D1124–D1132 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 19 November 2013*
doi:10.1093/nar/gkt1133

Updates on the web-based VIOLIN vaccine database and analysis system

Yongqun He[1,2,3,4,*], Rebecca Racz[1,5], Samantha Sayers[1,6], Yu Lin[1], Thomas Todd[1,7], Junguk Hur[8], Xinna Li[1], Mukti Patel[9], Boyang Zhao[10], Monica Chung[9], Joseph Ostrow[9], Andrew Sylora[9], Priya Dungarani[9], Guerlain Ulysse[9], Kanika Kochhar[9], Boris Vidri[9], Kelsey Strait[9], George W. Jourdian[11,12] and Zuoshuang Xiang[1]

[1]Unit for Laboratory Animal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA, [2]Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA, [3]Center for Computational Medicine and Biology, University of Michigan Medical School, Ann Arbor, MI 48109, USA, [4]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA, [5]College of Pharmacy, University of Michigan, Ann Arbor, MI 48109, USA, [6]School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA, [7]Division of Comparative Medicine, University of South Florida, Tampa, FL 33612, USA, [8]Department of Neurology, University of Michigan, 48109, Ann Arbor, MI, USA, [9]College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI 48109, USA, [10]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI 48109, USA, [11]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA and [12]Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Received September 21, 2013; Revised October 20, 2013; Accepted October 24, 2013

ABSTRACT

The integrative Vaccine Investigation and Online Information Network (VIOLIN) vaccine research database and analysis system (http://www.violinet.org) curates, stores, analyses and integrates various vaccine-associated research data. Since its first publication in NAR in 2008, significant updates have been made. Starting from 211 vaccines annotated at the end of 2007, VIOLIN now includes over 3240 vaccines for 192 infectious diseases and eight noninfectious diseases (e.g. cancers and allergies). Under the umbrella of VIOLIN, >10 relatively independent programs are developed. For example, Protegen stores over 800 protective antigens experimentally proven valid for vaccine development. VirmugenDB annotated over 200 'virmugens', a term coined by us to represent those virulence factor genes that can be mutated to generate successful live attenuated vaccines. Specific patterns were identified from the genes collected in Protegen and VirmugenDB. VIOLIN also includes Vaxign, the first web-based vaccine candidate prediction program based on reverse vaccinology. VIOLIN collects and analyzes different vaccine components including vaccine adjuvants (Vaxjo) and DNA vaccine plasmids (DNAVaxDB). VIOLIN includes licensed human vaccines (Huvax) and veterinary vaccines (Vevax). The Vaccine Ontology is applied to standardize and integrate various data in VIOLIN. VIOLIN also hosts the Ontology of Vaccine Adverse Events (OVAE) that logically represents adverse events associated with licensed human vaccines.

INTRODUCTION

Vaccination is one of the most significant inventions in modern medicine. It has been used to dramatically improve human health. However, our efforts to develop vaccines to protect against many diseases have not been successful. For example, the infectious diseases AIDS, tuberculosis and malaria are three of the top five threats to human health (1), but there is not an effective and safe vaccine available against any of these diseases. Vaccines can also be developed against many noninfectious diseases, including cancer, allergy and autoimmune diseases. More funding has been added to extensive vaccine research from governments and nonprofit foundations. For example, Gates Foundation has donated billions of dollars to invest in vaccine research and development. It has been anticipated that the Gates Foundation donation, combined with commitments

*To whom correspondence should be addressed. Tel: +1 734 615 8231; Fax: +1 734 936 3235; Email: yongqunh@umich.edu

from the USA and other governments could prevent the deaths of 8 million children from 2010 to 2019 (2). Resulting from intensive vaccine research and development, a large volume of data and publications has been published. A recent study has confirmed that the records of vaccine-related literature stored in PubMed (3) are increasing at an exponential rate (4).

To address the challenge of integrating and analyzing published vaccine-related results, we have developed the Vaccine Investigation and Online Information Network (VIOLIN, http://www.violinet.org) (5). VIOLIN has become the largest, web-based vaccine database and analysis system for vaccine researchers. The vaccines annotated in VIOLIN include licensed vaccines, vaccines being tested in clinical trials and vaccines that have been studied in research and experimentally verified effective in at least one laboratory animal model. The vaccine data collected for each vaccine covers vaccine components (e.g. vaccine adjuvants), protection efficacy and host immune responses. VIOLIN also contains many specific software programs such as several for vaccine literature mining and vaccine design. The first VIOLIN paper was published in the Database Issue of the *Nucleic Acids Research* journal in 2008 (5). Since its first publication, dramatic progress has been made. This article aims to summarize the major changes and updates since 2008.

OVERALL SYSTEM DESIGN, ANNOTATION PIPELINE AND STATISTICS

VIOLIN is currently implemented using a three-tier architecture built on two HP ProLiant DL380 G6 servers that run a Red Hat Linux operating system (Red Hat Enterprise Linux ES 4). Users submit database queries through the web. These queries are processed using PHP/SQL (middle-tier, application server based on Apache) against a MySQL (version 5.0) relational database (back-end, database server). The result of each query is then presented to the user through the web browser. The data stored in these two servers are routinely backed up by each other and with additional data storage space available at the University of Michigan.

All the annotated information in VIOLIN is obtained by manual curation from peer-reviewed literature or other reliable websites. PubMed (3) is our major source for obtaining peer-reviewed publications. Manual curation emphasizes the retrieval of experimental evidence of vaccine efficacy and ensures the accuracy of vaccine information in VIOLIN. Additionally, a web-based curation and literature mining system (Limix) was used (5,6). The interactive Limix system allows data curators to search literature, copy and edit text, add references, submit data to the database and check submission history. Limix also provides the data reviewers a platform and tools to review, edit and approve the curated data. All these features are implemented and available for use on a user-friendly web interface. Data submitted to VIOLIN is reviewed by an expert and, upon approval, is published to the VIOLIN website.

As shown in our original VIOLIN publication, VIOLIN contained ~200 vaccines or vaccine candidates against 18 pathogens in the end of 2007 (5). After 5 years of diligent work, VIOLIN now includes over 3200 vaccines or vaccine candidates against 192 infectious diseases and eight noninfectious diseases. Table 1 summarizes a list of most annotated pathogens and diseases and related vaccine information stored in VIOLIN. More details can be found on the VIOLIN statistics page: http://www.violinet.org/stat.php.

VIOLIN includes many relatively independent programs (e.g. Protegen and Vaxjo) targeting specific vaccine-related domains (e.g. protective antigens and vaccine adjuvants) (Figure 1 and Table 2). Their relations are described in Figure 1. Specifically, a vaccine is developed to protect (or treat) a host against a disease. The disease can be an infectious disease caused by an infectious microbe or a noninfectious disease such as cancer or autoimmune disease. A vaccine has different components such as protective antigen and vaccine adjuvant. Depending on the classification criteria, different vaccine types exist, such as live attenuated vaccines and DNA vaccines. Once administered, a vaccine will induce specific immune responses including humoral antibody response and cell-mediated immunity. The level of protection is considered as the gold standard for evaluating the efficacy of a vaccine (Figure 1).

The individual VIOLIN programs share common annotated data in the general VIOLIN database. However, these individual programs often contain additional data that is unique for the program such as vaccine adjuvant-specific data (e.g. adjuvant structure), and is not typically found in the description of a specific vaccine. These individual programs also include their own query interfaces, as well as other related information including BLAST analysis (7) and websites that provide our annotated data for downloading.

VACCINE-RELATED PATHOGEN GENES/PROTEINS

In modern vaccine research, it is critical to identify genes and proteins that can be directly used for vaccine development. Two types of pathogen genes or proteins are used for developing vaccines. One type is the protective antigens that are able to induce antigen-specific protective immunity. Another type is microbial virulence factors that can be mutated in virulent pathogens to make live attenuated vaccines. VIOLIN has incorporated individual programs specifically targeted to these two types of genes and proteins.

Protegen: database of protective antigens

Protegen was developed in 2010 to store and analyze protective antigens (Table 2) (8). To be qualified as a protective antigen and included in Protegen, it is required that this antigen is used for development of an experimentally verified vaccine or has been experimentally shown to induce an immune response (e.g. production of neutralization antibody) that correlates with protection. This is a key difference between the antigens collected in Protegen

Table 1. Representative VIOLIN statistics as of 18 October 2013

Pathogen (disease)	No. of vaccines and licensed vaccines	No. of protective antigens used	No. of virmugens used
Gram-positive bacteria			
Clostridium tetani (tetanus)	60 (57)[a]	2	0
Mycobacterium tuberculosis (Tuberculosis)	43 (2)	26	15
Erysipelothrix rhusiopathiae (Erysipelas)	29 (28)	1	0
Bacillus anthracis (Anthrax)	26 (4)	13	1
Corynebacterium diphtheriae (Diphtheria)	22 (22)	1	0
Gram-negative bacteria			
Leptospira spp. (Leptospirosis)	132 (130)	2	0
Salmonella spp. (Salmonellosis)	62 (19)	6	46
Brucella spp. (Brucellosis)	60 (7)	25	15
Escherichia coli (Hemorrhagic colitis)	40 (14)	17	4
Haemophilus influenzae (Meningitis)	30 (11)	14	0
Viruses			
Bovine herpesvirus 1 (Infectious bovine rhinotracheitis)	159 (146)	7	2
Influenza virus [Influenza (flu)]	153 (89)	49	2
Bovine viral diarrhea virus 1 [Bovine viral Diarrhea (BVD)]	129 (128)	0	0
Bovine Parainfluenza 3 Virus (BPIV-3)	108 (108)	0	0
Newcastle disease virus (Newcastle disease)	97 (95)	3	0
Parasite			
Plasmodium spp. (Malaria)	36 (0)	33	7
Leishmania donovani (Visceral leishmaniasis)	15 (1)	12	1
Toxoplasma gondii (Toxoplasmosis)	14 (0)	12	3
Trypanosoma cruzi (Chagas disease)	14 (0)	16	0
Eimeria spp. (Coccidiosis)	11 (8)	1	0
Fungi			
Coccidioides spp. (Coccidioidomycosis)	4 (0)	9	0
Noninfectious disease			
Cancer	52 (2)	72	0
Arthritis	4 (0)	4	0
Diabetes	4 (0)	5	0
Atherosclerosis (Atherosclerosis, arteriosclerotic vascular disease)	2 (0)	0	0
Allergy	1 (0)	15	0

[a]The number in parentheses corresponds to the number of licensed vaccines.

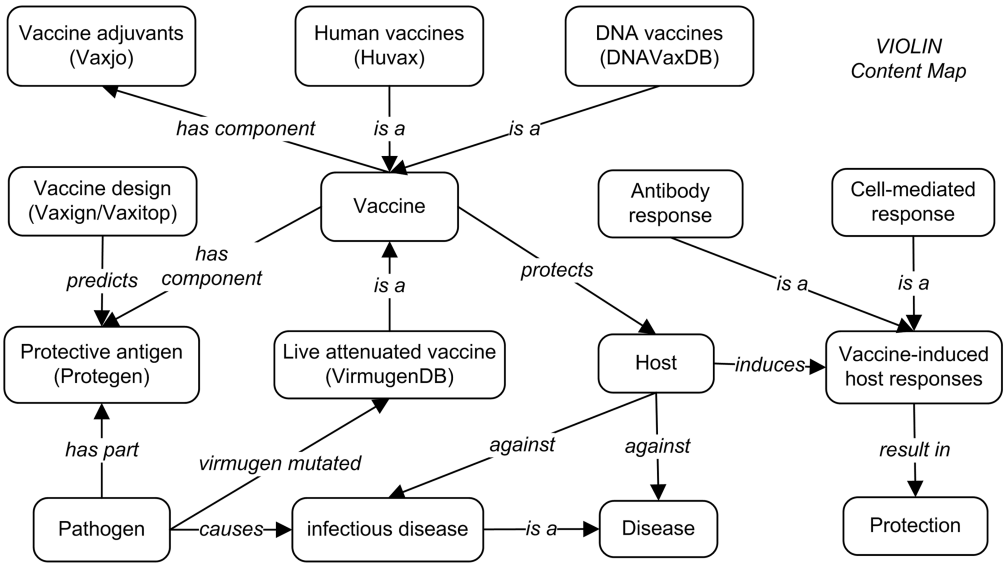

Figure 1. Overview of major VIOLIN components and their relations. The program names inside parentheses are databases or tools. For components with no quoted program names, the information is listed in the general VIOLIN database.

and those in some other databases, such as AntigenDB that focuses on the induction of immune responses without requiring protection (9). Currently, Protegen holds over 800 protective antigens from 200 pathogens (Table 2). Over 200 protective antigens have been added since the Protegen paper was published in 2010. The protective antigens collected in Protegen have been used in development of different types of vaccines, particularly protein subunit vaccines, DNA vaccines and recombinant vector vaccines (8).

Table 2. Statistics of specific VIOLIN programs as of 18 October 2013

Program	Description	URL	Release year	Reference	New or old	Comment
Pathogen genes/proteins and vaccine candidate prediction						
Protegen	857 protective antigens	http://violinet.org/protegen	2009	(8)	new	–
VirmugenDB	225 virmugens	http://violinet.org/virmugendb	2011	(12)	new	–
VBLAST	BLAST vaccine genes	http://violinet.org/blast	2007	(5)	old	
Vaxign	vaccine candidate prediction	http://violinet.org/vaxign	2009	(15,16,40)	new	precomputed or dynamic
Vaxitop	Immune epitope prediction	http://violinet.org/vaxign/vaxitop	2009	(16,40)	new	–
Vaccine components (other than protective antigens)						
Vaxjo	93 vaccine adjuvants	http://violinet.org/vaxjo	2010	(22)	new	–
DNAVaxDB	144 DNA vaccine plasmids	http://violinet.org/dnavaxdb	2010	(23)	new	–
Vaxvec	Vaccine vectors	http://violinet.org/vaxvec	2010	–	new	
Specific vaccine types						
DNAVaxDB	419 DNA vaccines	http://violinet.org/dnavaxdb	2012	(23)	new	for 99 pathogens
VirmugenDB	207 attenuated vaccines	http://violinet.org/virmugendb	2011	(12)	new	for 57 pathogens
Vevax	>1000 licensed veterinary vaccines	http://violinet.org/vevax	2011	–	new	for 106 pathogens
Huvax	184 licensed human vaccines	http://violinet.org/huvax	2011	–	new	for 29 pathogens
Vaccine literature mining						
VO-SciMiner	*Brucella* vaccine-gene interaction network	http://violinet.org/vo-sciminer	2011	(32)	new	–
Vaxmesh	MeSH-based literature mining	http://violinet.org/litesearch/meshtree/meshtree.php	2007	(5)	old	–
Litesearch	Vaccine literature search	http://violinet.org/litesearch/keywords_search.php	2007	(5)	old	–
Hosting of community efforts						
VO	Community-based VO	http://violinet.org/vaccineontolog	2008	(26,27)	new	>4800 terms
OVAE	Ontology of Vaccine Adverse Events	http://www.violinet.org/ovae/	2013	(41)	new	>1300 terms
ICoVax2012	2012 Computational vaccinology workshop	http://violinet.org/icovax2012	2012	(35)	new	–
ICoVax2013	2013 Computational vaccinology workshop	http://violinet.org/icovax2013	2013	–	new	–
Data retrieval (other than protective antigens)						
V-Utilities	Data query and retrieval utilities	http://violinet.org/v-utilities	2010	–	new	for software programming

We have also performed data analysis to find specific patterns among the Protegen data (10). For example, among 201 protective protein antigens from Gram-negative bacteria, 48% of antigens are extracellular or cell wall proteins and ~40% of protective antigens are adhesins or adhesin-like proteins. Among 69 protective protein antigens from Gram-positive bacteria available in Protegen, 64% of protective antigens belong to extracellular or outer membrane proteins, and 54% of protective antigens are adhesins or adhesin-like proteins. Many conserved domains, including autotransporter and TonB domains, are enriched in these bacterial protective antigens (10).

In addition to our own data analysis, the data in Protegen have been used as the gold standard for evaluation of different vaccine candidate protection methods (11). In the study conducted by Jaiswal *et al.* (11), the Protegen data were used to evaluate four software programs, including Vaxign, in prediction of protein vaccine candidates. Many nonadhesin and nonsurface bacterial vaccine candidates collected in Protegen has been the challenge for prediction by different vaccine candidate prediction programs.

VirmugenDB: database of 'virmugens'

Various virulence factors exist as part of a pathogen, and not all virulence factors can be knocked out to make an effective live attenuated vaccine. The term 'virmugen' was coined by Dr Yongqun He to represent a gene encoding a virulent factor of a pathogen that, when mutated, has been proven feasible in laboratory animal studies to create a live attenuated vaccine (12). Currently, VirmugenDB contains over 220 virmugens that were mutated to make more than 200 vaccines experimentally verified as useful for vaccination against over 50 bacterial, viral and protozoal pathogens (Table 2). Significant patterns were identified from analysis of virmugen data. For example, the *aroA* gene has been used in 10 Gram-negative and one Gram-positive bacteria as a virmugen. The *aroA* gene sequences in the 10 Gram-negative bacteria share at least 50% identity (12). This gene encodes for a key enzyme involved in aromatic amino acid biosynthesis. This finding suggests that interference of the aromatic amino acid biosynthesis pathway provides a good strategy for live attenuated vaccine development. Indeed, the analysis of all virmugens found that virmugens tend to involve metabolism of nutrients (e.g. amino acids, carbohydrates, nucleotides) and cell membrane formation. Compared with other virulence factors, it is likely that virmugens have specific characteristics, the study of which deserves more investigation. Host genes whose expressions were regulated by virmugen mutation vaccines or wild-type virulent pathogens were also annotated and compared with an ultimate aim to identify the protective immune mechanisms specifically targeted by vaccines (12).

Customized BLAST analysis programs

A commonly used tool for gene or protein sequence similarity search is BLAST (7). Several customized BLAST programs are available in VIOLIN. Protegen and

VirmugenDB include customized BLAST libraries of DNA and protein sequences of protective antigens and virmugens, respectively, for sequence similarity searches. DNAVaxDB also includes a BLAST search program for comparing user-provided gene or protein sequence(s) with protective antigens used in development of DNA vaccines. In addition, VIOLIN also has a VBLAST program for searching all pathogen genes and proteins annotated in the VIOLIN system. Different BLAST programs (e.g. blastn, blastp and RPS BLAST) are included as well.

Vaxign/Vaxitop: genome sequence-based vaccine candidate prediction

As an emerging vaccine development strategy in the genomics era, reverse vaccinology initiates vaccine development from bioinformatics analysis of genome sequences (13). Over the last decade, different parameters have been included for vaccine candidate prediction. The Vaxign program in VIOLIN is the first web-based vaccine design software program based on the reverse vaccinology strategy (14,15). The Vaxign pipeline predicts potential vaccine protein candidates based on the prediction of the following criteria: subcellular localization, transmembrane helices, adhesin probability, microbial sequence conservation by ortholog analysis, exclusion of proteins having orthologs in selected genome(s), similarity to host proteins and identification of major histocompatibility complex (MHC) Class I or Class II immune epitopes. The MHC Class I and II binding epitope prediction is performed using Vaxitop, an internally developed tool. Vaxitop is a position-specific scoring matrix (PSSM)-based epitope prediction program. Whereas other tools use an arbitrary percentage or rank cutoff, Vaxitop relies on a statistical *P*-value to examine the likelihood of a candidate peptide being an immune epitope (14,16).

Vaxign has been demonstrated to effectively predict vaccine candidates for *Brucella* spp. (15,17,18), uropathogenic *Escherichia coli* (14) and human herpesvirus 1 (16). The Vaxign program has also been used for genome reannotation and prediction of virulence factors. For example, Vaxign has been applied to reannotate genome and predict virulence factor genes using the genome sequences of *Campylobacter fetus* subspecies (19) and *Corynebacterium diphtheriae* NCTC13129 (20). Currently over 350 genomes have been precomputed using the Vaxign pipeline (Table 2). The predicted results can be queried using a user-friendly web interface. Vaxign also allows users to perform dynamic vaccine candidate predictions by inputting specific sequences of up to 500 proteins.

VACCINE COMPONENTS

Vaccines typically contain multiple components. One of the most critical components is the protective antigen(s). As described above, VIOLIN Protegen contains the information of known protective antigens. Live attenuated vaccines contain mutations of virulence factors (i.e. virmugens) leading to the attenuation phenotype. The VIOLIN VirmugenDB collects a list of known virmugens. However, a virmugen is not considered as a vaccine component since it is not part of the vaccine recipe. There are many other types of vaccine components (21). Below we introduce two more VIOLIN programs collecting two specific types of vaccine components:

Vaxjo: database of vaccine adjuvants

A vaccine adjuvant is a vaccine component used to accelerate, prolong or enhance host immune responses to coadministered protective antigens in a vaccine. Vaccine adjuvants have different actions *in vivo*. They may modify the cytokine immune network, deliver and present antigens to appropriate immune effector cells, induce CD8+ cytotoxic T-lymphocyte (CTL) responses or generate a short-term or long-term depot to give a continuous or pulsed release. Vaxjo is a vaccine adjuvant database that annotates and stores various vaccine adjuvants and their usage in different vaccines (22). Currently, Vaxjo contains 93 vaccine adjuvants used in 378 vaccines for over 70 pathogens (Table 2). For each vaccine adjuvant, Vaxjo introduces its name, components, preparation, vaccines in VIOLIN that utilize each adjuvant and at least one reliable reference. Different types of vaccine adjuvants are collected. The commonly identified vaccine adjuvant types with highest numbers of adjuvants include 28 synthetic adjuvants, 18 microorganism-derived adjuvants, 15 emulsion adjuvants and 13 mineral salt adjuvants. Aluminum hydroxide is the most common adjuvant found, with 62 associated vaccines collected in VIOLIN. Freund's complete and incomplete adjuvants are also commonly used with each being associated with 42 vaccines (22).

DNA vaccine plasmids collected in DNAVaxDB

As of 14 September 2013, 141 DNA vaccine plasmids have been annotated in the DNAVaxDB (23) (Table 2). These plasmids have been used in generation of over 400 DNA vaccines. Among the most commonly used plasmids were pcDNA3.1, pcDNA3, pVAX1, pVR1012 and pCI. Specific patterns have been identified by analyzing the plasmids collected in VIOLIN. The most commonly used promoter is the human cytomegalovirus virus (CMV) immediate-early promoter that elicits higher expression levels. Some plasmids have been more frequently used for development of DNA vaccines against one type of pathogen than the others. For example, 10 Gram-negative bacterial DNA vaccines use the plasmid pCMVi-UB, but this plasmid has not been used in DNA vaccines against any Gram-positive bacteria, viruses or parasitic pathogens (23).

The VIOLIN database has also included the information of other vaccine components such as Vaxvec for collection and analysis of vaccine vectors (e.g. bacterial vaccine vectors, viral vaccine vectors). More work is necessary to systematically annotate and classify such information.

SPECIFIC VACCINE TYPES

VIOLIN includes commercial vaccines as well as those still undergoing preclinical or clinical trials. Here we introduce two programs targeting two different types of vaccines:

Huvax: licensed human vaccines databases

Huvax collects and allows query of licensed human vaccine data. Huvax has curated all 104 human vaccines currently licensed in the USA and Canada, including 27 bacterial vaccines, 47 viral vaccines and 30 combination vaccines. The annotated data for each licensed human vaccine cover vaccine types, preparation, adjuvants, preservatives, allergens, age groups, administration routes, manufacturers, immune responses and adverse events (AEs). Different patterns have been found from the analysis of data for all human licensed vaccines. For example, aluminum salts, including $Al(OH)_3$ and $Al(PO)_4$, have been found to be the most commonly used adjuvants. In addition, several preservatives, including phenol, thimerosal and 2-phenoxyethanol (24), have been commonly used in human vaccine preparation.

DNAVaxDB: DNA vaccines

DNAVaxDB is designed to store and analyze specifically DNA vaccines and their related plasmids and protective antigens. Currently, DNAVaxDB holds over 417 DNA vaccines using 141 DNA vaccine plasmids and 375 protective antigens (Table 2). These vaccines are developed against 99 infectious and noninfectious diseases (including arthritis, cancer and diabetes). To meet the needs for many researchers who are only interested in DNA vaccines, independent web query interfaces have also been developed to query the DNA vaccines, plasmids and protective antigens used in DNA vaccines.

APPLICATIONS OF VACCINE ONTOLOGY ON VIOLIN DATA INTEGRATION AND LITERATURE MINING

Application of VO on VIOLIN data exchange and integration

Originally VIOLIN used VIOLINML, an eXtensible Markup Language (XML)-based format for VIOLIN data exchange (5). Over the past few years, we have switched to rely on the community-based Vaccine Ontology (VO) for data exchange. A biomedical ontology is a set of terms and relations that represent entities in a biomedical domain and how they relate to each other, and terms in ontologies are typically expressed in computer and human interpretable formats to support automated reasoning (25). VO is a biomedical ontology in the vaccine and vaccination domain (21,26,27). The development of VO follows the OBO Foundry principles, including openness, collaboration, and using a common shared syntax (28). Using the Web Ontology Language (OWL) format (http://www.w3.org/TR/owl2-quick-reference/), VO is developed to support machine processing and automated reasoning. In order to properly and efficiently develop and analyze VO, we have also developed several software programs (25,29,30), which have been widely used by the ontology community for development and analysis of other biomedical ontologies.

As demonstrated in the Ontobee program (30), currently VO has over 4800 ontology terms (http://www.ontobee.org/ontostat.php?ontology=VO). These terms cover most vaccines, vaccine components (e.g. protective antigens, adjuvants), virmugens and vaccination types stored in VIOLIN. Other top-level terms and term relations are also included in VO. Through systematic alignments with top level ontologies, VO logically represents these vaccine-specific terms and the relations among them and other terminologies, such as pathogens, diseases and vaccines.

VO-SciMiner: VO-based literature mining

The SciMiner literature mining program supports literature indexing and gene name tagging (31). By integrating VO and SciMiner, VO-SciMiner was developed to retrieve, store and analyze vaccines, microbial genes and vaccine-gene interaction networks based on literature mining of PubMed articles (32). VO-SciMiner was first evaluated using the bacterial model of *Brucella*, a Gram-negative bacterium that causes zoonotic brucellosis in humans and various animals (6). A set of rules was set up for term expansion and literature indexing of VO terms. Using 100 manually annotated biomedical articles, VO-SciMiner demonstrated high recall (91%) and precision (99%) for indexing PubMed papers. The asserted and inferred VO hierarchies provide semantic support. As a result, VO-SciMiner indexing exhibited superior performance in retrieving *Brucella* vaccine-related papers over the MeSH-based PubMed literature search method. Using extracted abstracts for all *Brucella*-related papers, VO-SciMiner identified 140 *Brucella* genes associated with *Brucella* vaccines. These *Brucella* genes included protective antigens, virulence factors, and other vaccine-related genes. The enriched biological functional categories of these genes were also identified. An integrative interaction network of *Brucella* vaccines and genes were constructed and used to address different questions. A web-based query interface has been developed to facilitate its use (Table 2). Our study shows that VO-SciMiner can be possibly developed to improve the efficiency for PubMed searching in the vaccine domain. The expansion of VO-SciMiner to other pathogens is underway.

ONTOLOGY-BASED REPRESENTATION AND ANALYSIS OF VACCINE ADVERSE EVENTS

Although licensed vaccines are in general very safe, they sometimes induce different types of adverse events (AEs) in vaccine recipients. In the USA, the Vaccine Adverse Event Reporting System (VAERS) has been used for decades for collecting different vaccine AE (VAEs) cases (33). The Ontology of Adverse Events (OAE) is a community-based biomedical ontology in the area of AEs (33,34). OAE has been used to analyze VAERS AE data (33). Furthermore, to better represent and analyze vaccine AEs, we developed the Ontology of Vaccine Adverse

Events (OVAE; http://www.violinet.org/ovae) by extending the OAE and the VO. OVAE was used to represent and classify the AEs recorded in package insert documents of commercial vaccines licensed in the USA. With over 1300 terms, OVAE includes 87 distinct types of VAEs associated with 63 licensed human vaccines licensed in the USA. The OAE can be used to answer different questions such as the top 10 vaccines associated with the highest numbers of VAEs and the top 10 VAEs most frequently observed among vaccines. More efforts will be made to use OVAE for better analysis of VAERS data.

VIOLIN VACCINE DATA QUERY

Vaxquery (http://www.violinet.org/vaxquery) is the primary data query system developed to search curated vaccine data and related information stored in the VIOLIN system. The default keyword search provides four sections of output containing the keyword(s): vaccines, pathogens, vaccine-related genes and vaccine-related literature. Vaxquery also provides a set of advanced searching programs (http://www.violinet.org/vaxquery/adv_vaxquery.php). The advanced Vaxquery search can be performed in three ways: a vaccine search, a pathogen search or a hierarchical data comparison. For each of these methods, a user can type keywords for specific parameters (e.g. vaccine trade name, antigen, adjuvant). The advanced hierarchical search and comparison program provides a hierarchical structure of the VIOLIN data and allows users to display selected vaccine information. These query and visualization approaches offer the users to customize their search for vaccine-related information.

In addition to Vaxquery, different VIOLIN programs (e.g. Protegen, Huvax) have their own query interfaces. These specific query programs search only the information in specific domains (e.g. protective antigens, human licensed vaccines).

OTHER VIOLIN PROGRAMS

Several other VIOLIN programs have been developed. For example, in addition to VO-SciMiner, three other vaccine literature mining programs exist: Litesearch, Vaxmesh and Vaxlert. These programs exist in the original VIOLIN paper published in 2008 (5). Litesearch is a simple literature search of vaccine-related publications. Vaxlert is a program that provides newly published vaccine papers and literature email alerts. Vaxmesh includes a MeSH tree hierarchy and publication records related to each MeSH term in the tree hierarchy.

Several new VIOLIN programs are being developed. Vevax is a licensed veterinary vaccine database. Compared with human vaccines, many more animal vaccines have been licensed. Currently Vevax contains over 1000 licensed vaccines for 17 animal species. For analysis and study of vaccine-related molecular mechanisms, VIOLIN provides two programs Vaxism and Vaxar. Vaxism focuses on introducing basic information of microbial pathogenesis, protective immunity and animal models. Vaxar targets the classification and analysis of animal responses to vaccinations. Based on Vaxar, so far we have collected vaccine-induced responses from 35 host species. For software programmers to query and retrieve data, VIOLIN also provides a programming utility service (V-Utilities; http://www.violinet.org/v-utilities).

The VIOLIN website has also been used to host several community-based efforts. For example, VIOLIN is the website that hosts the project of VO (http://www.violinet.org/vaccineontology). VIOLIN also hosts the official websites for two International Computational Vaccinology workshops (ICoVax): ICoVax 2012 (http://www.violinet.org/icovax2012) (35) and ICoVax 2013 (http://www.violinet.org/icovax2013).

DISCUSSION

The VIOLIN development in the past 5 years has proven very productive. According to Google Analytics, VIOLIN has been visited by ~60 000 unique visitors since 2008, and more visits have been seen in the last 2 years. This article introduces many individual VIOLIN programs (e.g. Protegen, VirmugenDB, Vaxjo and Vaxign), most of which have newly been developed since 2008. These programs can also be integrated for the study of a specific pathogen. For example, we have previously reported the application of different VIOLIN programs to simultaneously study vaccines for *Brucella* (17). Similar approaches can be used to study other pathogens.

The mechanisms of vaccine-induced protections to various diseases remain unclear. One largely ignored research area is the identification of mechanisms by which successful vaccines stimulate protective immunity. Systems vaccinology provides a feasible strategy to tackle this problem (36,37). Systematic annotation of host genes whose expressions are induced by vaccines allows for the collection and meta-analysis of experimentally verified results identified from a large volume of peer-reviewed publications. Our previous ontology-based meta-analysis study allowed the identification of experimental factors that significantly contribute to the protection efficacy of whole organism *Brucella vaccines* (42,43). Analysis of omics data from publically available high-throughput data repositories can also provide valuable novel insights regarding mechanism. As such, we are currently exploring these possibilities to gain better understanding of vaccine-specific protective immunity and to potentially allow the identification of early innate signatures for immunogenicity of vaccines, discover novel immune regulation mechanisms and support rational vaccine design.

The semantic web aims at extending the existing web of documents into a web of data designed to be processed automatically (38). Within the Semantic Web framework, a movement known as 'Linked Open Data' (LOD) has emerged with the goal of publishing various open datasets using machine-parsable language such as Resource Description Framework (RDF) on the web and establishing good practices for sharing this data (30,39). The VO provides a foundation for integrating

various vaccine datasets. Based on VO and other related ontologies, we have planned to develop a 'Linked Open Vaccine Data' (LOVD) system to support deep data integration and sharing. Such a LOVD system will promote further basic and translational vaccine research and development.

One limitation of our manual curation of vaccine-related information is that often we cannot contain up-to-date and complete information in such a time of fast-growing publications. A potential time delay in updating our database is expected. This is why we will frequently miss newly developed vaccines or vaccine components published in peer-reviewed journals. As a result, a failure to find some information in the database does not mean such information does not exist in the literature. Although there exists literature mining programs to automatically extract relevant information, we find that there are no programs with sufficient high quality and accuracy when compared with manual curation by trained researchers. Nevertheless, we do provide some literature mining and curation programs, as shown in our Limix and VO-SciMiner programs, to facilitate manual curation. We recognize this tradeoff, but ultimately we envision VIOLIN to be a resource that provides high-quality information about various aspects of vaccine, at the expense of potential delay in providing the most up-to-date information. In addition, one major goal of our study is to identify scientifically sound patterns and hypotheses from curated data, which often does not require an inclusion of all possible data. For example, our VirmugenDB study shows that many genes encoding for enzymes involving the metabolism of nutrients (e.g. amino acids, carbohydrates and nucleotides), such as the *aroA* gene encoding for key enzyme important for the aromatic amino acid biosynthesis, have been frequently knocked out for making live attenuated vaccines (12). Such a finding could be generated with all the possible papers that we had found but might not be complete.

Over the past years, we have focused on annotation and analysis of preventive vaccines against infectious diseases. In the future, we will expand to cover vaccines against other types of diseases and expand the coverage of existing vaccine types (e.g. cancer vaccines). As therapeutic vaccine development becomes more extensive in those research fields, manual annotation and analysis of data on therapeutic vaccines will become a primary research topic in our continued VIOLIN project development. We anticipate that VIOLIN will continue to be a comprehensive and crucial source for vaccine knowledge collection, vaccine data analysis and rational vaccine design.

FUNDING

Supported by NIHNIAIDR01 grant [#R01AI081062 to Y.H., for R.R., S.S., Y.L., X.L. and Z.X.]; Startup Funding (to Y.H.) from the University of Michigan; Supported by the Undergraduate Research Opportunity Program (UROP) (to Y.H.) at the University of Michigan (for J.O., A.S., P.D., G.U., K.K., B.V. and K.S.). Funding for open access charge: National Institutes of Health R01 grant [R01AI081062].

Conflict of interest statement. None declared.

REFERENCES

1. Feachem,R.G. (2004) The research imperative: fighting AIDS, TB and malaria. *Trop. Med. Int. Health*, **9**, 1139–1141.
2. D'Argenio,D.A. and Wilson,C.B. (2010) A decade of vaccines: integrating immunology and vaccinology for rational vaccine design. *Immunity*, **33**, 437–440.
3. Coordinators,N.R. (2013) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
4. He,Y., Rappuoli,R., De Groot,A.S. and Chen,R.T. (2010) Emerging vaccine informatics. *J. Biomed. Biotechnol.*, **2010**, 218590.
5. Xiang,Z., Todd,T., Ku,K.P., Kovacic,B.L., Larson,C.B., Chen,F., Hodges,A.P., Tian,Y., Olenzek,E.A., Zhao,B. *et al.* (2008) VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res.*, **36**, D923–D928.
6. Xiang,Z., Zheng,W. and He,Y. (2006) BBP: *Brucella* genome annotation with literature mining and curation. *BMC Bioinformatics*, **7**, 347.
7. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
8. Yang,B., Sayers,S., Xiang,Z. and He,Y. (2011) Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res.*, **39**, D1073–D1078.
9. Ansari,H.R., Flower,D.R. and Raghava,G.P. (2010) AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res.*, **38**, D847–D853.
10. He,Y. and Xiang,Z. (2012) Bioinformatics analysis of bacterial protective antigens in manually curated Protegen database. *Procedia Vaccinol.*, **6**, 3–9.
11. Jaiswal,V., Chanumolu,S.K., Gupta,A., Chauhan,R.S. and Rout,C. (2013) Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics*, **14**, 211.
12. Racz,R., Chung,M., Xiang,Z. and He,Y. (2013) Systematic annotation and analysis of "virmugens" - virulence factors whose mutants can be used as live attenuated vaccines. *Vaccine*, **31**, 797–805.
13. Rappuoli,R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol.*, **3**, 445–450.
14. He,Y., Xiang,Z. and Mobley,H.L. (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.*, **2010**, (2010), Article ID 297505, 15 pages.
15. Xiang,Z. and He,Y. (2009) Vaxign: a web-based vaccine target design program for reverse vaccinology. *Procedia Vaccinol.*, **1**, 23–29.
16. Xiang,Z. and He,Y. (2013) Genome-wide prediction of vaccine targets for human herpes simplex viruses using Vaxign reverse vaccinology. *BMC Bioinformatics*, **14**, S3.
17. He,Y. and Xiang,Z. (2010) Bioinformatics analysis of *Brucella* vaccines and vaccine targets using VIOLIN. *Immunome Res.*, **6(Suppl. 1)**, S5.
18. Gomez,G., Pei,J., Mwangi,W., Adams,L.G., Rice-Ficht,A. and Ficht,T.A. (2013) Immunogenic and invasive properties of *Brucella melitensis* 16M outer membrane protein vaccine candidates identified via a reverse vaccinology approach. *PLoS One*, **8**, e59751.
19. Ali,A., Soares,S.C., Santos,A.R., Guimaraes,L.C., Barbosa,E., Almeida,S.S., Abreu,V.A., Carneiro,A.R., Ramos,R.T., Bakhtiar,S.M. *et al.* (2012) *Campylobacter fetus* subspecies: comparative genomics and prediction of potential virulence targets. *Gene*, **508**, 145–156.

20. D'Afonseca,V., Soares,S.C., Ali,A., Santos,A.R., Pinto,A.C., Magalhaes,A.A.C., de Jesus Faria,C., Barbosa,E., Guimaraes,L.C., Eslabao,M. *et al.* (2012) Reannotation of the Corynebacterium diphtheriae NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria. *Open Access Bioinform.*, **4**, 1–13.
21. Lin,Y. and He,Y. (2012) Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses. *J. Biomed. Semantics*, **3**, 17.
22. Sayers,S., Ulysse,G., Xiang,Z. and He,Y. (2012) Vaxjo: a web-based vaccine adjuvant database and its application for analysis of vaccine adjuvants and their uses in vaccine development. *J. Biomed. Biotechnol.*, **2012**, 831486.
23. Racz,R., Li,X., Patel,M., Xiang,Z. and He,Y. (2013) DNAVaxDB: the first web-based DNA vaccine database and its data analysis. *BMC Bioinformatics*, In press.
24. Coudeville,L., Bailleux,F., Riche,B., Megas,F., Andre,P. and Ecochard,R. (2010) Relationship between haemagglutination-inhibiting antibody titres and clinical protection against influenza: development and application of a bayesian random-effects model. *BMC Med. Res. Methodol.*, **10**, 18.
25. Xiang,Z., Courtot,M., Brinkman,R.R., Ruttenberg,A. and He,Y. (2010) OntoFox: web-based support for ontology reuse. *BMC Res. Notes*, **3**, 175.
26. He,Y., Cowell,L., Diehl,A.D., Mobley,H.L., Peters,B., Ruttenberg,A., Scheuermann,R.H., Brinkman,R.R., Courtot,M., Mungall,C. *et al.* (2009) VO: Vaccine Ontology. *The 1st International Conference on Biomedical Ontology (ICBO 2009)*. Buffalo, NY, USA, *Nat. Preced.*, doi:10.1038/npre.2009.3553.1.
27. Ozgur,A., Xiang,Z., Radev,D.R. and He,Y. (2011) Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J. Biomed. Semantics*, **2(Suppl. 2)**, S8.
28. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
29. Xiang,Z., Lin,Y. and He,Y. (2012) Ontorat web server for automatic generation and annotations of new ontology terms. *Proceedings of the International Conference on Biomedical Ontologies (ICBO), University of Graz, Graz, Austria, July 24–27, 2012*.
30. Xiang,Z., Mungall,C., Ruttenberg,A. and He,Y. (2011) Ontobee: A Linked Data Server and Browser for Ontology Terms. *Proceedings of the International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26-30.*, 279–281.
31. Hur,J., Schuyler,A.D., States,D.J. and Feldman,E.L. (2009) SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics (Oxford, England)*, **25**, 838–840.
32. Hur,J., Xiang,Z., Feldman,E.L. and He,Y. (2011) Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network. *BMC Immunol*, **12**, 49.
33. Sarntivijai,S., Xiang,Z., Shedden,K.A., Markel,H., Omenn,G.S., Athey,B.D. and He,Y. (2012) Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines. *PLoS One*, **7**, e49941.
34. He,Y., Xiang,Z., Sarntivijai,S., Toldo,L. and Ceusters,W. (2011) AEO: a realism-based biomedical ontology for the representation of adverse events. *Adverse Event Representation Workshop, International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26–30, 2011. Proceeding of ICBO-2011*, 309–315.
35. He,Y., Cao,Z., De Groot,A.S., Brusic,V., Schönbach,C. and Petrovsky,N. (2013) Computational vaccinology and the ICoVax 2012 workshop. *BMC Bioinformatics*, **14(Suppl. 4)**, I1.
36. Nakaya,H.I., Li,S. and Pulendran,B. (2012) Systems vaccinology: learning to compute the behavior of vaccine induced immunity. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **4**, 193–205.
37. Pulendran,B., Li,S. and Nakaya,H.I. (2010) Systems vaccinology. *Immunity*, **33**, 516–529.
38. Berners-Lee,T., Hendler,J. and Lassila,O. (2001) The Semantic Web. *Scientific American*, 29–37.
39. Bizer,C., Heath,T. and Berners-Lee,T. (2009) Linked Data - The Story So Far. *Int. J. Seman. Web Inform. Syst.*, **5**, 1–22.
40. He,Y., Xiang,Z. and Mobley,H.L. (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.*, **2010**, 297505.
41. Marcos,E., Zhao,B. and He,Y. (2013) The Ontology of Vaccine Adverse Events (OVAE) and its usage in representing and analyzing adverse events associated with US-licensed human vaccines. *J. Biomed. Semantics*, **4**, 40.
42. He,Y., Xiang,Z., Todd,T., Courtot,M., Brinkman,R., Zheng,J., Stoeckert,C.J., Malone,J., Rocca-Serra,P., Sansone,S. *et al.* (2010) Ontology representation and ANOVA analysis of vaccine protection investigation. *Proceeding of Bio-Ontologies 2010: Semantic Applications in Life Sciences, International Society for Computational Biology (ISMB), July 9-10, 2010. Boston, MA, USA*.
43. Todd,T.E., Tibi,O., Lin,Y., Sayers,S., Bronner,D.N., Xiang,Z. and He,Y. (2013) Meta-analysis of variables affecting mouse protection efficacy of whole organism *Brucella* vaccines and vaccine candidates. *BMC Bioinformatics*, **14(Suppl. 6)**, S3.