, for a given strain). Only two epitopes from the H5N1 avian influenza A/Viet Nam/1194/2004 are included in this database. These results suggest that more studies need to be focused on the identification of epitopes from the strains responsible for human infections and also point to the urgent need to identify epitopes recognized by responses directed against avian influenza strains. It is, of course, not surprising that the number of epitopes that have been described in either humans or animal models for avian influenza infections are going to be comparatively few compared with the circulating human strains. Some of the original work defining the very

Table 1. Total number of published influenza A epitopes by protein

Protein	Antibody	T cell		Total
		CD4	CD8	
HA	150	113	35	298
NP	3	44	49	96
PA	0	1	11	12
NA	24	7	8	39
M1	4	9	15	28
PB2	0	0	9	9
M2	9	0	3	12
PB1	0	0	10	10
NS1	0	1	7	8
NS2	0	0	1	1

nature of Ab and T cell epitopes used influenza as a model and, as such, these data have been generated for >30 years. The emergence of the avian strains in 1997 (and their reemergence in 2003) has provided far less time for their study; in addition, the increased pathogenicity of fresh isolates has led to their being classified as select agents, making immunological analysis more difficult because of the special containment facilities required. Our analysis demonstrates and underlines this fundamental weakness and gap in our collective knowledge.

Another issue of obvious relevance is the distribution of epitopes by the source proteins from which they are derived (Table 1). It is generally anticipated that Ab responses to vaccination or infection are directed mostly toward epitopes from viral surface-exposed proteins, whereas epitopes recognized by cellular immunity may be broadly derived from both internal and surface proteins. Because internal proteins are far more conserved among different influenza strains and thereby potentially offer the best choice for vaccines aimed at eliciting the broadest possible strain coverage, knowledge of the source proteins from which the epitopes are derived is particularly relevant. Ab epitopes have been identified from only 5 of the 10 viral proteins, and the majority are derived from the virus surface proteins HA, NA, and M2. Compared with HA, fewer Ab epitopes were derived from NA and M2 proteins. T cell epitopes have been identified from all 10 influenza proteins; the highest number of epitopes being derived from HA and NP. Indeed, most published CD4 T cell epitopes are derived from the HA protein, whereas most CD8 T cell epitopes are derived from the NP protein. It should be emphasized that this analysis cannot determine whether the uneven distribution of epitopes as a function of the protein of origin is reflective of poor immunogenicity of those proteins in certain contexts or, perhaps more likely, reflects a bias in the number of studies addressing the immunogenicity of different proteins.

The host species in which the epitopes are identified is shown in Table 2. The majority of Ab and T cell epitopes were identified in mouse, human, or rabbit hosts. Few epitopes are described in birds, which are relevant hosts to study virus evolution. Studies using ferrets, a commonly used experimental model, and non-human primates are also underrepresented. Furthermore, rather astonishingly, only one Ab epitope, compared with 160 T cell epitopes, has been identified by using human samples. Compared with other animal hosts, such as rodents, relatively few human host data available in the literature is probably a reflection of the inherent complexity in characterizing and interpreting immune epitope data from human models, because the repertoire of epitopes recognized in rodents and rabbits is almost invariably measured after a single exposure to influenza. By contrast, in adult humans, the immune response is the final product of a long series of repeated exposures to different viral influenza strains.

Table 2. Total number of published influenza A epitopes by host species

Protein	Ab	T cell	Total
Mouse	71	290	361
Rabbitt	35	0	35
Chicken	3	0	3
Human	1	160	161
Ferret	1	0	1
Goat	1	0	1
Rhesus monkey	1	0	1
Cotton-top tamarin	0	2	2

Responses induced in humans by previous influenza infections or vaccinations might significantly skew the repertoire of epitopes recognized upon infection or vaccination with a different influenza strain, a phenomenon termed "original antigenic sin" (31). This situation highlights the need for more studies defining the Ab epitopes recognized in humans, and the degree to which they overlap with those recognized in animal model systems.

Conservancy of Ab and T Cell Influenza A Epitopes. For the second part of the analysis, we are interested in evaluating conservancy of epitopes among various influenza strains in general and with H5N1 in particular, using the conservancy tool provided by the IEDB. As mentioned above, identification of conserved epitopes is of interest in terms of the prospect for development of broader-spectrum influenza vaccines. Conservancy analysis could also identify epitopes detected from previously vaccinated or infected individuals and associated with cross-reactivity and potential protection from the avian H5N1 strains. Conversely, epitopes that are specific for a given subtype can be used for monitoring responses, removing the confounding influence of immune responses derived from previous exposures to partially cross-reactive strains or subtypes.

To analyze epitope conservancy, we first assembled a collection of representative human and avian H5N1 influenza strains for the analysis, because inclusion of all available sequences would generate a biased conservancy picture reflective of relative abundance of available sequences from a given strain or subtype. For the human influenza strains, our strategy was to select viral strains that had been used for vaccination or were known to cause infection in the human population. A total of 17 influenza strains including 7 H1N1, 8 H3N2, and 2 H5N1 strains were selected for the analysis. Of these, 5 H1N1 and 6 H3N2 strains had been used in annual influenza vaccinations from 1968 to 2004 (SI Table 5). In addition, other pathogenic H1N1 and H3N2 human influenza strains of potential interest, such as A/Brevig Mission/1/18, which circulated in the 1918 pandemic, were also included. The two H5N1 strains that circulated in the 1997 and 2003–2004 H5N1 outbreaks, respectively, were also selected.

Next, using the epitope conservancy analysis tool provided in the analysis resources of the IEDB, we find that, overall, T cell epitopes are more conserved than Ab epitopes (SI Tables 6–8). For T cell epitopes, ≈50% and 30% are conserved at 80% and 90% identity levels, respectively, in both human (H1N1 and H3N2) and avian (H5N1) strains (SI Table 8). At the 100% identity level, 15.0% of T cell epitopes are conserved in the human strains, and 11.4% are also conserved in the avian H5N1 strains. In contrast, only 2.7% of Ab epitopes are conserved at 100% identity level, and <11% were conserved at 80% identity level. A possible reason for this difference is that ≈80% of the linear Ab epitopes, compared with only 40% of the T cell epitopes, are derived from the two most variable influenza proteins, HA and NA. In general, the results suggest that significant levels of interstrain cross-reactivity are likely for T cell epitopes, but much less so for Ab epitopes. Several highly conserved discontinuous conformational Ab epitopes are also identified (SI Table 9). However, their degree of conservation should be interpreted with caution, because pattern-wise conserved discontinuous sequences may not be cross-reactive because of the influence of unknown neighboring and interdispersed amino acids on protein 3D structures. Finally, it should be emphasized that the fact that an epitope is conserved does not necessarily imply that it is also cross-protective.

In this analysis we have organized the data around the subtypes in which the epitopes are found (e.g., H3N2 and H1N1). This is relevant for Ab epitopes for obvious reasons. But it is also relevant for T cell epitopes because, even if an epitope sequence is conserved in different subtypes, flanking regions and differences in the viral genome might affect whether the epitope is recognized as dominant in the context of a different subtype. Furthermore, our analysis will help to determine whether or not a given epitope could be used as a marker for a given subtype. In this context, whereas responses to conserved epitopes might be most useful with respect to vaccine development, subtype-specific epitopes might be most useful for diagnostic purposes and the study of viral evolution.

It should be noted here that the main purpose of the current study is to provide a resource analyzing and making accessible influenza information with potential implications in terms of future research in areas relevant to vaccine research, understanding the role of T cell immunity in influenza, and to highlight multiple pandemic influenza issues surrounding H1N1 and H5N1 viruses in particular. Our analysis is purely bioinformatics and can address only experiments that have been performed and published in peer-reviewed journals. However, the analysis also suggests possible experiments that could be conducted to further validate the epitopes and improve our understanding of the immune response to influenza or ability to combat influenza. For example, several linear and MAb-defined epitopes (SI Table 6) were shown to be highly conserved within H1 or H3 subtypes. Data like these maybe useful for identification of new vaccine targets, and experiments to demonstrate that one of these MAbs indeed neutralized virus *in vitro* or provided passive protection *in vivo* would be neither time-consuming nor technically difficult. Similarly, for the MAb-defined cross-reactive conformational epitopes (SI Table 9), it should be possible to test one or several of these MAbs for cross-reaction with intact viruses. As a result, the effort that has gone into the collection and assembly of influenza-specific information/reagents are justified by its utility and conceivable experimental applications.

Identification of Protective Ab and T Cell Influenza A Epitopes. It is well appreciated that not all Ab and T cell responses are protective. Indeed, responses directed against certain influenza-derived epitopes have been reported in a murine animal model to actually exacerbate disease (32, 33). To address this issue, we focus specifically on epitopes for which protective data are available. Protective epitopes are defined herein as those that tested positive in virus challenge or neutralization assays, even though we are aware that caution needs to be exercised in directly equating *in vitro* neutralization assays with *in vivo* protection. Only nine Ab and nine T cell epitopes are identified to meet this criterion (SI Table 10). As a result, these data emphasize the need for more studies that evaluate the protective and neutralizing efficacy of immune responses directed against different epitopes. In particular, focusing the immune response on relatively conserved epitopes is considered as an avenue to develop influenza vaccines, but their prophylactic efficacy as compared with nonconserved ones must be established.

All data presently available are derived from animal models in

Table 3. Proposed research agenda toward a more systematic and comprehensive collection of influenza immune epitopes

Knowledge gap	Proposed research agenda
Only a few protective Ab and T cell epitopes were reported in the literature	Focus on determining protective Ab and T cell epitopes
Paucity of Ab epitopes in comparison with T cell epitopes	Promote and increase Ab epitope identification studies
Limited spectrum of animal hosts (currently predominantly mouse) used for epitope identification	Expand and balance the repertoire of tested host species, especially avian, nonhuman primates, and human
Limited number of epitopes reported for avian influenza strains/subtypes	Focus on identifying epitopes derived from avian influenza viruses
Limited number of epitopes reported from proteins other than HA and NP	Identify epitopes derived from all 10 influenza proteins

hosts such as mice, rabbits, and macaques. To the best of our knowledge, no study defining human protective epitopes has been conducted, most likely because of ethical reasons. The degree of conservation of protective epitopes across different avian H5N1 and human influenza viral strains is also calculated. In general, protective T cell epitopes are highly conserved between human and avian influenza strains. Protective Ab epitopes are, as expected, less conserved. However, one protective Ab epitope from the M2 protein shows appreciable conservation among the selected human influenza strains and H5N1. Because M2 is a relatively conserved protein, identification of protective Ab epitopes derived from this protein, as has been pointed out, holds promise for the future development of a universal influenza epitope-based vaccine (34). However, it has been shown that even the limited degree of sequence variation between this epitope and the homologous H5N1 sequences might result in lack of cross-reactivity (35). Nevertheless, whether these epitopes could be used to induce cross-reactive responses and also confer protection in humans needs to be addressed experimentally.

An important issue that influenza epitope research must address is which epitopes are likely to confer greatest protection. Cross-protective cytotoxic T lymphoctes (CTL) have been the focus of many studies over the last decade, but their impact on influenza infection in human *in vivo* still needs to be conclusively established. Influenza virus appears to be most sensitive to neutralizing Abs, and Abs to HA are more effective than those specific for NA and M2, perhaps the reason why the virus has evolved to evade such responses, just like herpes viruses have evolved strategies to evade CTL responses. In that respect, it could be difficult to find broadly cross-reactive epitopes.

Conclusions

In summary, a comprehensive analysis of influenza A Ab and T cell epitopes indicates that a large set of influenza epitope data exists for researchers to use in their studies. To the best of our knowledge, all characterized epitopes, defined as presented above, were included in the analysis. If however, inadvertently omitted data were brought to our attention, we would be grateful to update IEDB accordingly. Nevertheless, given the present focus of the scientific community on influenza viruses, the amount of data are likely to increase in the near future. Therefore, we are continually updating the IEDB with new epitope information as it becomes available in the literature. These results are publicly accessible to the scientific community, and we are working to integrate our efforts with other bioinformatics resources such as BioHealthBase (28). Several different protective epitopes are found to be conserved, highlighting how the collation of relevant data from disparate sources, and the integration of immunological data with sequence variability information can yield results of great potential impact.

From our perspective, significant knowledge gaps and opportunities for future research in influenza A epitope identification also became apparent, including (*i*) Determination of protective Ab and T cell epitopes (only a few were reported in the literature), (*ii*) paucity of Ab epitopes in comparison with T cell epitopes, (*iii*) limited spectrum of animal hosts used for epitope identification, (*iv*) a limited number of epitopes reported for avian influenza strains/subtypes, and (*v*) a limited number of epitopes reported from proteins other than HA and NP. Based on these gaps, a proposed research agenda toward a more systematic and comprehensive collection of influenza immune epitopes is tabulated in Table 3.

This is a comprehensive analysis of the world-wide knowledge in a given research area, with the specific intent of not only making curated information accessible to the scientific community, but also with the specific goal of revealing gaps and consequent potential vulnerabilities in the available aggregated knowledge. Some of the results are unexpected and illustrate the power of the approach. Future similar analyses may encompass different disease targets of immunological relevance. The results could assist in correlating the amount of knowledge available with the actual importance of a particular disease, analyzing the impact of funding initiatives and other related topics, and transcending basic research and impacting global research and scientific policies.

In conclusion, influenza research is currently of high general interest. This analysis provides researchers with information that can be used to evaluate different vaccine concepts and design new basic studies. The study also provides the general scientific audience with an objective evaluation of which information is well represented within our current literature, which gaps exist, and what might be addressed by future investigations. In addition, the availability of databases relating to influenza A, as recently pointed out, is an important component of our strategy to combat seasonal outbreaks and a potential pandemic (36). Therefore, the influenza A epitope data analysis reported herein represents an important step in this direction. Specifically, the revealed gaps in our collective knowledge might inspire and guide directions for future research in the study of immunity against the influenza A virus.

Materials and Methods

Selection of IEDB-Curated Influenza A Epitopes. To maximize the immunological relevance of the study, IEDB-curated records were filtered to exclude data in which only the epitope was used as both immunogen and test antigen, because such information does not provide data on recognition of the epitope in the context of the whole virus or protein. T cell epitopes identified by MHC binding alone were also excluded from further analyses, because peptide MHC binding implies only that there exists a potential for immunogenicity but does not prove that this potential has or will be realized. This ability to select records

based on relevant assays is a key example of the IEDB flexibility. It should be noted that because an epitope is defined as a distinct molecular structure that interacts with specific immune receptors, largely overlapping or nearly identical structures were counted as separate entries. Similarly, for Ab conformational epitopes, single residues identified by mutant studies were also considered as separate entries. We have recently developed a computer algorithm to specifically cluster similar and related entries, thus mapping to a single structure or "antigenic site" largely overlapping or homologous (>80%) structures. The results obtained after clustering were qualitatively the same, even though the total number of epitopes was reduced by approximately a third. Development of such a filter was important because the inclusion of the extra sequence changed the "conservation score" between subtypes where the actual epitope is conserved. These duplicate entries could be a hindrance to effective use of the database, especially because these longer sequences for Class I "epitopes" include a sequence that is not actually part of the epitope, that is, actually being trimmed off before presentation. However, the database needs to record the original data as reported to avoid bias and data corruption.

Epitope Conservancy Analysis. To determine the conservation of continuous linear Ab and T cell influenza epitopes, we used the epitope conservancy-analysis tool provided in the analysis resources of IEDB. Using an epitope sequence and a set of protein sequences of a given influenza strain, this tool computes the maximum identity level at which the epitope can be found in the given protein sequence set or the influenza strain. For each epitope, the highest epitope identity level in each influenza strain was calculated. For discontinuous Ab epitopes, the algorithm was implemented to identify a matching epitope discontinuous-sequence pattern in a given protein sequence or set. For example, given the epitope discontinuous sequence "A1,B3,C6", its matching sequence pattern is **AXBXXC**, where X is any amino acid residue, and the number of Xs between two nearest known amino acid residues is equal to the gap distance between them. If an epitope's pattern is found within a protein sequence/set, the epitope is considered to be conserved within that protein sequence/set. In addition, the identity level was also calculated based on the known epitope residues. For patternwise matching sequences, the identity level is 100%. To obtain meaningful results, only discontinuous sequences consisting of at least three identified residues were used in the analysis. We emphasize that the algorithm developed here does not predict cross-reactivity but merely detects whether the residues involved in a conformational epitope are conserved in different sequences. Whether this conservancy would translate in Ab cross-reactivity should be experimentally determined. It should also be noted that, in this analysis, conservancy was calculated and reported for all epitope entries even though similar entries may be related to a single "epitope" (the difference being whether flanking residues were included). Because conservancy is not calculated based on the shared epitope subsequence, different conservancy values are expected in the context of different flanking regions for a single epitope. The differences are due to the variations in sizes and amino acid compositions of flanking regions considered by the algorithm in its calculation.

We thank the IEDB curation team (John Mokili, Jong de Castro, Julia Ponomarenko, Laura Zarebski, Michael Alexander, Muthuraman Sathiamurthy, Nima Salimi, Randi Vita, Russell Chan, Leora Zalman, Huda Makhluf, Michael Lyman, and Kerrie Vaughan) for their excellent work in curating influenza A epitope information and Alison Deckhut Augustine, Howard Gray, Ward Fleri, Muthuraman Sathiamurthy, and Randi Vita for helpful suggestions and critical review of the manuscript. This work was supported by the National Institutes of Health's (National Institute of Allergy and Infectious Disease) Contract HHSN26620040006C (Immune Epitope Database and Analysis Program), Contract N01 AI30039 (Epitope-Based Multipeptide Vaccines for Expanded Coverage Against Inter-Pandemic Influenza), and Kirin pharmaceutical division. This is LIAI publication number 768. E.A. was supported by the Wenner-Gren Foundations.

1. Choi YK, Ozaki H, Webby RJ, Webster RG, Peiris JS, Poon L, Butt C, Leung YH, Guan Y (2004) *J Virol* 78:8609–8614.
2. Fouchier RA, Schneeberger PM, Rozendaal FW, Broekman JM, Kemink SA, Munster V, Kuiken T, Rimmelzwaan GF, Schutten M, Van Doornum GJ, *et al.* (2004) *Proc Natl Acad Sci USA* 101:1356–1361.
3. Guo Y, Li J, Cheng X (1999) *Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi* 13:105–108.
4. Koopmans M, Wilbrink B, Conyn M, Natrop G, van der Nat H, Vennema H, Meijer A, van Steenbergen J, Fouchier R, Osterhaus A, Bosman A (2004) *Lancet* 363:587–593.
5. Peiris M, Yuen KY, Leung CW, Chan KH, Ip PL, Lai RW, Orr WK, Shortridge KF (1999) *Lancet* 354:916–917.
6. World Health Organization (WHO) (2006) www.who.int/csr/disease/avian_influenza/country/cases_table_2006_10_03/en/index.html.
7. Enserink M (2006) *Science* 311:932.
8. Fauci AS (2006) *Cell* 124:665–670.
9. Fauci AS (2006) *Emerg Infect Dis* 12:73–77.
10. U.S. Department of Health and Human Services (HHS) (2005) www.hhs.gov/pandemicflu/plan/.
11. Doherty PC, Turner SJ, Webby RG, Thomas PG (2006) *Nat Immunol* 7:449–455.
12. Enserink M (2005) *Science* 309:996.
13. Wadman M (2005) *Nature* 438:23.
14. Sambhara S, Poland GA (2006) *Lancet* 367:1636–1638.
15. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, *et al.* (2005) *PLoS Biol* 3:e91.
16. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, *et al.* (2005) *Immunogenetics* 57:326–336.
17. Vita R, Vaughan K, Zarebski L, Salimi N, Fleri W, Grey H, Sathiamurthy M, Mokili J, Bui HH, Bourne PE, *et al.* (2006) *BMC Bioinformatics* 7:341.
18. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A (2006) *BMC Bioinformatics* 7:153.
19. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG (2003) *J Immunol* 171:1741–1749.
20. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O (2003) *Protein Sci* 12:1007–1017.
21. Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothe BR, Chisari FV, Watkins DI, Sette A (2005) *Immunogenetics* 57:304–314.
22. Peters B, Sette A (2005) *BMC Bioinformatics* 6:132.
23. Janeway CA, Travers P, Walport M, Shlomchik MJ (2005) *Immunobiology* (Garland Science, New York).
24. Gupta V, Earl DJ, Deem MW (2006) *Vaccine* 24:3881–3888.
25. Munoz ET, Deem MW (2005) *Vaccine* 23:1144–1148.
26. Macken C, Lu H, Goodman J, Boykin L (2001) in *Options for the Control of Influenza*, eds Osterhaus ADME, Cox N, Hampson AW (Elsevier Science, Amsterdam), pp 103–106.
27. National Center for Biotechnology Information (NCBI) (2004) www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html.
28. BioHealthBase (2006) http://www.biohealthbase.org.
29. Johnson PR, Feldman S, Thompson JM, Mahoney JD, Wright PF (1986) *J Infect Dis* 154:121–127.
30. Palese P, Tumpey TM, Garcia-Sastre A (2006) *Immunity* 24:121–124.
31. Fazekas de St. Groth S, Webster RG (1966) *J Exp Med* 124:331–345.
32. Crowe SR, Miller SC, Woodland DL (2006) *Vaccine* 24:452–456.
33. Crowe SR, Miller SC, Shenyo RM, Woodland DL (2005) *J Immunol* 174:696–701.
34. Kaiser J (2006) *Science* 312:380–382.
35. Fan J, Liang X, Horton MS, Perry HC, Citron MP, Heidecker GJ, Fu TM, Joyce J, Przysiecki CT, Keller PM, *et al.* (2004) *Vaccine* 22:2993–3003.
36. Editorial (2006) *Nature* 440:255–256.