Software

PHIDIAS: a pathogen-host interaction data integration and analysis system

Zuoshuang Xiang*†‡, Yuying Tian§ and Yongqun He*†‡

Addresses: *Unit for Laboratory Animal Medicine, University of Michigan, 1150 W. Medical Dr., Ann Arbor, MI 48109, USA. †Department of Microbiology and Immunology, University of Michigan, 1150 W. Medical Dr., Ann Arbor, MI 48109, USA. ‡Center for Computational Medicine and Biology, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, USA. §Medical School Information Services, University of Michigan, 535 W. William St., Ann Arbor, MI, USA.

Correspondence: Yongqun He. Email: yongqunh@umich.edu

Published: 30 July 2007

Genome ***Biology*** 2007, **8:**R150 (doi:10.1186/gb-2007-8-7-r150)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/7/R150

Received: 23 March 2007
Revised: 8 June 2007
Accepted: 30 July 2007

Abstract

The Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) is a web-based database system that serves as a centralized source to search, compare, and analyze integrated genome sequences, conserved domains, and gene expression data related to pathogen-host interactions (PHIs) for pathogen species designated as high priority agents for public health and biological security. In addition, PHIDIAS allows submission, search and analysis of PHI genes and molecular networks curated from peer-reviewed literature. PHIDIAS is publicly available at http://www.phidias.us.

Rationale

An infectious disease is the result of an interactive relationship between a pathogen and its host. According to estimations of the World Health Organization, infectious diseases caused 14.7 million deaths in 2001, accounting for 26% of the total global mortality [1]. Integration and analysis of various data related to pathogens and pathogen-host interactions (PHIs) will yield a better understanding of, and means for, the control of infectious diseases induced by such pathogens.

Completely sequenced genomic information provides valuable information for gene and protein functions, and intra-organismic processes. Pathogen genome information also lays a foundation for the study of the interactions between host and microbial organisms. Several genome data resources, such as the National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB), are available to the public. However, data obtained from these sources often are not integrated. Lack of such integration prompted us to develop the *Brucella* Bioinformatics Portal (BBP) [2]. This program allows integration of data from more than 20 sources including information on the *Brucella* genome. The same strategy can be expanded to include other pathogens, thereby enhancing our ability to conduct comparative studies. The program can be modified to include additional features not yet available in BBP. For example, protein conserved domains (distinct units of molecular evolution usually associated with particular molecular functions) could be listed. The NCBI Conserved Domain Database (CDD) mirrors several collections, including the Protein families database of alignments (Pfam) [3], Simple Modular Architecture Research Tool (SMART) [4], and Clusters of Orthologous Groups (COG) [5], and thus provides comprehensive information about conserved protein domains. Conserved domains are critical for protein functions and provide important clues about microbial pathogenesis and interactions between pathogens and hosts.

While CDD contains conserved domains derived from various eukaryotic and prokaryotic organisms [6], it is difficult to compare and analyze pathogen-specific conserved domains. The availability of a program that permits the acquisition and storage of pathogen-specific domain information in an integrated system would be extremely useful, as would the combination of such a database with BLAST search programs and other programs for the determination of sequence analyses. To facilitate comparison and better understanding of pathogens and fundamental PHI mechanisms, it is necessary to integrate genome information from publicly important pathogens with effective tools for browsing, searching, and analyzing annotated genome sequences and conserved domains. Such an integrated system would also benefit from the inclusion of large amounts of published literature data relating to pathogens and their interactions with host immune systems. To allow machine-readable data exchange of the now voluminous pathogen information, He *et al.* [7] developed an Extensible Markup Language (XML)-based Pathogen Information Markup Language (PIML). PIML contains comprehensive pathogen-oriented information, including pathogen taxonomy, genomic information, life cycle, epidemiology, induced diseases in host, diagnosis, treatment, and relevant laboratory analysis. A list of PIML documents addressing pathogens deemed of high priority for public health and biological defense have been created and are available on the worldwide web or through a web service [7]. However, compared to relational databases, XML databases do not efficiently support query functions and scalability. These deficiencies prompted us to design a web-based relational database system to store and query PIML data. The database system can also integrate efficiently other PHI-related data, including manually curated information related to the pathobiology and management of laboratory animals that are given high priority pathogens [8].

The molecular functions of pathogen and host genes as well as their roles in specific PHI pathways have been extensively studied. Molecules that play important roles in the virulence of pathogens and in the host immune defense are particularly important for PHI. A systematic collation from the literature of these molecules and their functions is lacking. Once PHI-related molecules are collated, the next step is to illustrate molecular interactions and pathways involving these molecules. Existing pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9], BioCyc [10,11], and Biomolecular Interaction Network Database (BIND) [12], contain pathways for various metabolic and molecular interactions of different organisms. Although richly documented, the networks of microbial and host molecular and cellular interactions that occur during pathogenic infections of hosts are underrepresented in current database systems. He and colleagues [13] developed the Molecular Interaction Network Markup Language (MINetML, previously called ProNetML) to summarize information related to microbial pathogenesis. However, MINetML cannot be exchanged with other standard data exchange formats such as the Biological Pathways Exchange format (BioPAX) [14]. This deficiency prevents active data exchange and communication with biological pathway databases. In addition, there is no effective MINetML visualization tool available.

Experimental methodologies, including microarrays and mass spectrometry, provide abundant sources of gene expression data. Publicly available gene expression data repositories, including the NCBI Gene Expression Omnibus (GEO) [15] and the EBI ArrayExpress [16] store large amounts of gene expression data, much of which is related to interactions between pathogens and hosts. Summaries of gene expression experiments and gene profiles allow querying and comparison of PHI-related gene expression patterns.

To better understand the intricate interactions between pathogens and hosts, we have now developed a web-based PHI data integration and analysis system (PHIDIAS) that permits integration and analysis of genome sequences, curated literature data for general PHI information and PHI networks, and PHI-related gene expression data. PHIDIAS currently targets 42 pathogens. These include most category A, B, and C priority pathogens identified by the National Institute of Allergy and Infectious Diseases (NIAID) and the Centers for Disease Control and Prevention (CDC) in the USA, and other pathogens deemed of high priority with regards to public health, such as the human immunodeficiency virus (HIV) and *Plasmodium falciparum* (Table 1).

System design

PHIDIAS is implemented using a three-tier architecture built on two Dell Poweredge 2580 servers that run the Redhat Linux operating system (Redhat Enterprise Linux ES 4). Users can submit database or analysis queries through the web. These queries are then processed using PHP/Perl/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 in the web browser. Two servers are scheduled to regularly backup each others' data.

PHIDIAS includes six components that search and analyze annotated genome sequences, curated PHI data, and PHI-related gene expression data (Figure 1a). Pathogen genomes are displayed and analyzed by PGBrowser, Pacodom, and BLAST searches. The PGBrowser has been developed to browse and analyze the gene and protein sequences of 77 genomes from 42 bacterial, viral, and parasitic pathogens (Table 1). Although PHDIAS does not include non-pathogenic species, PHIDIAS includes genomes from both pathogenic strains (for example, *Escherichia coli* O157:H7 strain Sakai) and non-pathogenic strains (for example, *E. coli* strain K12) in the same pathogen species. Pacodom is used to search and analyze conserved protein domains of the pathogen genomes.

Table 1

Forty-two pathogens included in PHIDIAS

	Pathogens (disease)	CDC/NIAID category	No. of genomes	Phinfo	Pacodom	Phinet
1	*Bacillus anthracis* (anthrax)	A/A	3	√	4,588	√
2	*Brucella* spp. (brucellosis)	B/B	4	√	4,267	√
3	*Burkholderia mallei* (glanders)	B/B	1	√	4,679	√
4	*Burkholderia pseudomallei* (Melioidosis)	B/B	2	√	5,093	
5	*Campylobacter jejuni* (food safety threat)	/B	2		3,235	
6	*Clostridium botulinum* (botulism)	A/A	0	√	N/A	√
7	*Clostridium perfringens* (epsilon toxin)	B/B	1		3,770	
8	*Coxiella burnetii* (Q fever)	B/B	1	√	3,032	√
9	*Escherichia coli* (food safety threat)	B/B	6	√	5,440	√
10	*Francisella tularensis* (tularemia)	A/A	2	√	3,057	√
11	*Helicobacter* spp. (gastric ulcer)		5		3,374	
12	*Legionella pneumophila* (legionnaires' disease)		3		3,974	
13	*Listeria monocytogenes* (food safety threat)	/B	2		3,999	
14	*Mycobacterium tuberculosis* (tuberculosis)	/C	2	√	3,991	
15	*Rickettsia prowazekii* (typhus fever)	/C	1	√	2,129	√
16	*Rickettsia rickettsii* (Rocky Mountain spotted fever)	/C	0	√	N/A	√
17	*Salmonella enterica* (food safety threat)	B/B	4	√	5,150	√
18	*Shigella* spp. (food safety threat)	B/B	5	√	5,211	√
19	*Vibrio* spp. (water safety threat)	B/B	5		5,449	
20	*Yersinia pestis* (plague)	A/A	5	√	4,828	√
21	Crimean-Congo hemorrhagic fever virus (tickborne hemorrhagic fever)	C/C	1	√	4	√
22	Eastern equine encephalitis virus (encephalitis)	B/B	0	√	N/A	√
23	Foot-and-mouth disease virus (foot-and-mouth disease)		7	√	3	
24	Guanarito virus (viral hemorrhagic fever)	A/A	1	√	0	√
25	Human immunodeficiency virus (AIDS)		2	√	8	
26	Junin virus (viral hemorrhagic fever)	A/A	1	√	0	√
27	Lassa virus (viral hemorrhagic fever)	A/A	1	√	0	√
28	Louping ill virus (encephalomyelitis)		1	√	6	√
29	Machupo virus (viral hemorrhagic fever)	A/A	1	√	0	√
30	Marburg virus (viral hemorrhagic fever)	A/A	1	√	N/A	√
31	Measles virus (measles)		1	√	0	√
32	Newcastle Disease Virus (Newcastle disease)		0	√	N/A	
33	Powassan virus (encephalitis)		0	√	N/A	√
34	Reston ebola virus (viral hemorrhagic fever)	A/A	1	√	1	√
35	Rift Valley fever virus (Rift Valley fever)	/A	1	√	3	√
36	Variola virus (smallpox)	A/A	2	√	129	
37	Venezuelan equine encephalitis virus (viral encephalitis)	B/B	1	√	8	√
38	Yellow fever virus (yellow fever)	/C	1	√	5	√
39	*Cryptosporidium parvum* (cryptosporidiosis)	B/B	0	√	N/A	
40	*Coccidioides immitis* (meningitis)		0	√	N/A	
41	*Phakopsora pachyrhizi* (soybean rust)		0	√	N/A	√
42	*Plasmodium falciparum* (malaria)		0	√	N/A	
	Total (42 pathogens)		77	37	75,433	27

The program includes 20 bacteria (54 genomes), 18 viruses (23 genomes), and 4 parasites. The database contains 75,433 conserved domains (7,919 unique PSSMs) and PHI network information for 27 pathogens.

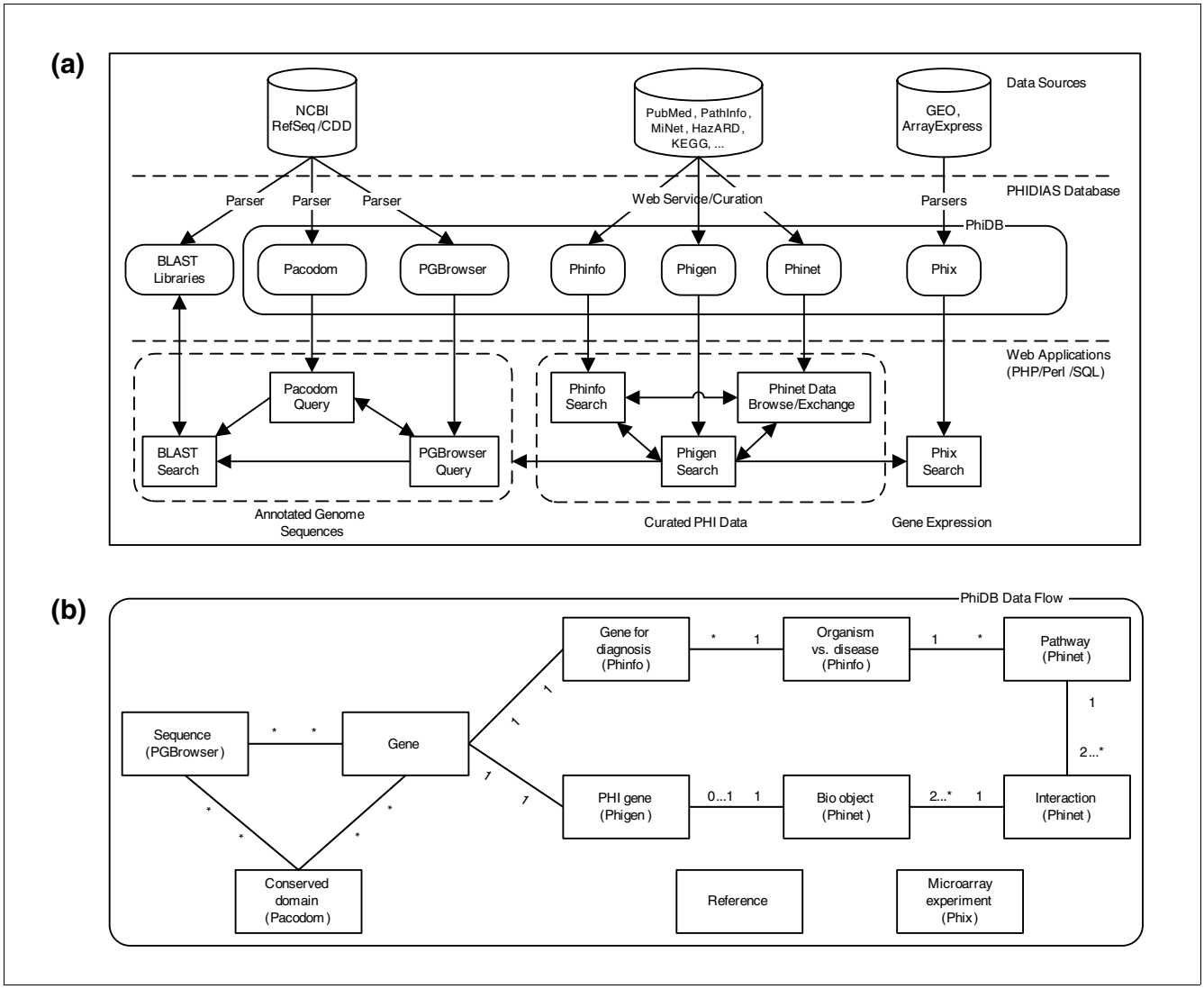


Figure 1
PHIDIAS data flow. **(a)** The PHIDIAS system architecture. **(b)** PhiDB data flow among key elements of different PhiDB database modules. The relationships among these elements are represented by the following signs: *, zero or more; 1, one; and 2...*, two or more. For example, the labeling of a pathway with '1' and '2...*' indicates that one pathway includes two or more interactions.

Customized BLAST programs allow users to perform similarity searches on pathogen genome sequences. Curated PHI data are separated into Phinfo, Phigen and Phinet, based on general PHI information, PHI molecules and networks, respectively. PHI gene expression experiments and gene profiles are searched through the Phix database system.

PhiDB is the PHIDIAS relational database that integrates different PHIDIAS components. Figure 1b illustrates the relationship and data flow among different database modules and PHIDIAS components. PhiDB integrates PHI-related data from more than 20 public databases (Table 2) and from data curated by the PHIDIAS curation team. PhiDB contains gene information, including sequences, conserved domains from pathogen genomes as well as gene information for PHI and diagnosis of pathogen infections. The biological objects (Bio Object) in the data flow diagram are flexible, that is, they can be a gene or gene product, or any other molecular or cellular entity, including metabolites, cell membrane, mitochondria and so on. The Bio Object element also enables representation of a cluster or group of molecules such as virulent factors and protective antigens. Each interaction includes two or more Bio Objects that function as input or output objects. Each pathway contains more than one interaction. General information pertaining to each pathogenic organism and each disease is available and integrates with pathway and gene information. PHI-related gene expression experiments are also recorded. Detailed information for references, including peer-reviewed journal publications, reliable websites and databases for each of the components is also stored. Each of

Table 2

Public databases and software programs integrated in PHIDIAS

Resources	Databases and analysis programs	Comments
Databases		
NCBI	RefSeq	Reference sequences
	Genome	Genome summary
	Gene	Gene information
	Protein	Protein information
	Nucleotide	Nucleotide information
	CDD	Conserved domains
	COGs	Clusters of orthologous groups
	Taxonomy	*Brucella* taxonomy information
	PubMed	Biomedical publications
	GEO	Gene expression database
EBI and SIB	ArrayExpress	Gene expression database
	Swissprot	Annotated protein data
	TrEMBL	Protein data
	InterPro	Protein families, domains and functions
	PROSITE	Protein families and domains
VBI	PathInfo	PIML documents via web service
	MiNet	MiNetML documents via web service
TIGR	CMR	Comprehensive microbial resource
	TIGRfam	TIGRfam assignments
	GO	Gene ontology
	KEGG	Pathways
	BioCyc	Biological pathways
	PFam	Protein domains and families
	ProDom	Protein domain families
	PDB	Protein database
University of Michigan	BBP	*Brucella* bioinformatics portal
	HazARD	Hazards in animal research database
Software programs integrated		
NCBI	BLAST	Blastn, blastp, blastx, tblastn, tblastx, PSI/PHI Blast, Mega Blast, Blast 2 sequences
GMOD	GBrowse	Genome browsing and analysis
	BioPerl	Programming tools
	BioPAX	Biological pathway data exchange format

CMR, TIGR Comprehensive Microbial Resource; GO, Gene Ontology; MeSH, Medical Subject Headings; PDB, Protein Data Bank.

the PHIDIAS components focuses on different PhiDB elements. All of these components are integrated together and readily available for biomedical researchers working on different pathogens and PHI systems.

To illustrate the features of data integration and comparative analyses using PHIDIAS, the pathogenic *Brucella* serves as an example and demonstrates how PHIDIAS can promote *Brucella* research. *Brucella* species are Gram-negative, facultative intracellular bacteria that cause brucellosis in humans and animals [17]. *B. melitensis*, *B. suis*, *B. abortus*, and *B. canis* are human pathogens in decreasing order of severity. *Brucella* species have been identified as priority agents amenable for use in biological warfare and bioterrorism and are listed as USA NIAID category B priority pathogens. The genomes of *B. melitensis* strain 16 M [18], *B. suis* strain 1330 [19], and *B. abortus* strain 994-1 [20] and strain 2308 [21] have been sequenced and published.

PHIDIAS components

PGBrowser: pathogen genome browser

Pathogen genomes serve as the foundation for the study of PHI in the post-genomic era. PGBrowser integrates data from

more than 20 different sources, including NCBI, EBI, and The Institute for Genomic Research (TIGR) (Table 2). Currently, PGBrowser stores 77 genome sequences and 203,297 features from 42 pathogens. NCBI Entrez Programming Utilities are used to download genome information for the pathogens selected from Reference Sequences (RefSeq) and other NCBI databases. The information obtained is formatted in XML. A script has been developed to parse all the protein/gene features, including raw sequences. These are stored in the PhiDB database. Another script has also been developed to query UniProt and other EBI databases, and to download all of the protein information that relates to the 42 pathogens using the SwissProt format. The information is then parsed and stored in a database based on Locus Tag matches. The molecular weights and isoelectric points (pI) are calculated from the protein sequences using the modules (Bio::Tools::pICalculator and Bio::Tools::SeqStats) from BioPerl [22]. In order to enhance the query process, all pathogen sequences and annotation information for PGBrowser are stored in the database server instead of flat files.

The genome browser web interface of PGBrowser was developed based on the Generic Genome Browser (GBrowse) available at the Generic Software Components for Model Organism Databases (GMOD), a popular genome browser tool because of its portability, simple installation, convenient data input and easy integration with other software programs [23]. The GBrowse program has been used to display genome information about the bacterial pathogens *Brucella* spp. [2] and *Pseudomonas aeruginosa* [24]. PGBrowser modifies GBrowse and allows simultaneous query and analysis for any bacterial or viral gene across all 77 genomes of the 42 pathogens. For example, a query for *sodC* in PGBrowser results in 32 *sodC* hits from 32 genomes in 11 bacterial species, among which are four *Brucella sodC* genes from four *Brucella* genomes (Figure 2a). One can query any *Brucella* gene (for example, *sodC*) among the different *Brucella* genomes, analyze the gene sequences before and after a particular gene (Figure 2b), and obtain gene DNA, RNA, and protein sequences, and perform sequence analyses (for example, finding restriction enzyme digestion sites). As a feature inherited from GBrowse, PGBrowser also provides means for annotating restriction sites, finding short oligonucleotides, and downloading protein or DNA sequence files. PGBrowser can also be directly accessed from other PHIDIAS components such as Pacodom.

A detailed page of pathogen gene information has been developed to summarize integrative information about a specific pathogen gene, such as *sodC* in *B. melitensis* strain 16 M (Figure 3). It not only provides web links to various databases but also lists detailed protein annotation from authorized databases (for example, UniProt). Additionally, this page includes PHI specific information curated internally by the PHIDIAS curation team. A curator is also prompted to provide additional information using an online submission system. This page also provides DNA and protein sequences in FASTA format. The sequences can be directly linked to a customized BLAST search to find similar sequences from other pathogens. The references for curated PHI information are listed. A PubMed link is available for searching more related peer-reviewed articles. Figure 3 shows that Cu/Zn superoxide dismutase (SOD) encoded by the *B. abortus sodC* gene is required for *Brucella* protection from endogenous superoxide stress [25]. The *B. abortus sodC* mutant is attenuated in macrophages and mice [25]. Figure 3 also indicates that *Brucella* Cu/Zn SOD induces protective Th1 type immune responses and has been used for *Brucella* vaccine development [26]. For comparative purposes, one may examine *sodC* genes from other bacterial pathogens, such as *Bacillus anthracis*. Passalacqua *et al.* [27] recently showed that *B. anthracis* Cu/Zn SOD plays only a trivial role in protecting against endogenous superoxide stress. This indicates that the same gene may have different roles in microbial pathogenesis, suggesting that it is important to analyze pathogen genes individually, particularly in terms of the interactions between pathogens and hosts.

While PHIDIAS is pathogen-oriented and focuses on functional analysis of pathogen genes during PHI, host genome sequences may be requested for gene level PHI analyses. Since GBrowse-based human and mouse genome browsers are publicly available, PGBrowser contains a web interface that allows users to conveniently search the host genome sequence browsers by linking them to the websites.

Pacodom: pathogen protein conserved domains

The conserved domain data from completely sequenced pathogenic organisms provide valuable information for the identification of protein functions and for the study of PHI. Currently, the NCBI CDD database contains 12,589 position-specific score matrix (PSSM) models that are commonly used representations of motifs present in biological sequences. However, the PSSM models cover a broad range of organisms and, therefore, it is difficult to compare conserved domains from select priority pathogens. To circumvent this problem, a pathogen-specific protein conserved domains database module called Pacodom was developed. This program contains all possible conserved domains found in the 77 pathogen genomes of 42 pathogens. To build this system, a local reverse-position-specific (RPS) CDD library was constructed based on the CDD conserved domain data downloaded from NCBI [28]. The RPS BLAST program (downloaded from the NCBI toolkit distribution) [29] was run for each protein sequence against the RPS CDD library with an expectation value of 10^{-6}. The domain alignments obtained from the RPS BLAST search are used to calculate the PSSM. A Perl script was developed to store non-redundant PSSM models [30] in the Pacodom MySQL database module. Currently, the Pacodom database contains 7,919 PSSMs found in 151,787 protein sequences. This value comprises 76.4% of a total of 198,696 proteins from all genomes available in PhiDB.

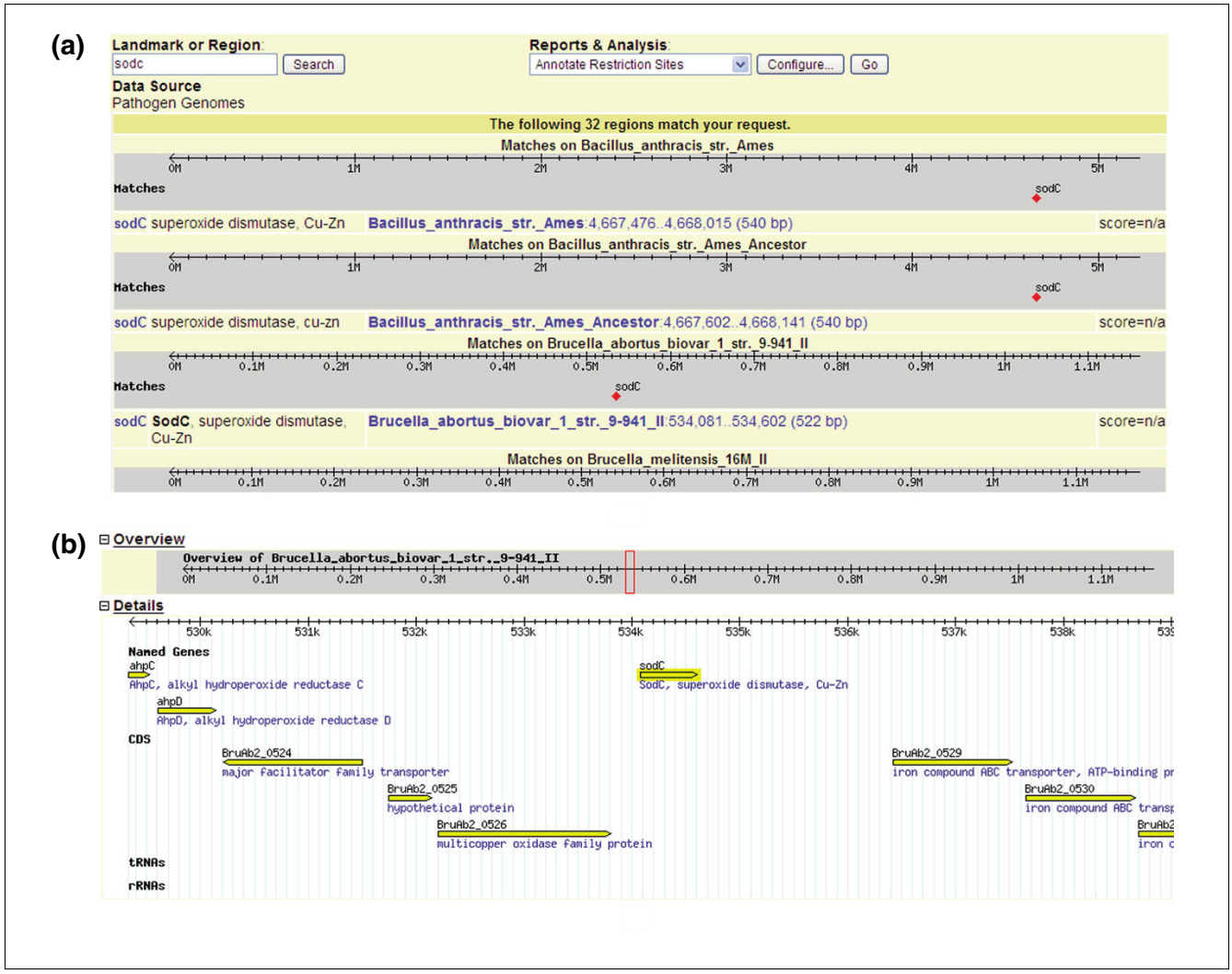
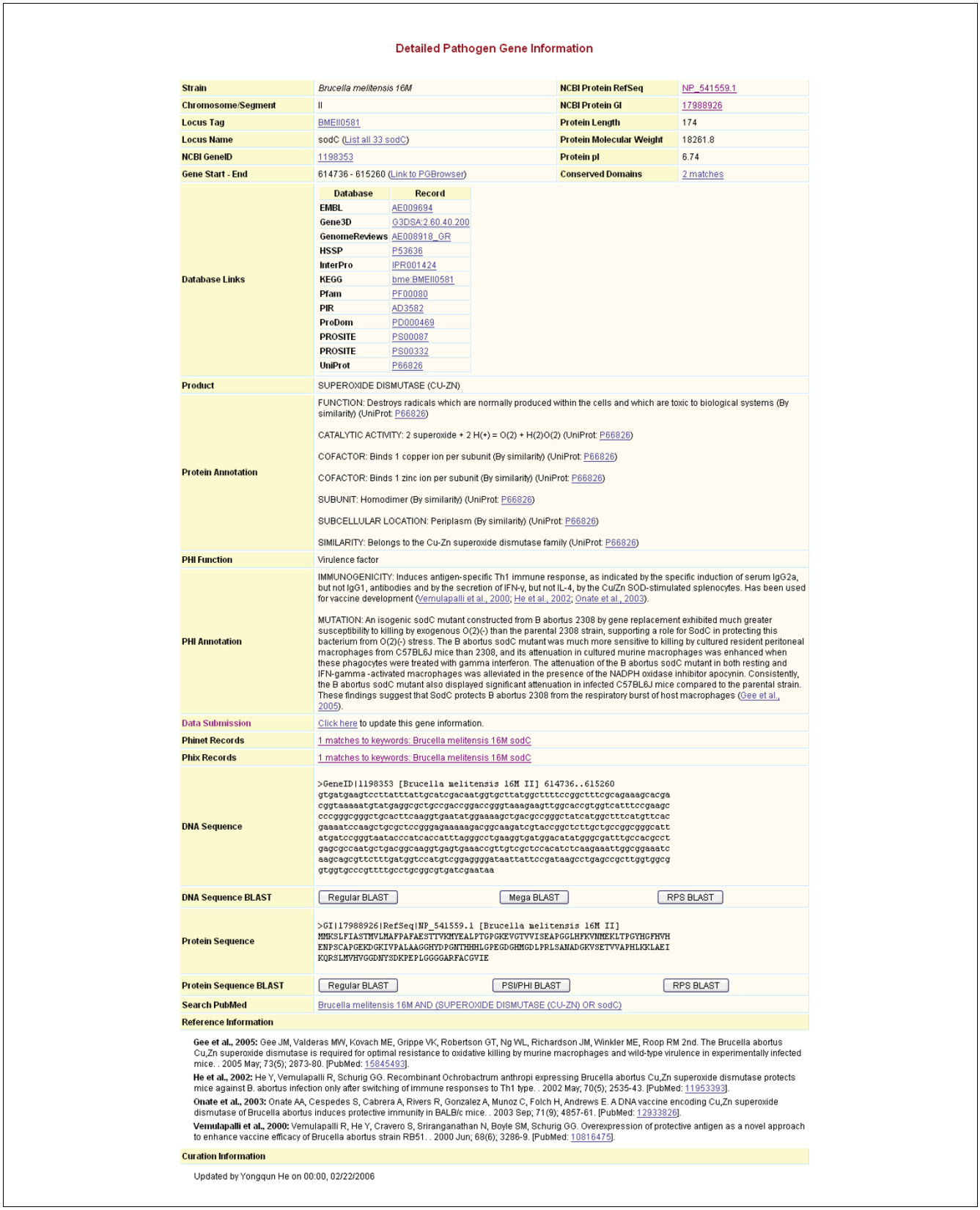


Figure 2
Comparison and analyses of *sodC* genes in the PGBrowser. Thirty two *sodC* genes are found in 32 genomes from 11 bacteria species **(a)**, including sodC from *B. abortus* strain 9-941 **(b)**.

The conserved domain data from completely sequenced pathogenic organisms provide valuable information for comparative analysis of functional roles of pathogen proteins and their involvement in the interactions between host and microbial organisms. For example, conserved domain data can be used to study phagocytosis, a process where host phygocytic cells (for example, macrophages) engulf pathogen cells (for example, *Brucella*). A search for 'phagocytosis' in Pacodom yields 14 domains; 13 domains do not match any protein from any PhiDB pathogen genome (Figure 4a). However, one domain, 'Nramp' (pfam01566), matches 42 pathogen proteins (Figure 4b). As summarized in the Pfam description of this domain (available in Pacodom), the natural resistance-associated macrophage protein (Nramp) family consists of Nramp1 and Nramp2 in human and mouse systems. Nramp1 plays an important role in phagocytosis and the macrophage activation pathway and regulates the interphagosomal replication of bacteria. Nramp2 is a transporter of multiple divalent cations (for example, Fe^{2+}, Mn^{2+} and Zn^{2+}) and is involved in a major transferrin-independent iron uptake system in mammals. The Pfam summary does not list any related microbial Nramp proteins. However, a Pacodom search shows Nramp is very common in the bacterial pathogens listed in PHIDIAS. Those 42 proteins containing the Nramp domain come from many bacterial species, such as *Brucella* spp., *Mycobacterium tuberculosis*, and *Salmonella enterica*. Nramp exists in all strains from these bacteria, whether the strain is pathogenic or non-pathogenic. In contrast, Nramp does not exist in the following species: *Campylobacter jejuni*, *Clostridium perfringens*, *Coxiella burnetii*, *Francisella tularensis*, and *Rickettsia prowazekii*. The Nramp domain has been investigated in depth in mycobacteria [31]. Since pathogenic mycobacteria survive within phagosomes, a nutrient-restricted environment, divalent cation transporters of the Nramp



Detailed Pathogen Gene Information

Strain	*Brucella melitensis 16M*	**NCBI Protein RefSeq**	NP_541559.1
Chromosome/Segment	II	**NCBI Protein GI**	17988926
Locus Tag	BMEII0581	**Protein Length**	174
Locus Name	sodC (List all 33 sodC)	**Protein Molecular Weight**	18261.8
NCBI GeneID	1198353	**Protein pI**	6.74
Gene Start - End	614736 - 615260 (Link to PGBrowser)	**Conserved Domains**	2 matches

Database Links

Database	Record
EMBL	AE009694
Gene3D	G3DSA:2.60.40.200
GenomeReviews	AE008918_GR
HSSP	P53636
InterPro	IPR001424
KEGG	bme:BMEII0581
Pfam	PF00080
PIR	AD3582
ProDom	PD000469
PROSITE	PS00087
PROSITE	PS00332
UniProt	P66826

Product

SUPEROXIDE DISMUTASE (CU-ZN)

Protein Annotation

FUNCTION: Destroys radicals which are normally produced within the cells and which are toxic to biological systems (By similarity) (UniProt: P66826)

CATALYTIC ACTIVITY: 2 superoxide + 2 H(+) = O(2) + H(2)O(2) (UniProt: P66826)

COFACTOR: Binds 1 copper ion per subunit (By similarity) (UniProt: P66826)

COFACTOR: Binds 1 zinc ion per subunit (By similarity) (UniProt: P66826)

SUBUNIT: Homodimer (By similarity) (UniProt: P66826)

SUBCELLULAR LOCATION: Periplasm (By similarity) (UniProt: P66826)

SIMILARITY: Belongs to the Cu-Zn superoxide dismutase family (UniProt: P66826)

PHI Function

Virulence factor

PHI Annotation

IMMUNOGENICITY: Induces antigen-specific Th1 immune response, as indicated by the specific induction of serum IgG2a, but not IgG1, antibodies and by the secretion of IFN-γ, but not IL-4, by the Cu/Zn SOD-stimulated splenocytes. Has been used for vaccine development (Vemulapalli et al., 2000; He et al., 2002; Onate et al., 2003).

MUTATION: An isogenic sodC mutant constructed from B abortus 2308 by gene replacement exhibited much greater susceptibility to killing by exogenous O(2)(-) than the parental 2308 strain, supporting a role for SodC in protecting this bacterium from O(2)(-) stress. The B abortus sodC mutant was much more sensitive to killing by cultured resident peritoneal macrophages from C57BL6J mice than 2308, and its attenuation in cultured murine macrophages was enhanced when these phagocytes were treated with gamma interferon. The attenuation of the B abortus sodC mutant in both resting and IFN-gamma -activated macrophages was alleviated in the presence of the NADPH oxidase inhibitor apocynin. Consistently, the B abortus sodC mutant also displayed significant attenuation in infected C57BL6J mice compared to the parental strain. These findings suggest that SodC protects B abortus 2308 from the respiratory burst of host macrophages (Gee et al., 2005).

Data Submission

Click here to update this gene information.

Phinet Records

1 matches to keywords: Brucella melitensis 16M sodC

Phix Records

1 matches to keywords: Brucella melitensis 16M sodC

DNA Sequence

```
>GeneID|1198353 [Brucella melitensis 16M II] 614736..615260
gtgatgaagtccttatttattgcatcgacaatggtgcttatggctttccggctttcgcagaaagcacga
cggtaaaaatgtatgaggcgctgccgaccggaccgggtaaagaagttggcaccgtggtcatttccgaagc
cccgggcgggctgcacttcaaggtgaatatggaaaagctgacgccgggctatcatggctttcatgttcac
gaaaatccaagctgcgctccgggagaaaaagacggcaagatcgtaccggctcttgctgccggcgggcatt
atgatccgggtaataccccatcaccatttagggcctgaaggtgatggacatatgggcgatttgccacgcct
gagcgccaatgctgacggcaaggtgagtgaaaccgttgtcgctccacatctcaagaaattggcggaaatc
aagcagcgttctttgatggtccatgtcggaggggataattattccgataagcctgagccgcttggtggcg
gtggtgcccgttttgcctgcggcgtgatcgaataa
```

DNA Sequence BLAST

Regular BLAST | Mega BLAST | RPS BLAST

Protein Sequence

```
>GI|17988926|RefSeq|NP_541559.1 [Brucella melitensis 16M II]
MMKSLFIASTMVLMAFPAFAESTTVKMYEALPTGPGKEVGTVVISEAPGGLHFKVNMEKLTPGYHGFHVH
ENPSCAPGEKDGKIVPALAAGGHYDPGNTHHHLGPEGDGHMGDLPRLSANADGKVSETVVAPHLKKLAEI
KQRSLMVHVGGDNYSDKPEPLGGGGARFACGVIE
```

Protein Sequence BLAST

Regular BLAST | PSI/PHI BLAST | RPS BLAST

Search PubMed

Brucella melitensis 16M AND (SUPEROXIDE DISMUTASE (CU-ZN) OR sodC)

Reference Information

Gee et al., 2005: Gee JM, Valderas MW, Kovach ME, Grippe VK, Robertson GT, Ng WL, Richardson JM, Winkler ME, Roop RM 2nd. The Brucella abortus Cu,Zn superoxide dismutase is required for optimal resistance to oxidative killing by murine macrophages and wild-type virulence in experimentally infected mice. . 2005 May; 73(5); 2873-80. [PubMed: 15845493].

He et al., 2002: He Y, Vemulapalli R, Schurig GG. Recombinant Ochrobactrum anthropi expressing Brucella abortus Cu,Zn superoxide dismutase protects mice against B. abortus infection only after switching of immune responses to Th1 type. . 2002 May; 70(5); 2535-43. [PubMed: 11953393].

Onate et al., 2003: Onate AA, Cespedes S, Cabrera A, Rivers R, Gonzalez A, Munoz C, Folch H, Andrews E. A DNA vaccine encoding Cu,Zn superoxide dismutase of Brucella abortus induces protective immunity in BALB/c mice. . 2003 Sep; 71(9); 4857-61. [PubMed: 12933826].

Vemulapalli et al., 2000: Vemulapalli R, He Y, Cravero S, Sriranganathan N, Boyle SM, Schurig GG. Overexpression of protective antigen as a novel approach to enhance vaccine efficacy of Brucella abortus strain RB51. . 2000 Jun; 68(6); 3286-9. [PubMed: 10816475].

Curation Information

Updated by Yongqun He on 00:00, 02/22/2006

Figure 3
Integrative pathogen gene information in PHIDIAS.

family in phagosomes and mycobacteria may compete for metals that are crucial for bacterial survival [31]. However, inactivation of mycobacterial Nramp, called Mramp, does not affect virulence in mice, suggesting a sufficient redundancy in the cation acquisition systems [32]. A more recent report [33] demonstrated that the *Salmonella enterica* serovar typhimurium (*S. typhimurium*) requires both of the divalent cation transport systems, MntH (Nramp1 homolog) and SitABCD (putative ABC iron and/or manganese transporter), for full virulence in congenic Nramp1-expressing mice. These results suggest that bacterial Nramp is required for pathogenesis in *S. typhimurium* and probably other bacteria by synchronizing with other redundant cation transport system(s) to compete for divalent cations with host cells. The role of *Brucella* Nramp in pathogenesis remains unclear and deserves further analysis. This example demonstrates how Pacadom can be used to find valuable information and form testable hypotheses by comparative analysis of conserved domains.

It is noted that the Nramp domain (pfam01566), while found in a list of pathogens in Pacodom, is also found in many bacterial species that are not pathogens. Therefore, it may be important for investigators to cross reference PHIDIAS search results against databases that contain both pathogen and non-pathogen species. Since Pacodom includes conserved domains from both pathogenic strains and non-pathogenic strains of the same microbial species, it can be used to find domains shown in pathogenic but not in non-pathogenic strains. For example, a query of 'bacteriophage' in Pacodom results in many conserved domains being found, such as Phage_Mu_Gp45 (pfam06890) and Phage_Mu_F (pfam04233), which exist in pathogenic *E. coli* O157:H7 strain Sakai but not in the benign K12 strain. Such domains have previously been reported as required for pathogenesis [34].

BLAST searches

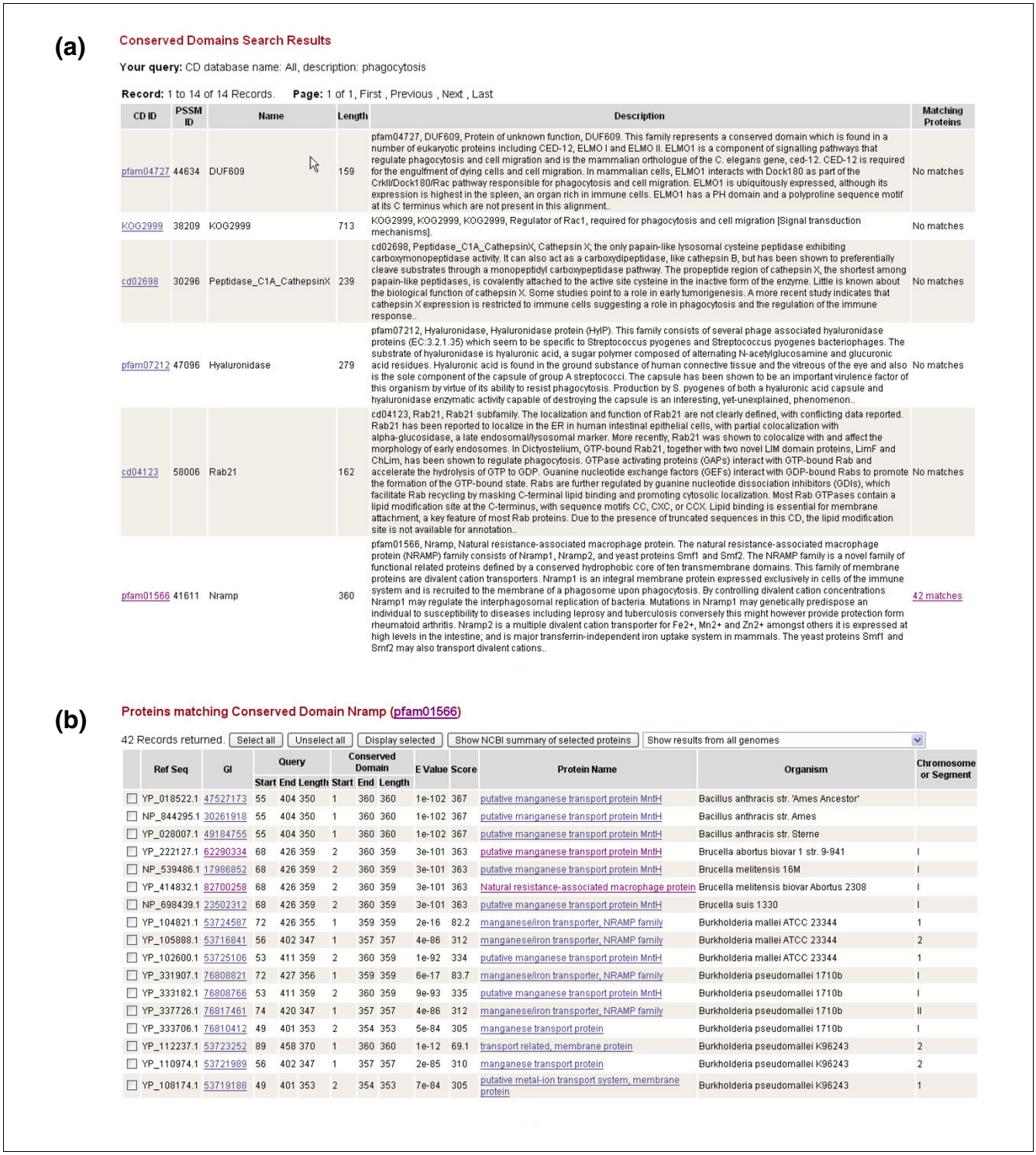
Gene or protein sequences among different pathogen genomes can be analyzed by different BLAST search approaches. PHIDIAS BLAST uses the latest web server version of BLAST obtained from NCBI [35]. It includes regular BLAST services (blastn, blastp, blastx, tblastn, tblastx), PSI/PHI BLAST, Mega BLAST, RPS BLAST, and BLAST 2 sequences. The nucleotide and protein BLAST libraries contain sequences from all the 77 genomes of the 42 pathogens (Table 1). The 7,919 PSSMs available in Pacodom are combined to form a customized RPS BLAST library specifically used for the RPS BLAST program. The sequence libraries are updated periodically to reflect newly curated annotations and the addition of new genomes.

The approaches used with BLAST greatly help comparative studies for all the genes available in PhiDB. However, some gene annotations from certain genomes are not satisfactory. Based on sequence similarity, these are readily detected with BLAST. The PHIDIAS BLAST methods can also be used to find a group of pathogen genes using a seeding DNA or protein sequence. For example, a PHIDIAS blastp search for the protein sequence of human Nramp1 (also known as SLC11A1, RefSeq#: NP_000569) yields 65 hits from 77 pathogen genomes, most of which are attributable to a single putative manganese transport protein (MntH, which belongs to the Nramp family) found in different pathogens, including four *Brucella* strains. A blastp search using human Nramp2 (also known as SLC11A2, RefSeq#: NP_000608) as input yields similar hits. The BLAST search results are consistent with the analysis of conserved domains as described in the section on Pacodom above.

Phinfo: curated pathogen-host interaction general information

The Phinfo database module stores pathogen and PHI information curated from the biomedical literature and other curated databases. A major source of Phinfo data are PIML documents available from Virginia Bioinformatics Institute (VBI) [7]. A Java program was developed to extract PIML documents from the ToolBus/PathPort PIML XML database via the PathInfo web service [36]. An Extensible Stylesheet Language for Transformations (XSLT) script was developed to parse the PIML documents into a text-based SQL script. This in turn was used to insert the parsed data into a predesigned MySQL database system. Phinfo also integrates data manually curated by the PHIDIAS curation team from PubMed literature and other databases such as KEGG [9]. Phinfo links to the Hazards in Animal Research Database (HazARD). This database was developed internally at the University of Michigan [8]. Pathobiology and management of laboratory animals administered USA NIAID/CDC priority pathogens are subjects of the HazARD database and can be searched with Phinfo [8]. Currently, Phinfo includes information for 36 pathogens and corresponding PHI information supported by 2,894 references.

Phinfo provides an integrative web interface for user-friendly querying and display of curated pathogen and PHI information. Two query programs are available in Phinfo: Keyword Search and Topic Search. The Keyword Search program allows queries for specific pathogen and PHI information. Such information is displayed with the searched keywords highlighted in color. The Topic Search program searches for one or many of 47 topics listed in the hierarchical structure (Figure 5). Compared to the native PIML XML database [7], the relational Phinfo database system provides secure storage, efficient querying, and database extendibility (that is, the ability to add new data categories). In addition, Phinfo provides links to public databases (for example, NCBI taxonomy, NCBI Gene database, and PubMed). Phinfo is also integrated with other PHIDIAS components. For example, Phinfo of *Brucella* spp. indicates that a PCR assay based on the *B. abortus* gene *wboA* (forward primer: TTAAGCGCTGATGCCATTTCCTTCAC, reverse primer: GCCAACCAACCCAAATGCTCACAA) has been used to



(a)

Conserved Domains Search Results

Your query: CD database name: All, description: phagocytosis

Record: 1 to 14 of 14 Records. **Page:** 1 of 1, First , Previous , Next , Last

CD ID	PSSM ID	Name	Length	Description	Matching Proteins
pfam04727	44634	DUF609	159	pfam04727, DUF609, Protein of unknown function, DUF609. This family represents a conserved domain which is found in a number of eukaryotic proteins including CED-12, ELMO I and ELMO II. ELMO1 is a component of signalling pathways that regulate phagocytosis and cell migration and is the mammalian orthologue of the C. elegans gene, ced-12. CED-12 is required for the engulfment of dying cells and cell migration. In mammalian cells, ELMO1 interacts with Dock180 as part of the Crkll/Dock180/Rac pathway responsible for phagocytosis and cell migration. ELMO1 is ubiquitously expressed, although its expression is highest in the spleen, an organ rich in immune cells. ELMO1 has a PH domain and a polyproline sequence motif at its C terminus which are not present in this alignment..	No matches
KOG2999	38209	KOG2999	713	KOG2999, KOG2999, KOG2999, Regulator of Rac1, required for phagocytosis and cell migration [Signal transduction mechanisms].	No matches
cd02698	30296	Peptidase_C1A_CathepsinX	239	cd02698, Peptidase_C1A_CathepsinX, Cathepsin X; the only papain-like lysosomal cysteine peptidase exhibiting carboxymonopeptidase activity. It can also act as a carboxydipeptidase, like cathepsin B, but has been shown to preferentially cleave substrates through a monopeptidyl carboxypeptidase pathway. The propeptide region of cathepsin X, the shortest among papain-like peptidases, is covalently attached to the active site cysteine in the inactive form of the enzyme. Little is known about the biological function of cathepsin X. Some studies point to a role in early tumorigenesis. A more recent study indicates that cathepsin X expression is restricted to immune cells suggesting a role in phagocytosis and the regulation of the immune response..	No matches
pfam07212	47096	Hyaluronidase	279	pfam07212, Hyaluronidase, Hyaluronidase protein (HylP). This family consists of several phage associated hyaluronidase proteins (EC:3.2.1.35) which seem to be specific to Streptococcus pyogenes and Streptococcus pyogenes bacteriophages. The substrate of hyaluronidase is hyaluronic acid, a sugar polymer composed of alternating N-acetylglucosamine and glucuronic acid residues. Hyaluronic acid is found in the ground substance of human connective tissue and the vitreous of the eye and also is the sole component of the capsule of group A streptococci. The capsule has been shown to be an important virulence factor of this organism by virtue of its ability to resist phagocytosis. Production by S. pyogenes of both a hyaluronic acid capsule and hyaluronidase enzymatic activity capable of destroying the capsule is an interesting, yet-unexplained, phenomenon..	No matches
cd04123	58006	Rab21	162	cd04123, Rab21, Rab21 subfamily. The localization and function of Rab21 are not clearly defined, with conflicting data reported. Rab21 has been reported to localize in the ER in human intestinal epithelial cells, with partial colocalization with alpha-glucosidase, a late endosomal/lysosomal marker. More recently, Rab21 was shown to colocalize with and affect the morphology of early endosomes. In Dictyostelium, GTP-bound Rab21, together with two novel LIM domain proteins, LimF and ChLim, has been shown to regulate phagocytosis. GTPase activating proteins (GAPs) interact with GTP-bound Rab and accelerate the hydrolysis of GTP to GDP. Guanine nucleotide exchange factors (GEFs) interact with GDP-bound Rabs to promote the formation of the GTP-bound state. Rabs are further regulated by guanine nucleotide dissociation inhibitors (GDIs), which facilitate Rab recycling by masking C-terminal lipid binding and promoting cytosolic localization. Most Rab GTPases contain a lipid modification site at the C-terminus, with sequence motifs CC, CXC, or CCX. Lipid binding is essential for membrane attachment, a key feature of most Rab proteins. Due to the presence of truncated sequences in this CD, the lipid modification site is not available for annotation..	No matches
pfam01566	41611	Nramp	360	pfam01566, Nramp, Natural resistance-associated macrophage protein. The natural resistance-associated macrophage protein (NRAMP) family consists of Nramp1, Nramp2, and yeast proteins Smf1 and Smf2. The NRAMP family is a novel family of functional related proteins defined by a conserved hydrophobic core of ten transmembrane domains. This family of membrane proteins are divalent cation transporters. Nramp1 is an integral membrane protein expressed exclusively in cells of the immune system and is recruited to the membrane of a phagosome upon phagocytosis. By controlling divalent cation concentrations Nramp1 may regulate the interphagosomal replication of bacteria. Mutations in Nramp1 may genetically predispose an individual to susceptibility to diseases including leprosy and tuberculosis conversely this might however provide protection form rheumatoid arthritis. Nramp2 is a multiple divalent cation transporter for Fe2+, Mn2+ and Zn2+ amongst others it is expressed at high levels in the intestine; and is major transferrin-independent iron uptake system in mammals. The yeast proteins Smf1 and Smf2 may also transport divalent cations..	42 matches

(b)

Proteins matching Conserved Domain Nramp (pfam01566)

42 Records returned. Select all | Unselect all | Display selected | Show NCBI summary of selected proteins | Show results from all genomes

Ref Seq	GI	Query			Conserved Domain			E Value	Score	Protein Name	Organism	Chromosome or Segment
		Start	End	Length	Start	End	Length					
YP_018522.1	47527173	55	404	350	1	360	360	1e-102	367	putative manganese transport protein MntH	Bacillus anthracis str. 'Ames Ancestor'	
NP_844295.1	30261918	55	404	350	1	360	360	1e-102	367	putative manganese transport protein MntH	Bacillus anthracis str. Ames	
YP_028007.1	49184755	55	404	350	1	360	360	1e-102	367	putative manganese transport protein MntH	Bacillus anthracis str. Sterne	
YP_222127.1	62290334	68	426	359	2	360	359	3e-101	363	putative manganese transport protein MntH	Brucella abortus biovar 1 str. 9-941	I
NP_539486.1	17986852	68	426	359	2	360	359	3e-101	363	putative manganese transport protein MntH	Brucella melitensis 16M	I
YP_414832.1	82700258	68	426	359	2	360	359	3e-101	363	Natural resistance-associated macrophage protein	Brucella melitensis biovar Abortus 2308	I
NP_698439.1	23502312	68	426	359	2	360	359	3e-101	363	putative manganese transport protein MntH	Brucella suis 1330	I
YP_104821.1	53724587	72	426	355	1	359	359	2e-16	82.2	manganese/iron transporter, NRAMP family	Burkholderia mallei ATCC 23344	1
YP_105888.1	53716841	56	402	347	1	357	357	4e-86	312	manganese/iron transporter, NRAMP family	Burkholderia mallei ATCC 23344	2
YP_102600.1	53725106	53	411	359	2	360	359	1e-92	334	putative manganese transport protein MntH	Burkholderia mallei ATCC 23344	1
YP_331907.1	76808821	72	427	356	1	359	359	6e-17	83.7	manganese/iron transporter, NRAMP family	Burkholderia pseudomallei 1710b	I
YP_333182.1	76808766	53	411	359	2	360	359	9e-93	335	putative manganese transport protein MntH	Burkholderia pseudomallei 1710b	I
YP_337726.1	76817461	74	420	347	1	357	357	4e-86	312	manganese/iron transporter, NRAMP family	Burkholderia pseudomallei 1710b	II
YP_333706.1	76810412	49	401	353	2	354	353	5e-84	305	manganese transport protein	Burkholderia pseudomallei 1710b	I
YP_112237.1	53723252	89	458	370	1	360	360	1e-12	69.1	transport related, membrane protein	Burkholderia pseudomallei K96243	2
YP_110974.1	53721989	56	402	347	1	357	357	2e-85	310	manganese transport protein	Burkholderia pseudomallei K96243	2
YP_108174.1	53719188	49	401	353	2	354	353	7e-84	305	putative metal-ion transport system, membrane protein	Burkholderia pseudomallei K96243	1

Figure 4
Example of Pacodom applications. **(a)** Pacodom search of 'phagocytosis'. **(b)** There are 42 Nramp protein matches from 42 pathogen genomes of 15 microbial species available in Pacodom.

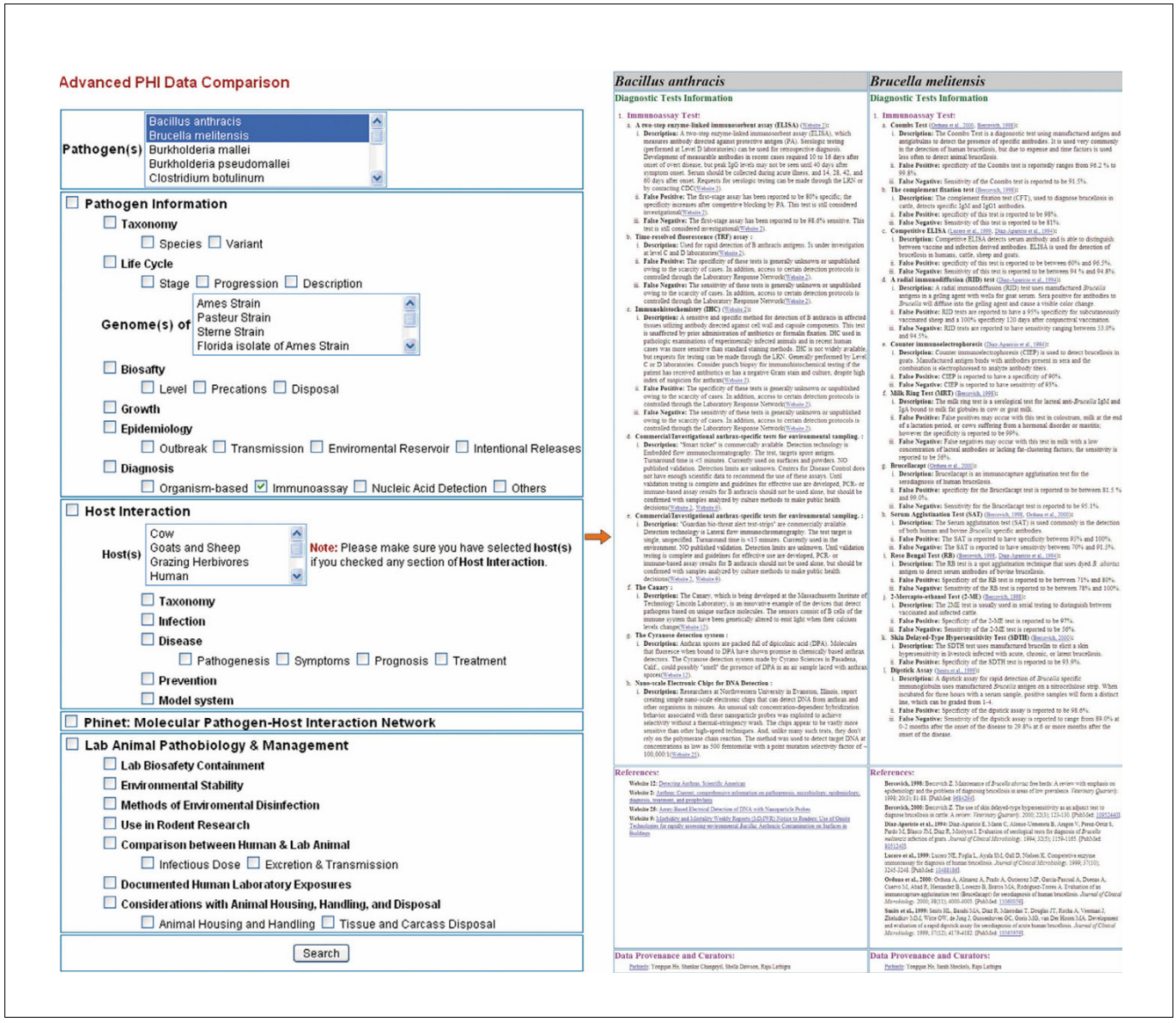


Figure 5
PhiDB Topic Search. The PhiDB Topic Search web interface is shown on the left and a comparison of immunoassays for diagnosis of *B. melitensis* and *B. anthracis* is shown on the right.

differentiate *B. abortus* vaccine strain RB51 from other *Brucella* strains. Either of the primer sequences can be linked directly by clicking to local nucleotide BLAST analysis. Genes found from local BLAST searches are also linked to the PHIDIAS gene table (Figure 3). The *wboA* genes from four *Brucella* genomes are always the first four hits. Other microbial genes (for example, from *Vibrio* and *Yersinia*) are also found, indicating a possible cross-reaction during PCR assays and/or functional similarities among these genes.

Phigen: pathogen-host interaction genes

The interactions between pathogen and host genes have been extensively studied in the post-genomic era [37]. However, most databases of genes and proteins focus on sequence annotation and function in a single cell species. Phigen focuses on functional annotation of pathogen genes and their interaction with host genes during the process of pathogen-host reactions. The main source of the PHI-related gene annotation comes from literature curation and data integration. The information about genes and/or proteins required for virulence, able to induce protective immune responses in hosts, or used for diagnosis, has been annotated and stored in the Phigen system. Phigen consists of two parts, pathogen gene search and manual curation submission.

Every pathogen gene may be involved in an interaction between the pathogen and its host. The pathogen gene search interface of Phigen allows users to search for any pathogen

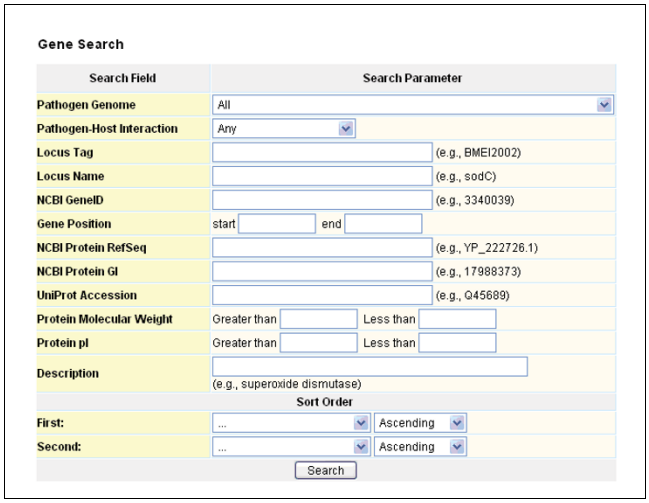


Figure 6
Gene search web interface in Phigen.

genes from the 77 genomes of the 42 pathogens available in PhiDB (Table 1). The Phigen search has a function for simple Boolean-powered keyword searches and an advanced topic search (Figure 6). The advanced topic search allows searching for PHI-specific information and generic features, including chromosomes and chromosomal position, RefSeq identifier, GenBank accession number, locus tag and name, molecular weight, pI, and description. Searched results can also be sorted in ascending or descending order. Molecular weight and pI data obtained in each search may be used to aid the interpretation of two-dimensional mass spectrometry data for proteomics analyses.

Phigen provides an efficient online submission system for submitting of data for curation of pathogen genes, especially their roles in PHI. The information is fully referenced from peer-reviewed publications, with direct links to PubMed paper abstracts and full texts for additional details. Submitted information is critically reviewed and verified by reviewers prior to acceptance. Currently, Phigen has manually curated and stored more than 400 genes from 42 pathogens. Instead of altering records from other public databases, the curation is currently focusing on adding PHI-related information, such as host immune responses, gene mutations and resultant pathogenic changes in the host. In addition to integrated gene information, the PHI-specific information assists researchers in surveying, comparing, and studying gene-specific PHI mechanisms.

Phinet: pathogen-host interaction network curation, data exchange, and visualization

PHI has the ability to reveal complicated networks between pathogen and host molecules. Phinet is targeted at analyzing molecular networks responsible for PHI. Phinet data are stored in PhiDB and are derived from the MINetML XML database extracted through the web service, other curated databases (for example, KEGG), and manual annotation based on literature curation. Similar to that implemented in Phinfo, a Java program was developed to extract MINetML documents from the ToolBus/PathPort MINetML XML database via the MINet web service [38]. An XSLT script was further developed to parse the MINetML documents into a text-based SQL script, which is used to insert the parsed data into a pre-designed MySQL database system. Data from the KEGG pathway database are manually curated and added to Phinet. Phinet also includes a web-based data submission system that permits internal or external curators to submit PHI-related network data. The Phinet data submission follows a similar curation policy as described for Phigen online submission above. If conflicts exist for data from different sources, those records with the strongest reference support are selected, or in some circumstances, conflicting data were included with well-documented references. Currently, Phinet includes PHI network information for 21 pathogens.

A Graphviz-based visualization software program has been developed internally to dynamically display all the biological interactions in Phinet (Figure 7). The visualization program effectively displays all pathway data for each pathogen available in Phinet. The user can select to view information about a biological object or the interaction between biological objects (Figure 7).

Data exchange among different pathway databases is critical for data sharing and integration. BioPAX is a community-supported data exchange format for biological pathway data [14]. Current BioPAX Level 2 covers metabolic pathways, molecular interactions and protein post-translational modifications. Compared to the model representation format SBML, BioPAX focuses on molecule and interaction classification schemes and database cross-referencing for pathway components. PHI networks involve complex signaling pathways and gene regulatory networks that are similar to BioPAX, although they are not supported in their entirety by the current BioPAX version. A program was developed to transform Phinet data to the closest BioPAX OWL format using the current BioPAX Level 2 format. These BioPAX documents can be used to communicate with other biological pathway databases and, additionally, provide input files for other software programs.

Phix: pathogen-host interaction gene expression

Gene expression data for pathogens and/or hosts during PHIs comprise important data for analysis of pathogen pathogenesis and host defense mechanisms. The NCBI GEO [15] and EBI ArrayExpress [39] are the two biggest repositories that store publicly available microarray and proteomics data, many of which relate to PHI. The Phix database stores all gene expression experiment records for the targeted 42 pathogens and their infected hosts from the GEO and ArrayExpress databases. Since new gene expression experiments are frequently submitted to these databases, a Linux

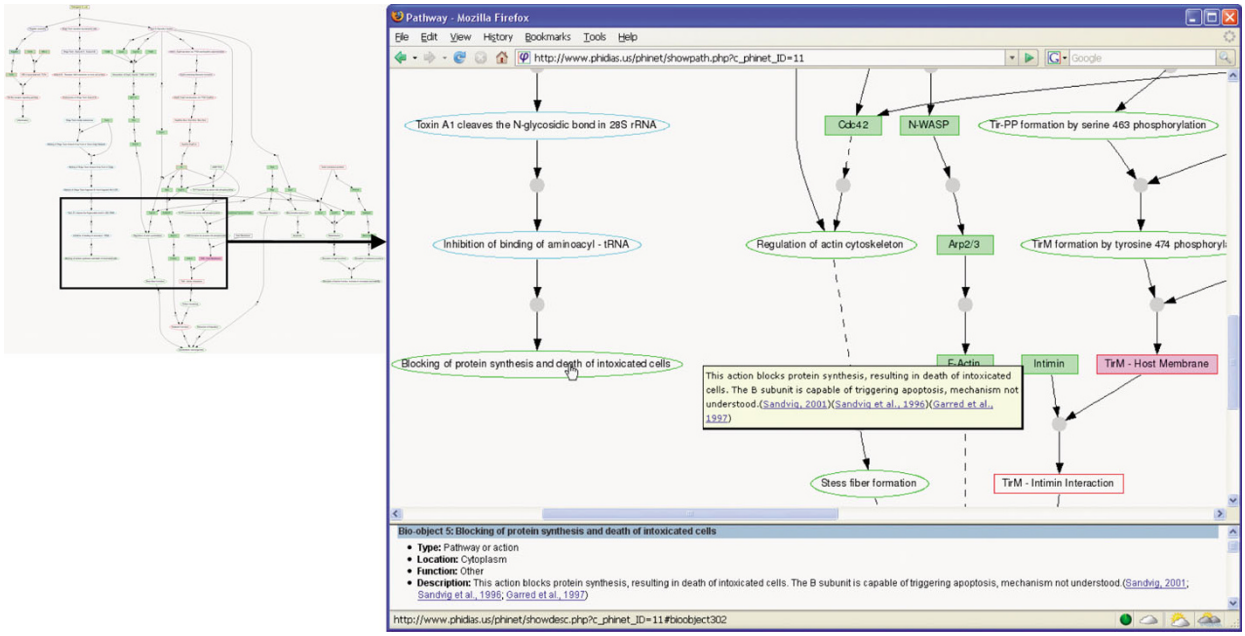


Figure 7
Visualization of an *E. coli* pathogenesis network in Phinet. A click on each node provides detailed information about a biological object in the bottom frame. When a mouse cursor moves over a node, a brief description of the biological object will appear. An interaction between biological objects is represented by a centered gray ball and arrows between nodes. Once the centered gray ball is clicked, details about the specific interaction appear in the bottom frame. Subcellular locations of biological objects are differentiated by the node border colors. The biological object types (for example, protein or gene) are represented by a combination of the node background colors and shapes. The program also displays different interactions, such as inhibition (solid T sign), activation (solid arrow), and indirect effects (dashed line).

cron job [40] was developed to check daily for any new information; if found, the new data are added to the database. The Phix module currently stores 187 GEO records and 79 ArrayExpress records. The Phix gene expression search program provides a one-step system for users to query PHI gene expression experimental data. For example, a query of 'macrophage' in Phix leads to 13 search hits representing various experimental studies involving pathogen-infected macrophages. Each hit links to detailed information in GEO or ArrayExpress. These results are particularly useful for comparing different pathogen-macrophage interaction systems. Finally, Phix also includes a gene profile search engine for query and comparison of expression profiles of specific genes from one, or all, of the pathogen genomes selected from the GEO and ArrayExpress databases. In contrast to the general GEO and ArrayExpress gene profile search engines, this program is specifically targeted to pathogen and PHI studies.

To improve further integration of different PHIDIAS components, the PHIDIAS web site contains a keyword search engine that simultaneously allows searching for information from all PHIDIAS components. All results are sorted based on the components and displayed in one page for convenient data analysis (data not shown).

Discussion

A deeper understanding of PHI is required for effectively combating infectious diseases. To efficiently analyze the ever-increasing amount of PHI data in the post-genomics era, PHIDIAS was developed. This program permits integration of PHI related data from genome sequences, the biomedical literature, curated databases, and gene expression experiments. PHIDIAS covers 42 microbial and viral pathogens of high priority for public heath and security. The gene and protein sequences from each genome are available for browsing and analysis using PGBrowser and customized BLAST searches. The conserved domains are analyzed and stored in Pacadom. PHI data extracted from existing databases, or internally manually curated, are stored in Phinfo (general PHI information), Phigen (PHI genes) and Phinet (PHI networks). PHI-related gene expression experiment records and profiles from public GEO and ArrayExpress repositories can be directly searched in Phix. The PHIDIAS components are interconnected (Figure 1). Scenarios have been used in this report to show that PHIDIAS greatly helps *Brucella* research by allowing users to search and analyze integrative *Brucella* data derived from different sources and compare these data with those from other pathogens.

Similar PHI-related biological programs exist. PHI-base is a web-accessible database devoted to the identification and presentation of information on fungal and oomycete pathogenicity genes and their host interactions [41]. PathoPlant deals with plant-pathogen interactions, signal transduction reactions, and microarray gene expression data from *Arabidopsis thaliana* subjected to pathogen infection and elicitor treatment [42]. In contrast to PHI-base and PathoPlant, which target the interactive relationships between pathogens and hosts, PHIDIAS includes a list of other bacterial, viral and parasitic pathogens and their interactions with hosts. Similar to PHIDIAS, PHI-base and PathoPlant contain manually curated information supported by strong experimental evidence (gene disruption experiments) and literature references. Each system allows interlinking of gene information with external data sources. However, PHIDIAS integrates more data sources for a broader scope of data integration and analysis. PHIDIAS also provides on-line submission systems for curators to submit annotated data for genes as well as genetic interactions and pathways.

Many biological systems allow systematic genome comparison. MicrobesOnline is a publicly available suite of web-based comparative genomic tools designed to facilitate multispecies comparison among prokaryotes [43]. The database PRODORIC systematically organizes information about the prokaryotic gene expression of multiple prokaryotic species, and integrates this information into regulatory networks [44]. As does PHIDIAS, these systems contain many comparative analysis and visualization tools. However, while MicrobesOnline and PRODORIC target more general prokaryotic species, PHIDIAS focuses on pathogenic bacteria as well as viral and parasitic pathogens important for biodefense and/or human health. PHIDIAS also emphasizes interactions between pathogens and hosts, which MicrobesOnline and PRODORIC currently lack. PHIDIAS also contains manually curated data for functional annotation of genes and genetic networks in pathogen genomes.

Eight Bioinformatics Resource Centers (BRCs), sponsored by the USA NIAID, provide web-based resources for organisms that are considered potential agents of biowarfare or bioterrorism or cause emerging or re-emerging diseases [45]. Each BRC is targeted to maintain and annotate genomes from a selected list of pathogens. Each BRC contains a web site to display the data and analyses for these pathogens. BRC Central [46] serves as a repository linking these eight BRCs. Many of the pathogens contained in the BRCs are also found in PHIDIAS. However, PHIDIAS also targets non-biodefense pathogens (for example, HIV) not included in the BRCs. Additionally, PHIDIAS includes not only data analysis and search functions found in the BRC resources, but also provides tighter integration of various data types. Finally, PHI and literature data curation are emphasized in PHIDIAS but not in the BRCs.

PHIDIAS is unique in that it integrates existing knowledge about a broad range of human or zoonotic priority pathogens, and focuses on efficient searching, visualization, comparison, and analysis of pathogen genes and their interactions with their hosts using genome sequences, manually curated literature data, and gene expression data from public resources. PHIDIAS utilizes online data submission systems for efficient data curation, making integrative PHI data more comprehensive. All the PHIDIAS components are scalable, and more pathogens and PHI systems may be added to the system. Due to inclusion of an ever increasing number of pathogens in PHIDIAS and in view of the dramatically increasing amount of literature information, it will be an ongoing challenge to curate all the significant genes and keep the PHI-related information in PhiDB current. Therefore, one of our future directions will be to explore ontology-based natural language processing and statistical methods for efficient literature acquisition and curation. In this regard, we have now developed a literature mining and curation system (Limix). This system has been used efficiently for literature mining and curation for four *Brucella* genomes [2]. Systematic curation and incorporation of *Brucella*-specific mutation and genetic interaction information has allowed a comprehensive investigation of *Brucella* pathogenesis [2]. Limix is currently being expanded to annotate literature for other pathogens and PHI systems. Finally, future plans for expanding PHIDIAS include development of a web-based database and an analysis pipeline that permit storage, processing, and modeling of PHI-related gene expression data. This approach will allow researchers to address scientific PHI questions with the ultimate goal of successfully fighting infectious diseases.

Acknowledgements

We thank the authors of published data in various programs (for example, RefSeq, CDD, Pfam, PubMed, PathInfo, MINet, HazARD, KEGG, and so on) for making them available to the public. We also acknowledge the public availability of many open-source programs (for example, GBrowse and NCBI BLAST) that have allowed the integration and extension into PHIDIAS. The critical review and editing of this manuscript by Drs L Colby and GW Jourdian from the University of Michigan Medical School is gratefully acknowledged.

References

1. Becker K, Hu Y, Biller-Andorno N: **Infectious diseases - a global challenge.** *Int J Med Microbiol* 2006, **296:**179-185.
2. Xiang Z, Zheng W, He Y: **BBP: *Brucella* genome annotation with literature mining and curation.** *BMC Bioinformatics* 2006, **7:**347.
3. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32:**D138-141.
4. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Res* 2006, **34:**D257-260.
5. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al.*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
6. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, *et al.*: **CDD: a conserved domain database for interactive domain**

family analysis. *Nucleic Acids Res* 2007, **35:**D237-240.
7. He Y, Vines RR, Wattam AR, Abramochkin GV, Dickerman AW, Eckart JD, Sobral BW: **PIML: the Pathogen Information Markup Language.** *Bioinformatics* 2005, **21:**116-121.
8. He Y, Rush HG, Liepman RS, Xiang Z, Colby LA: **Pathobiology and management of laboratory rodents administered CDC Category A agents.** *Comparative Med* 2007, **57:**18-32.
9. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32:**D277-280.
10. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database.** *Nucleic Acids Res* 2002, **30:**56-58.
11. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-442.
12. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31:**248-250.
13. **The Molecular Interaction Network Markup Language(MINetML)** [http://pathport.vbi.vt.edu/xml/molecules/molecules.dtd]
14. Stromback L, Lambrix P: **Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX.** *Bioinformatics* 2005, **21:**4401-4407.
15. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles - database and tools.** *Nucleic Acids Res* 2005, **33:**D562-566.
16. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, *et al.*: **ArrayExpress - a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2005, **33:**D553-555.
17. Roop RM 2nd, Bellaire BH, Valderas MW, Cardelli JA: **Adaptation of the brucellae to their intracellular niche.** *Mol Microbiol* 2004, **52:**621-630.
18. DelVecchio VG, Kapatral V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharyya A, Lykidis A, *et al.*: **The genome sequence of the facultative intracellular pathogen *Brucella melitensis*.** *Proc Natl Acad Sci USA* 2002, **99:**443-448.
19. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Umayam L, Brinkac LM, Beanan MJ, *et al.*: **The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts.** *Proc Natl Acad Sci USA* 2002, **99:**13148-13153.
20. Halling SM, Peterson-Burch BD, Bricker BJ, Zuerner RL, Qing Z, Li LL, Kapur V, Alt DP, Olsen SC: **Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*.** *J Bacteriol* 2005, **187:**2715-2726.
21. Chain PS, Comerci DJ, Tolmasky ME, Larimer FW, Malfatti SA, Vergez LM, Aguero F, Land ML, Ugalde RA, Garcia E: **Whole-genome analyses of speciation events in pathogenic brucellae.** *Infect Immun* 2005, **73:**8353-8361.
22. **BioPerl** [http://www.bioperl.org]
23. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.
24. Winsor GL, Lo R, Sui SJ, Ung KS, Huang S, Cheng D, Ching WK, Hancock RE, Brinkman FS: ***Pseudomonas aeruginosa* Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation.** *Nucleic Acids Res* 2005, **33:**D338-343.
25. Gee JM, Valderas MW, Kovach ME, Grippe VK, Robertson GT, Ng WL, Richardson JM, Winkler ME, Roop RM 2nd: **The *Brucella abortus* Cu, Zn superoxide dismutase is required for optimal resistance to oxidative killing by murine macrophages and wild-type virulence in experimentally infected mice.** *Infect Immun* 2005, **73:**2873-2880.
26. He Y, Vemulapalli R, Schurig GG: **Recombinant *Ochrobactrum anthropi* expressing *Brucella abortus* Cu, Zn superoxide dismutase protects mice against *B. abortus* infection only after switching of immune responses to Th1 type.** *Infect Immun* 2002, **70:**2535-2543.
27. Passalacqua KD, Bergman NH, Herring-Palmer A, Hanna P: **The superoxide dismutases of *Bacillus anthracis* do not cooperatively protect against endogenous superoxide stress.** *J Bacteriol* 2006, **188:**3837-3848.
28. **NCBI CDD Download** [ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/cdd.tar.gz]
29. **NCBI Toolkit Download** [ftp://ftp.ncbi.nlm.nih.gov/toolbox]
30. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure.** *Nucleic Acids Res* 2002, **30:**281-283.
31. Agranoff D, Monahan IM, Mangan JA, Butcher PD, Krishna S: ***Mycobacterium tuberculosis* expresses a novel pH-dependent divalent cation transporter belonging to the Nramp family.** *J Exp Med* 1999, **190:**717-724.
32. Boechat N, Lagier-Roger B, Petit S, Bordat Y, Rauzier J, Hance AJ, Gicquel B, Reyrat JM: **Disruption of the gene homologous to mammalian Nramp1 in *Mycobacterium tuberculosis* does not affect virulence in mice.** *Infect Immun* 2002, **70:**4124-4131.
33. Zaharik ML, Cullen VL, Fung AM, Libby SJ, Kujat Choy SL, Coburn B, Kehres DG, Maguire ME, Fang FC, Finlay BB: **The *Salmonella enterica* serovar typhimurium divalent cation transport systems MntH and SitABCD are essential for virulence in an Nramp1G169 murine typhoid model.** *Infect Immun* 2004, **72:**5522-5525.
34. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, *et al.*: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8:**11-22.
35. **NCBI BLAST Download** [http://www.ncbi.nih.gov/BLAST/download.shtml]
36. **PathInfo Web Service** [http://staff.vbi.vt.edu/pathport/services/wsdls/pathinfo.wsdl]
37. Forst CV: **Host-pathogen systems biology.** *Drug Discov Today* 2006, **11:**220-227.
38. **MINet Web Service** [http://www.vbi.vt.edu/~pathport/services/wsdls/pathway.wsdl]
39. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, *et al.*: **ArrayExpress - a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2003, **31:**68-71.
40. Petersen R: *Linux: The Complete Reference* 4th edition. Emeryville, CA: McGraw-Hill Osborne Media; 2000.
41. Winnenburg R, Baldwin TK, Urban M, Rawlings C, Kohler J, Hammond-Kosack KE: **PHI-base: a new database for pathogen host interactions.** *Nucleic Acids Res* 2006, **34:**D459-464.
42. Bulow L, Schindler M, Hehl R: **PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses.** *Nucleic Acids Res* 2007, **35:**D841-845.
43. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP: **The MicrobesOnline Web site for comparative genomics.** *Genome Res* 2005, **15:**1015-1022.
44. Munch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D: **PRODORIC: prokaryotic database of gene regulation.** *Nucleic Acids Res* 2003, **31:**266-269.
45. **NIAID Bioinformatics Resource Centers for Biodefense and Emerging or Re-emerging Infectious Diseases: an Overview** [http://www.niaid.nih.gov/dmid/genomes/brc/]
46. **BRC Central** [http://www.brc-central.org]