D754–D764 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 22 October 2009*
doi:10.1093/nar/gkp832

GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database

Lynn M. Schriml[1,*], Cesar Arze[1], Suvarna Nadendla[1], Anu Ganapathy[1], Victor Felix[1], Anup Mahurkar[1], Katherine Phillippy[2], Aaron Gussman[1], Sam Angiuoli[1], Elodie Ghedin[3], Owen White[1] and Neil Hall[4]

[1]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, [2]National Center for Biotechnology Information, Bethesda, MD, [3]University of Pittsburgh School of Medicine, Division of Infectious Diseases, Pittsburgh, PA, USA and [4]University of Liverpool, School of Biological Sciences, UK

Received August 12, 2009; Revised September 16, 2009; Accepted September 18, 2009

ABSTRACT

The Gemina system (http://gemina.igs.umaryland.edu) identifies, standardizes and integrates the outbreak metadata for the breadth of NIAID category A–C viral and bacterial pathogens, thereby providing an investigative and surveillance tool describing the Who [Host], What [Disease, Symptom], When [Date], Where [Location] and How [Pathogen, Environmental Source, Reservoir, Transmission Method] for each pathogen. The Gemina database will provide a greater understanding of the interactions of viral and bacterial pathogens with their hosts and infectious diseases through in-depth literature text-mining, integrated outbreak metadata, outbreak surveillance tools, extensive ontology development, metadata curation and representative genomic sequence identification and standards development. The Gemina web interface provides metadata selection and retrieval of a pathogen's Infection Systems (Pathogen, Host, Disease, Transmission Method and Anatomy) and Incidents (Location and Date) along with a hosts Age and Gender. The Gemina system provides an integrated investigative and geospatial surveillance system connecting pathogens, pathogen products and disease anchored on the taxonomic ID of the pathogen and host to identify the breadth of hosts and diseases known for these pathogens, to identify the extent of outbreak locations, and to identify unique genomic regions with the DNA Signature Insignia Detection Tool.

INTRODUCTION

GeMInA, Genomic Metadata for Infectious Agents (http://gemina.igs.umaryland.edu), is an open source web-based pathogen-centric tool designed to provide an integrated investigative and geospatial surveillance system connecting pathogens, pathogen products and disease metadata anchored on the taxonomic ID of the pathogen and host.

The Gemina project has developed a rigorous system of ontological standards that enable the tracking of pathogen related metadata, based on rigorous literature data mining. The Gemina system enables biomedical, bioforensics, and biodefense users to ask the questions of **Who** and **What** are these pathogens and hosts being affected, **When** and **Where** are the incidents occurring, and **What** diseases and symptoms are being reported in the current or month, year or decade.

The Gemina system links unique genomic representations of each pathogen with ontology regularized metadata for the associated epidemiological information. Gemina provides a metadata selection query interface to guide identification of the NIAID category A–C viral and bacterial pathogens (http://www3.niaid.nih.gov/biodefense/PDF/cat.pdf), connecting the pathogen metadata to a selection tool to calculate unique regions within the genomes of these pathogens identifying DNA signatures using the Insignia detection tool (1) (http://insignia.cbcb.umd.edu/).

The Gemina database and query web interface provide a greater understanding of the interactions of viral and bacterial pathogens, their hosts and infectious diseases through in-depth literature text-mining, integrated outbreak metadata, outbreak surveillance tools, extensive ontology development, metadata curation and

*To whom correspondence should be addressed. Tel: +1 410 706 6776; Fax: +1 410 706 6756; Email: lschriml@som.umaryland.edu

representative genomic sequence identification and standards development.

The Gemina system enables users to explore the diversity of outbreak data for each NIAID category A–C pathogen reported in literature including the published CDC's Morbidity and Mortality Weekly Reports (http://www.cdc.gov/mmwr/distrnds.html) that have been regularized through a set of mature community-adopted ontologies, to identify the breadth of hosts and diseases known for these pathogens, where these pathogens have been reported to occur in the world and to link to the Insignia detection tool to calculate the unique regions within the genomes of these pathogens.

Outbreak surveillance reporting sites, such as BioCaster (2), HealthMap (3), ProMed-mail (4) and the World Health Organization (WHO) Disease Outbreak News (http://www.who.int/csr/don/en/) identify outbreaks in real-time contributed by member institutions, news reports, and personal accounts. RSS feeds and online outbreak reports are rich resources of automated data feeds. Online reporting site data includes a mixture of suspected and documented outbreak cases and contains unfiltered data that requires additional quality control and data cleanup. Mining of the online data sets through the filter of controlled vocabularies has provided a rich resource of additional metadata for published outbreak cases. Gemina and the real-time reporting sites provide complimentary resources for outbreak surveillance.

DATABASE GENERATION AND CONTENT

Database development strategy

The Gemina Chado database (5) developed from early collaborations with Generic Model Organism Database (GMOD) system (http://www.gmod.org). Initial data structures were developed in collaboration with the Microbial Rosetta Stone (6,7). In 2005–2006 Gemina developed an open source schema, compiled data, developed standard vocabularies, and deployed the public Gemina site.

Infection module and Chado database

The concepts of the Gemina system are maintained in an infection module that has been developed for the Chado database. The Infection Module integrates with existing Chado modules allowing it to act as an extension of existing data. It links directly to the General, Controlled Vocabulary, and Publication modules and can enhance genome annotations in the Sequence Module. The data model places no direct limits on which controlled vocabularies may be associated with which fields; this is handled via agreed upon usage conventions. This allows a field to be constrained to terms from multiple controlled vocabularies or ontologies, e.g., a valid Source term could come from either the NCBI taxonomy or an environmental source controlled vocabulary. The usage conventions employed within Gemina direct the usage of vocabularies for each data type.

Data loading process

The Gemina system utilizes an automated loading pipeline which processes the infection and incident flat files of curated pathogen metadata and populates the various tables that act as the backend of the Gemina website interface. The pipeline can be segmented into the following sections: database initialization, infection and incident flat file loading, and database post-processing. The first step is to initialize the Gemina database by populating the database tables with the necessary ontologies/controlled vocabularies. This initial loading step ensures an essential high level of quality control as all subsequent data files will be loaded and checked against the ontologies from this first step. This flat file loading process is executed sequentially due to the connectivity of the data. First all infection files are loaded one at a time followed by the incident files. This ordered loading procedure must be followed because an infection system [defined as a pathogen, host and disease] plus a transmission [defined as a transmission method and an anatomy] must be present in order for an incident to load successfully. This additional level of data integrity ensures that there is always a direct correspondence of infection systems to incidents. The loading scripts have additional quality control checks that identify mismatches between loading files and ontology files to identify terms that have yet to be added to ontologies, misspelled terms or changes in taxonomic names. These 'to be reviewed' terms are moved to a 'unmapped' review queue at the time of loading for manual review. Following the data load several post-processing steps occur to tie the various vocabularies together (e.g. Reservoir) and allow for the robust search capabilities. The geocoding described in the GAZ and GEO Surveillance section is performed during this section of the load process. The automation of this loading process carries several benefits such as quick turn-around time when new data is curated and an easily repeatable (24–48 hr process) that can be enhanced by scheduling execution of the pipeline using CRON or any similar tools.

Database content

In the 23 June 2009 release of Gemina, the database contained 367 bacteria, 21 plant and bacterial toxins, and 10 991 viral strains including influenza A subtypes and strains, and influenza B and C strains and their associated chain of infection metadata. The Gemina's viral data set includes strains and metadata for each of the influenza A subtypes and strains and influenza B and C stains including the 2009 swine-origin H1N1 variant with updates posted monthly. Gemina's Influenza data set is curated from the literature, NCBI's Influenza virus resource's FTP (ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/) files and metadata (David Spiro and Elodie Ghedin, pers. comm.) of the Influenza Virus Genome Sequencing Project inclusive of the 2009 Swine Flu H1N1 outbreak. The current Gemina data set is reported on the Gemina Database Statistics (http://gemina.igs.umaryland.edu/cgi-bin/gemina/Stats.cgi) site which enumerates the list of pathogen strains grouped

by species, their category A–C status, number of Infections Systems (14 192) and Incidents (26 038) in Gemina. Trends in pathogen metadata are explored on the Metadata Summary Statistics pages. Links to each pathogens page, such as the *Bacillus anthracis* page: http://gemina.igs.umaryland.edu/cgi-bin/gemina/MetadataStats.cgi?pathogen=Bacillus%20anthracis, provide the total number of incidents, unique hosts (of the possible 793 hosts in Gemina), or unique locations that have been identified for each strain and allow a filter to explore the data for that strain. Comparisons between closely related strains may indicate emergence, novel hosts or geographical spread.

Infection systems and incidents data model

The Gemina system has utilized the five key chain of infection components—pathogen, host, disease, transmission method, and anatomy—to create a set of 'Infection Systems' that uniquely identify a pathogen. Specifically, a pathogen's set of Infection Systems define the pathogen's infectious fingerprint of **Who** or **What** was infected and **How** it was infected.

An infection system describes a pathogen, its hosts and reservoirs, and the methods by which it is transmitted between them. Both the anatomical component serving as the pathogen's point of entry into the host and the primary disease associated with infection as well as the pathogen's reservoir, defined as the habitat in which a pathogen naturally lives, grows and multiplies, are recorded. Each component of the infection data model is mapped to its respective controlled vocabularies and ontologies described in the *Ontologies data standard development section* (below) when terms are identified during curation. Additionally, loading scripts for the Gemina Chado database (5) confirm the mapping of infection and incident terms to the ontologies and vocabularies.

Outbreaks of pathogens, as reported in literature, are recorded as individual incidents in the Gemina data model, and are linked to a single Infection System (Pathogen/Host/Disease) and Transmission event [Transmission Method (type) and Anatomy (portal)]. Incidents are characterized by a geographic location and date (collection date). Incidents may also include Host Age and Host Gender metadata. Multiple incidents may be linked to a single infection system in the Gemina data model.

Metadata collection and curation

The Gemina system presents the diversity of *pathogen metadata types* (e.g. Pathogens, Toxins, Hosts, Reservoirs, Environmental Sources, Diseases, Transmission Methods, Symptoms, Location, Gender, Age and Date) collected from primary literature curation. Each pathogen's specific epidemiological chain of infection information regarding the diversity of known hosts, types of transmission, reservoir, and diseases are curated from authoritative sites such as the CDC, WHO and Medline in order to gain an initial understanding of the biology of each pathogen. Relevant PubMed articles are retrieved through multiple keyword searches (e.g. transmission, disease, outbreak), in combination with the PubMed query of the appropriate taxonomy name: 'pathogen taxonomy name [AND] epidemiology'. Relevant abstracts and papers are then reviewed, PDF URLs and PubMed IDs recorded. Extensive literature data mining for each pathogen has enabled the Gemina system to identify the diversity of diseases, hosts, reservoirs, transmission methods and portals of entry into the host (anatomy) specific to each pathogen. Epidemiologically, these chains of infection are specific to each pathogen.

Data processing and regularization

The tab-delimited Infection and Incident flat were devised for the collection and exchange of epidemiological metadata and to populate the Infection Module and Incident Module database tables in Gemina. As metadata is collected for each new pathogen, corresponding rows are added to the Infection and Incident Flat Files. A separate flat file is maintained on the Gemina SourceForge site (e.g. http://gemina.svn.sourceforge.net/viewvc/gemina/trunk/Gemina/data/incidents/) for infections and their related incidents for bacteria, viruses, influenza, toxins and reservoirs.

Ontology data standards development

Ontologies and controlled vocabularies have been developed to support the Gemina system for each metadata type by the Gemina team or in consortium and collaborative efforts. These vocabularies provide a metadata standard backbone for the Gemina system. The ontologies are utilized to ensure quality control of the data and for data filtering when data is loaded into the database. Synonymous terms identified during curation are stored in the ontologies and linked to the primary ontology term. These linked terms can be queried in the web interface connecting disparate data across PubMed references. Programmatic use of ontologies and controlled vocabularies has ensured that all data in Gemina is regularized at the time of curation and checked at the time of loading against the vocabularies. The rigorous ontology mapping and filtering methodology has allowed the Gemina system to track synonymous data over time, to semantically calculate relationships through the ontologies, to utilize these relationships to gain new knowledge regarding the transmission of pathogenic organisms, and to unify disparate data identified through the curation of primary literature. Additionally, for the query interface and Gemina user community, inclusion of synonymous terms in the ontologies bridges disparate data in unlinked sources and PubMed articles enabling it to be linked and retrieved, therefore enriching the utility of the ontologies and the data sets.

In order to regularize the chain of infection components (disease, symptom, anatomy and transmission method), we conducted a review of the available set of domain ontologies at the onset of the Gemina project (disease, cell, anatomy) in the Open Biomedical Ontologies (OBO) Foundry (8) (http://www.obofoundry.org) and

vocabularies (e.g. Medical Subject Headings (MeSH vocabulary (http://www.nlm.nih.gov/cgi/mesh/2007/MB_cgi), International Classification of Diseases (ICD)). We identified pertinent vocabularies which could be utilized for the project and began to develop the Gemina ontologies. The rules for building these ontologies followed those defined by the OBO Foundry: the ontology consists of terms (nodes) that are linked by common relations (is_a, part_of) (9); each term has a unique identifier with the ontology-specific syntax; collaborative ontology development; methods for community input; access to current and previous version; the ontology defines a distinct domain; and the ontologies are developed through collaborative efforts.

The current set of available ontologies and controlled vocabularies Gemina are listed on the Gemina Statistics site with the number of terms contained in each vocabulary. The term counts for each vocabulary are updated for each Gemina data load and presented on the Gemina Database Statistics (http://gemina.igs.umaryland.edu/cgi-bin/gemina/Stats.cgi) site. The ongoing ontology development has grown in the Gemina project into a collaborative effort of co-development and community-adopted ontologies (e.g., DO, EnvO, GAZ, Transmission Method ontology), with contribution of terms by collaborators (e.g. Symptom ontology). Here we describe the origins of the key vocabularies utilized by the Gemina project and Gemina's role in the development of some of our collaborative ontology projects.

Diseases

The communicable disease branch of the Disease ontology (v 2.1) (DO) (http://diseaseontology.sourceforge.net/) was initially utilized to create the gemina_disease ontology along with the addition of terms from MeSH and ICD-10. Gemina has enriched the content of infectious diseases in the DO in the role of curator of this branch as Gemina's role in the project grew from data contributor to collaborator. As the Disease Ontology is currently undergoing active updates and reorganization, the gemina_disease ontology terms and definitions will be incorporated into DO. When this work is complete Gemina will incorporate DO.

Anatomy

The Gemina data model includes human and animal hosts and their associated portals of entry. In order to map the diversity of anatomy terms identified from literature for the NIAID agents we developed an anatomical application ontology that includes animal, human and cell line terms compiled from the Anatomy branch of MeSH and select terms from the cell types Cell Ontology.

Symptoms

The Gemina symptom ontology was developed from an initial list of MeSH Sign and Symptoms terms and organized into a structured vocabulary of body system symptoms. Additional symptoms terms were identified and added from ICD-10s symptoms block R00-R99 (Chapter XVlll) (http://www.who.int/classifications/icd/en/). The symptom ontology was designed around the guiding concept of a symptom being: 'A perceived change in function, sensation or appearance reported by a patient indicative of a disease'. The symptom ontology was submitted to the OBO Foundry in July 2008 and continues to undergo active development to incorporate Basic Formal Ontology structure. In the Gemina system, symptoms are linked to the Pathogen–Host–Disease Infection System, describing the symptoms as they are associated in the referenced literature. Under this model, a new infection system is created when an outbreak occurs and where symptoms are reported which differ from the general set of symptoms associated with an Infection system conditions in order to adequately capture the novel infection system represented by the new outbreak situation.

Transmission methods

Direct and indirect methods of pathogen transmission between hosts, hosts and reservoirs were characterized in the Transmission Method ontology devised for the Gemina project from the epidemiological explanation of transmission methods in the CDC's Epidemiology Course Book (10). The Transmission method ontology has been further developed and was submitted to the OBO Foundry as the Pathogen Transmission ontology in February 2008.

Environmental sources and locations

A pathogenic agent may exist in the environment prior to being transmitted into a host. This environmental location, whether it is a reservoir or an intermediary non-living source of an agent, is an important epidemiological data type to capture, track and understand. Geographic location is a key data type linking place names and their associated outbreak data curated from thousands of references producing a worldwide representation of location information. Gemina's initial geographic location ontology and environmental habitat ontology were collaboratively joined into the Environmental Ontology consortium projects: EnvO ontology of environmental types and Gazetteer (GAZ) controlled vocabulary of geographic locations sponsored by the Genome Standard Consortium in 2007. The Gemina project utilizes both vocabularies and contributes to both projects as co-developer of the vocabularies working with Michael Ashburner on GAZ and Norman Morrison on EnvO, as well as the other members of the EnvO consortium.

GAZ and GEO surveillance

Each incident location in Gemina has been converted to its corresponding GIS coordinates, in decimal format. GIS coordinates have been identified using geocoding services provided by Geonames and Google Maps. These coordinates are curated to ensure correct place name to GIS mapping and then loaded into the Gemina database to allow easy access for mapping of incident data when needed. The raw coordinates pulled down from the geocoding process are stored in a local text file (e.g. /trunk/data/GIS/) which allows for incremental loading to

be performed when adding incidents that contain new locations. This data is presented in two formats from the Incidents Results page, as either single outbreak locations in Google Maps or as all outbreak locations of a pathogen in Google Earth. Both output types include additional Incident metadata such as: incident date, location place name, strain name and related resource links [NCBI's taxonomy database and NCBI's nucleotide database (GenBank)]. As part of the Google Earth surveillance tool, all incidents for a pathogen are viewable along the timeline for the available incidents. This timeline may be set in motion to see the progression of incidents. The Google Earth display enables exploration of the geospatial outbreak data of a strain with a global and time perspective.

Reservoirs

Reservoirs are defined, in the Gemina data model, as the habitat in which an infectious agent normally lives, grows, and multiplies. Infection systems are created for pathogens to document the epidemiological conditions involving a pathogen and its reservoir and to foster the characterization of the transmission of pathogens between reservoirs and hosts. Reservoir data has been identified from the CDC and literature, recorded in a reservoir data file (reservoir infection flat file) and curated into three distinct groups: human, environmental and animal reservoirs. *Human reservoirs* included anatomical structures or substances involved in person-to-person transmission without intermediaries. *Environmental reservoirs* included all non-living sources such as soil or water. *Animal reservoirs* included animals (non-human) which were the normal habitat for the pathogenic agent. Gemina maintains a curatorial ontology of these reservoir types that maps the NCBI taxonomic name, NCBI taxonomic ID and common name. The ontology file (reservoirs_gemina.obo) is archived on the Gemina SourceForge site in the Future directory as it is maintained for curatorial purposes. The public Reservoirs controlled vocabulary is auto-generated from the reservoirs infection flat file in order to properly recapitulate the structure of the source ontologies represented in the reservoir data types. The reservoir controlled vocabulary consists of Environments (EnvO terms), Organisms (NCBI taxonomy terms—Reservoirs), and environment matter and food (EnvO terms). Gemina programmatically extracts the reservoir data from the reservoir infection flat file and utilizes the knowledge of the ontological structure from the NCBI taxonomy and EnvO ontologies to properly structure the corresponding nodes to create semantically correct branches for the Reservoir Vocabulary.

Toxins: pathogen products as agents of disease

The Gemina toxin controlled vocabulary was developed in order to capture the relationship between the toxin product, such as Ricin toxin or *Clostridium botulinum* BoNT, and the disease caused by the bacterial, fungal and plant toxins. The Toxins vocabulary is under active development in collaboration with groups from MITRE and LANL as we co-develop a Virulence Factor Ontology (http://virulencewiki.igs.umaryland.edu/wiki/) that will incorporate toxins as well as other virulence factors.

NCBI taxonomy identifier data standards

The NCBI taxonomy controlled vocabulary was utilized for the pathogen, host, source and reservoir organisms in Gemina as a standard organism identifier. Utilizing the NCBI taxonomy we have created pathogen and host controlled vocabularies and have included an organisms-specific branch in the reservoir vocabulary. The Pathogen Taxonomy Controlled Vocabulary was initially built from the core set Microbial Rosetta Stone microbial and viral pathogens, expanded to include all pathogens to the level of strain for each NIAID viral and bacterial pathogen and completed by building the parental nodes based on the NCBI structure. The host and reservoir organism trees were built starting with the set of hosts or reservoirs identified from Gemina curation efforts in a bottom up approach of identifying the nodes of the NCBI Taxonomy tree that should be represented in the taxonomy tree.

Incident metadata

Inclusion of incident data in the Gemina database required determination of specifications for the addition of Gender, Age and Date (collection date) data types and mapping of input data to these specifications. **Gender** is represented on the website as Male or Female and in the data files as M or F. **Age** is represented, in years, in two formats as either a single year, e.g. 35, or in a range (e.g. 2–25). **Date** can be included in multiple day, month and year formats. Date may be queried as a single year (e.g. 1999), a single month (e.g. 1 December 2004), a single day (e.g. 5 May 1942), a range of years (1917–1999), a range of months (October 1989 to November 2008), or a range of dates (e.g. 14 January 1911 to 14 December 1951).

Genomic sequence standards development

Providing targeted DNA selection for the Gemina project has involved creating the methodology to identify the single best representative genomic sequence for each pathogen from the breadth of available sequences in GenBank's divisions and to periodically provide these sequences as a data set to the Insignia project for their pipeline updates. As the pace of microbial sequencing continues to accelerate, the public repositories continue to see a mixture of complete, incomplete and in-progress genome projects. With an understanding of this situation several years ago, Gemina devised a tiered selection strategy to identify a unique genomic sequence for each microbial pathogen that is the most 'complete' genomic sequence accessions for each pathogen to enable targeted DNA detection. To this end, we have developed an automated pipeline that mines NCBI genome sequences in a tiered system. Selecting the most recent, complete genomic sequence from the first tier, if it is present for each genome a representative sequence is identified for each pathogen in Gemina, if available. The selection process progresses down the tiers until a genomic

sequence is selected. After a genomic sequence is identified the taxonomic ID for the pathogen is removed from the query set. The NCBI genomic sequence tiers, from NCBI's ftp site (ftp://ncbi.nlm.nih.gov/) with the latest (July 2009) number of NCBI taxonomy IDs identified per tier are:

- Tier 1: RefSeq: an annotated non-redundant set of genomic sequences (3542 taxonomy IDs, e.g. NC_012564).
- Tier 2: GenBank (Complete Genome) (5914 taxonomy IDs, e.g. FJ560944).
- Genomes (Complete Genomes): GenBank genome complete sequences that have not undergone additional NCBI processing or analysis.
- Tier 3 Whole Genome Shotgun (WGS) (966 taxonomy IDs, e.g.ACOZ01000001).
- Tier 4 Genome Project sequences (incomplete genomes) (2840 taxonomy IDs, CY040535).

For genomes where only WGS data is available we would include all of the WGS data for the taxon id. Sequence selection for the bacterial genomes included both the chromosome and plasmid sequence, where available. The tiered selection process for the WGS and Complete Genomes (Tier 2, Tier 3) involves a rules based approach involving keywords and accessions to identify complete genomes and accessions belonging to the WGS division. Gemina then implemented a method to select these sequences and convert the GenBank flat files (GFF) into Bioinformatic Sequence Markup Language (BSML) sequence files. These files are loaded into the Gemina sequence database and annotated with their corresponding gene annotations. Through the Gemina database each genomic sequence is linked to a pathogen taxonomic ID in the Insignia database. The genomic sequences are uploaded to the Insignia project where they are utilized for their DNA pipeline. These sequences are updated periodically and are available on the Insignia project ftp site (ftp://ftp.cbcb.umd.edu/pub/software/insignia/).

Sequence selection and feature attributes

Sequences are identified as belonging to one type of genome category: (1) a whole genome project (**wgp**), (2) a complete genome (**cg**), (3) a **closed** genome which is considered a subset of a complete genome or (4) whole genome project (**wgp**) (Table 1).

Sequence type is also accessed from the record name including records of the type chloroplast, plasmid, mitochondrial (or containing name mitochondria). Chloroplast and mitochondrial sequences were excluded. Each sequence was also tagged with the source of the sequence: GenBank, Refseq and genomes. Additionally, data attributes of mol_type (ss, single strand; ds, double strand DNA or RNA), pos (positive or negative strand RNA), experimental (experimentally characterized), and complete (complete cds) were identified and used for annotation of the sequences. Experimental flag is determined based on the following rule: for GenBank data the presence of feature qualifier 'experiment' in CDS feature block means experimental. The source sequences referred to the ftp site utilized to retrieve the data: (i) GenBank data sets were obtained from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/genbank; (ii) RefSeq data sets were obtained from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/refseq; and GenBank genomes data sets were retrieved from NCBI's ftp site at ftp://ftp.ncbi.nih.gov/genomes and ftp://ftp.ncbi.nih.gov/genbank/genomes.

Table 1. Accession and keyword criteria

Genome category	Determined by	Keywords or accession patterns
Whole Genome Project (WGP)	Accession prefix	GenBank: AE, CP, CY EMBL: AL, BX, CR, CT, CU DDBJ: AP RefSeq: AC, NC, NT, NW WGS: AAAA–AZZZ, BAAA–BZZZ, CAAA–CZZZ, NZHTGS: AK, AC, DP
Complete Genomes (cg)	NCBIs genome database	Accessions identified by 'complete' [properties] search Accessions from NCBI's/genomes/IDs files Keyword 'complete genome' or 'complete chromosome' in definition line
	NCBIs closed genomes	Accession Patterns: GenBank: AE, CP, CY EMBL: AL, BX, CR, CT, CU DDBJ: AP RefSeq: AC, NC

Genome category, keywords and accession patterns for selection of WGS and Complete Genome sequences for Gemina.

Sequence load script development

Sequences were processed in a series of steps to convert the GenBank flat files. These included converting the GenBank sequence files to BSML formatted files (genbank2bsml conversion), loading the files into a Chado sequence database (bsml2chado script with chado to bsml component), then populating the cm_proteins table (chado_mart ergatis script) in the Chado database to annotate the genome features onto the sequences.

DATABASE ACCESS

Query interface

The Gemina database query web interface (http://gemina.igs.umaryland.edu) (Figure 1) provides a suite of metadata types as a query selection tool to explore the diversity of infectious pathogens, selects these pathogens to identify their associated DNA via the Insignia Selection pipeline or examines the pathogen's current outbreaks and outbreak history along the geospatial axis in Google Earth and Google Maps. Pathogens can be selected and explored based on their curated infection metadata (host, source, reservoir, disease, transmission method, or symptoms) or the incident metadata (location, date, gender or age). The Gemina system provides a standardized set of data types of infectious pathogen information. This standardization provides the research community with reliable, quality controlled data prepared in a format amenable for data

Figure 1. Gemina Database Query Web Interface. The Gemina web interface provides users with the option to select one or more vocabulary terms to build a query to submit against the Gemina database. This *Bacillus anthracis* example provides a demonstration of terms selected from two of Gemina's vocabularies. Synonyms, common names, taxonomy IDs and multiple terms (such as Anthrax 2012) may be used in the search box. A search of the Diseases vocabulary and the term Anthrax will initiate a broader search and will retrieve data associated with the Anthrax term and terms that are children of Anthrax as they are types of Anthrax.

exchange, comparisons and analysis. Therefore, data collected from diverse studies, over many years are comparable. The Gemina system is unique in that it contains the breadth of literature reported outbreak data for each pathogen standardized into a uniform format and set of vocabularies. Gemina's web interface provides a gateway to explore data in user specified ways.

User guidance on the query interface is provided with: a Quick Start guide on the Gemina Home Page, an online Tutorial, term selection and tree navigation guides provided at the bottom of the Tree View section of the Gemina Database Search Page, and a Help page on the Gemina Search Report page. The Tutorial link is available at the top of each Gemina page.

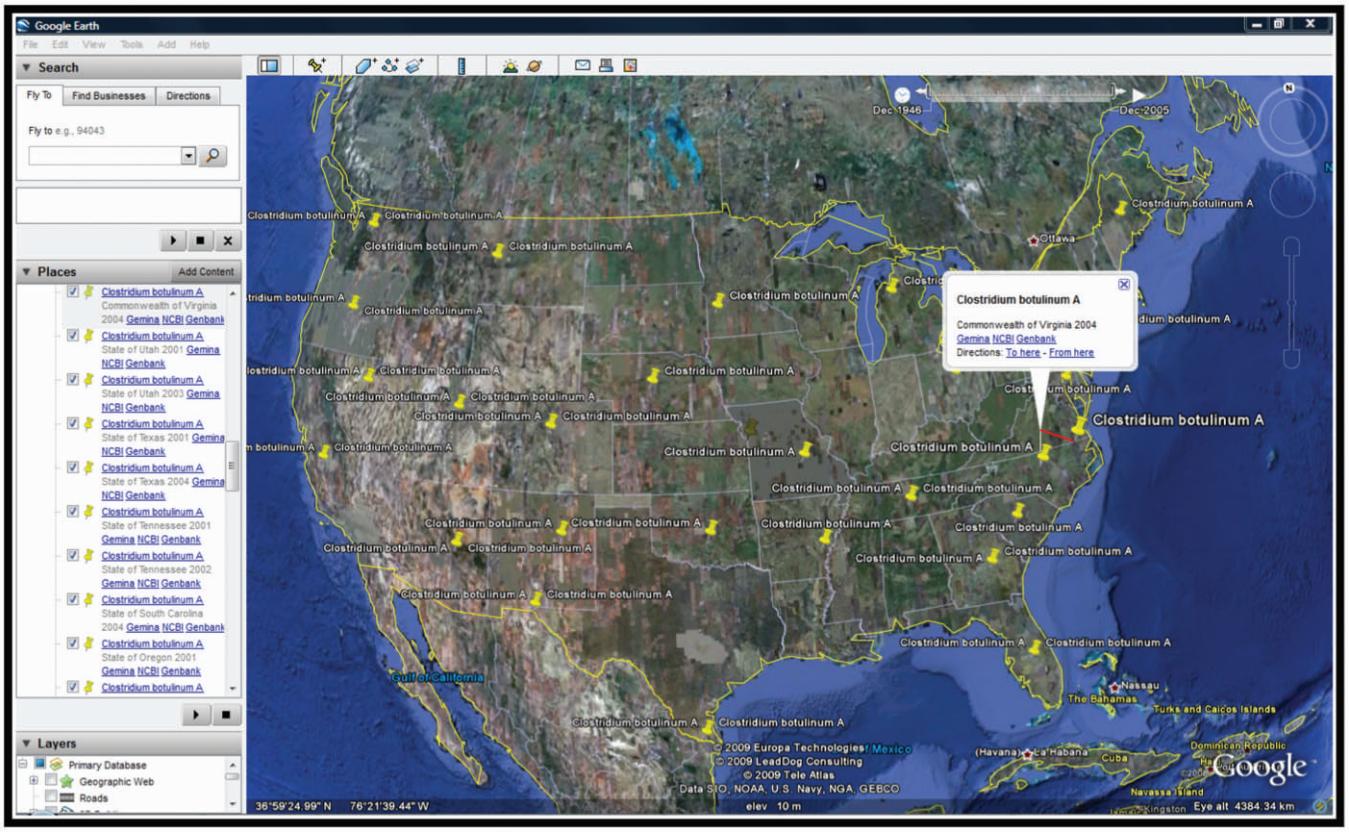

Figure 2. Gemina database geospatial surveillance resource. The Gemina Incident Search Report page provides geographic locations in GIS environment as demonstrated in this example of all incident locations for *Clostridium botulinum*. Links to related data at GenBank and NCBI's Taxonomy database are provided under 'Places' and as pop-ups for each incident.

To begin a search of the Gemina database: (i) select a Controlled Vocabulary; (ii) select one or more terms (from one or more vocabularies) to build their database query; and (iii) submit their database query by clicking on 'View Infections' or 'View Incidents'.

The Gemina Query interface is designed to provide users with an initial Tree View of the top level of each vocabulary. This view of the vocabulary provides the user with the opportunity to expand the branches of the tree, examine terms or groups of terms and to examine relatedness of terms and to select terms. This top level of each 'Tree View' also provides contextual knowledge of the terms, identifies related grouping of terms, parent-child term relationships and domain knowledge.

The Gemina web interface (Figure 1) presents data for each ontology and controlled vocabulary in a hierarchical tree format. The Tree View of each vocabulary can be search by entering text (proper names, common names, taxonomy IDs (where applicable) to identify matching terms within each tree. Multiple terms may be entered into the search box at one time, e.g. duck Alaska 1991, which are searched as 'duck AND Alaska AND 1991'. This query of the Pathogen vocabulary results in three influenza strains. The summary of the results is provided just below the search box as the number of results associated with data, e.g. *three results*. Clicking on this link will populate these results directly into the Selection Summary box for submission of these results for the 'View Infections' or 'View Incidents' query.

Some sample queries could include:

(1) select one or more types of metadata characterizing the pathogen–host–disease relationship; identify associated metadata; submit a refined search, narrowing the parameters by including a second or third vocabulary and terms by selecting 'Edit Query' from the top of the Gemina Search Report page;
(2) explore the breadth of outbreak data from published reports for one or more NIAID pathogens; select the Google Maps icon or select the Reference to read the associated PubMed article;
(3) identify a set of pathogens for DNA selection at Insignia based on their associated metadata; select the pathogens; select 'Submit all' from Tools to submit your selected pathogens to the Insignia Signature pipeline;
(4) view the geospatial patterns of outbreaks for an infectious disease or pathogen in Google Earth (Figure 2);
(5) view the precise location of outbreaks (Figure 3) for a subset of hosts in Google Maps; or
(6) select a subset of pathogens based on their means of transmission, known reservoirs, associated toxins, environmental habitats.

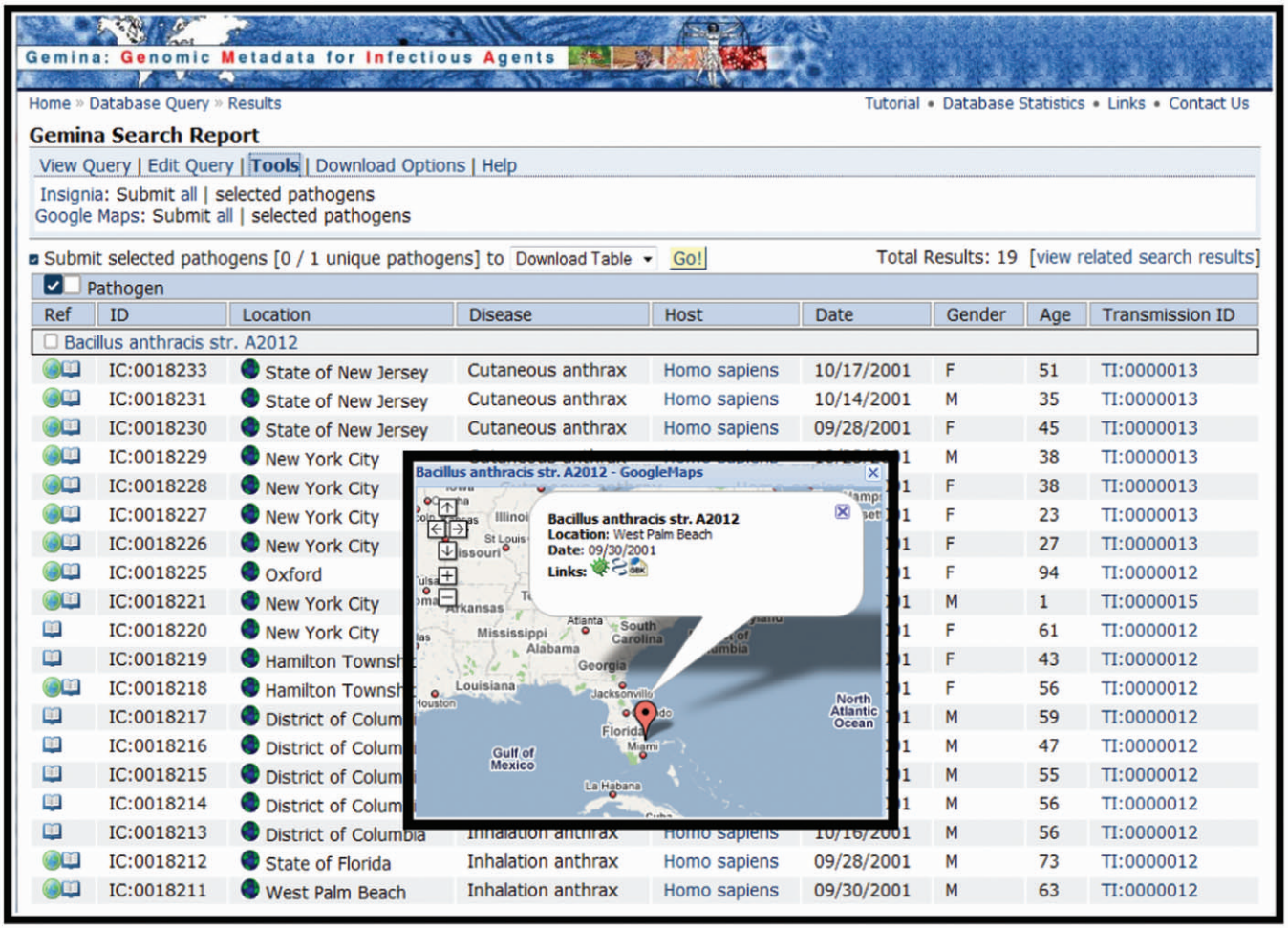

Figure 3. Incidents query results demonstration. Pathogen outbreak data in Gemina's Incident Results reporting the extent of outbreak locations for *Bacillus anthracis str*. A2012 (Florida strain) in 2001 provides an example of Gemina's incident metadata and Google Maps geographic location data display.

Data output

The Gemina Search Report results page (Figure 4) is pathogen-centric with the metadata for Infections Systems or Incidents reported, in separate columns, under the strain name of each pathogen. Each pathogen result set is presented as a separate block of results, with pathogens presented in alphabetical order from top to bottom. Data within each block are organized in descending order based on Gemina internal identifiers (infection system transmission IDs or incident IDs in column 2). The report page provides links to reservoir and toxin lists, curated references, options to select and submit pathogens to Insignia's DNA Sequence pipeline or download results in as a table or location data in KML format. Location data is viewable for each incident as place names, in Google Maps and Google Earth.

DISCUSSION

The Gemina system provides a geospatial surveillance tool to a diverse community of researchers enabling them to assess the emergence, host diversity, and global movement of viral and bacterial threat agents. The Gemina system has been designed to have a broad range of uses from the more research based applications such as disease modeling and comparative analysis between strains and species to applied applications such as disease surveillance and DNA based biomarker and diagnostic design. The Gemina system is accessed by the research community, on average 220 times per day by universities, government agencies, research institutes as well as the large scale data providers.

The Gemina system has fielded specific requests for biodefense NIAID priority pathogen surveillance research projects with a focus of identifying the diversity of locations, diseases and hosts for a particular pathogen responsible for a current outbreak. Gemina's standardization of a pathogen's 'chain of infection' and outbreak metadata provides the biodefense community with a outbreak history resource for data collection, comparison and integration.

The metadata standards developed in conjunction with the Gemina project provide a strong backbone for future

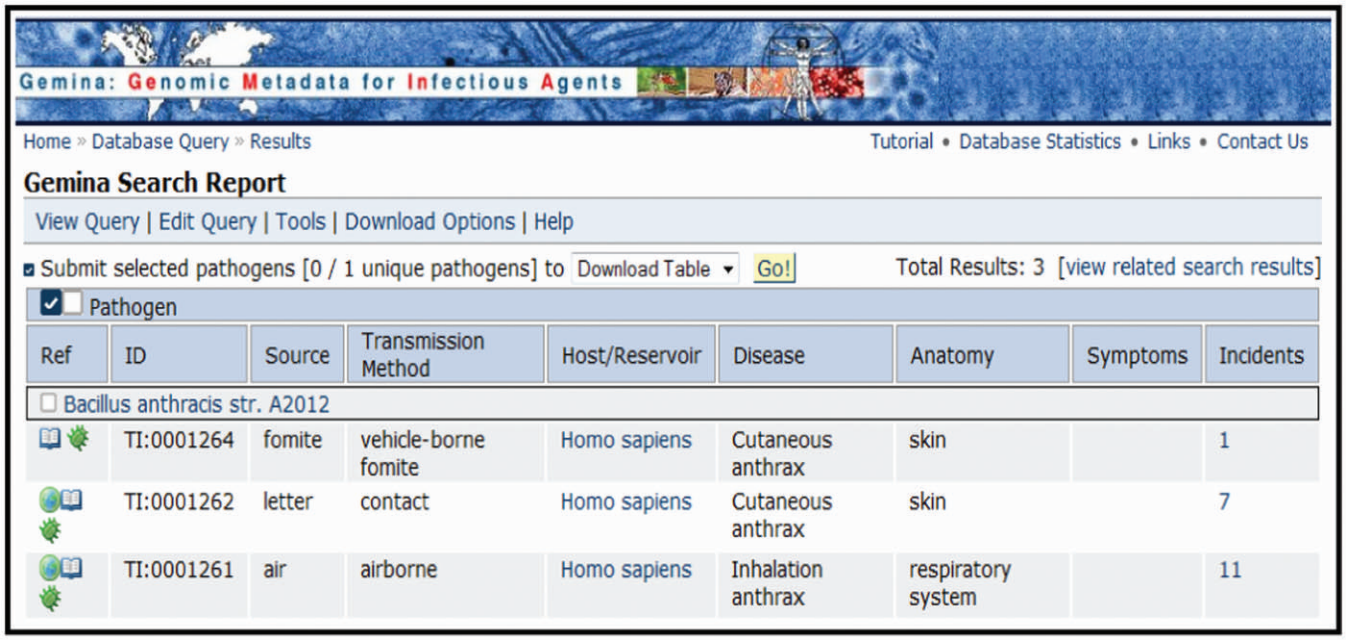

Figure 4. Gemina Infection results for multiple vocabulary query. Infection systems for *Bacillus anthracis str. A2012* (Florida strain) were identified from the Gemina database. The query included the pathogen name: *'Bacillus anthracis str. A2012';* the set of possible Anthrax diseases (disease ontology). The less restrictive search of '*Bacillus anthracis*' identified 19 unique pathogens and 68 infection systems and a wide range of hosts and 215 associated incidents.

data collection, database annotations and analysis as exemplified in ongoing food-borne pathogen collaborative projects with the Human Microbiome's Genomic Sequencing Center for Infectious Diseases (GSCID) and the FDA's Center for Food Security and Applied Nutrition. The Gemina user community additionally connects to Gemina through LinkOut links between NCBI Genome Projects database to the pathogen's results page, for example, http://gemina.igs.umaryland.edu/cgi-bin/gemina/GeneralQueryResults.cgi?pathogen=NCBITaxon:119857. Gemina surveillance data sets have been identified, for a research project at Los Alamos National Laboratory, based on a customized query using paired vocabulary-term tags. The surveillance project wanted to identify a set of standardized metadata for a subset of pathogens based on queries of disease, host and symptom. Based on URL query a text file representation of the Gemina tables can be downloaded with a number of download options and search terms. The base URL and some of the possible search terms are demonstration in this example:http://gemina.igs.umaryland.edu/cgi-bin/gemina/GeneralQueryResults.cgi?pathogen=!NCBITaxon:1392!&=&disease_names=&anatomy=&anatomy_namesview=incidents=&file_download=1&kml_download=0.

Gemina's focus on outbreaks of pathogens and the related epidemiological data has created a resource well tailored for the development of control measures and prevention strategies focused on breaking the chain of infection. Gemina's Infection Systems constitute a catalog of the key elements of the 'Chain of Infection' necessary for the spread of an infectious disease. Identification of these elements provides a repository of the focal points for the deployment of counter measures, as removal of one link in the chain could prevent infections from spreading and thereby break the chain of infection. The uniform information structure of Gemina's data allows users to search for several parameters simultaneously. In contrast to a standard PubMed literature search, Gemina searches provide a quick tool to investigate the multiple factors for infectious outbreaks documented in primary literature provided by biomedical, epidemiology and genomics researchers.

There have been a number of lessons learned in the creation of the Gemina database, retrieval of the data from primary literature and development of the suite of Gemina ontologies. We have found that data reporting biases have a definitive impact on the extent of data collected and recorded as ascertainment bias which is affected by the severity, location and time period of the event. Global outbreaks, such as the 2009 Swine Flu outbreak garnered an increase in reporting, data collection, sample collection and subsequent heightening of metadata output. Capturing metadata from primary published data and regularizing this metadata for inclusion in a relational database involves programmatic solutions for the diversity of factors involved that could confound term matching and term mapping. These factors have included: singular and plural terms, American and British English terms, non-English language terms, typos, common (colloquial) names, alternative names, former names, spelling variants [transliteration (between alphabets) variants and word order variants], formal names (e.g. used in nation states), and abbreviations.

FUTURE DIRECTIONS

The next stage of Gemina's development will involve the creation of an automated system for the identification of chain of infection elements from PubMed and other outbreak RSS feeds, additional consortium ontology development and expansion of the virulence factor ontology project. The foundation of the Gemina database will be augmented with daily updated PubMed articles for each pathogen–disease combination identified through PubMed abstracts pulled in from PubMeds RSS feed and identified based on key terms from the Gemina pathogen and disease vocabularies. As the backbone of the Gemina system has been built around the strength of the ongoing ontology development the Gemina system is well placed to provide the framework for cross database communication and data exchange to the biodefense and biomedical communities. The Gemina project will continue to build on this strong program of ontology and standards development partnering with a number of ontology consortium projects such as the Disease Ontology, the Environmental Habitats (EnvO) Ontology and Geographical Locations (GAZ) Gazetteer. These projects have strengthened the breadth and depth of ontology development for the Gemina project.

Availability

The Gemina system is an open source project. Gemina data sets, database, software, ontologies and controlled vocabularies are freely available on the Gemina website http://gemina.igs.umaryland.edu) and Gemina SourceForge site (http://gemina.sourceforge.net/). *Contact*: gemina@som.umaryland.edu, lschriml@som.umaryland.edu (PI)

Update frequency

The Gemina database is updated regularly on a cycle of primarily monthly updates driven by the generation of new data sets. The Gemina automated data update protocols process new data loads in a 48 h start to finish time frame.

ACKNOWLEDGEMENTS

The authors thank Matthew Davenport for his continue guidance and support of the Gemina project. They would like to thank Susan Bromberg and Mary Shimoyama at the Rat Genome Database for providing them with their disease ontology file. They are grateful to Michael Ashburner, Suzi Lewis, Warren Kibbe, Rex Chisholm, Norman Morrison, Dawn Field, Chris Mungall, Barry Smith, and the members of the OBO Foundry for their help in developing our ontology resources. They thank the following Gemina collaborators and colleagues for their continued support: Steven Salzberg, Adam Phillippy, Kunmi Ayanbule, Jay V. DePasse, Kumar Hari, Alan Goates, Ravi Jain, David Spiro, Naomi Sengamalay.

FUNDING

US Department of Homeland Security Science and Technology Directorate. [W81XWH-05-2-005, NBCH2070002]. Funding for open access charge: Institute for Genome Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Phillippy,A.M., Mason,J.A., Ayanbule,K., Sommer,D.D., Taviani,E., Huq,A., Colwell,R., Knight,I. and Salzberg,S. (2007) Comprehensive DNA discovery and validation. *PLoS Comput. Biol.*, **18**, 887–894.
2. Collier,N., Doan,S., Kawazoe,A., Goodwin,R.M., Conway,M., Tateno,Y., Ngo,Q.H., Dien,D., Kawtrakul,A., Takeuchi,K. *et al.* (2008) BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, **24**, 2940–2941.
3. Freifeld,C.C., Mandl,K.D., Reis,B.Y. and Brownstein,J.S. (2008) HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.*, **15**, 150–157.
4. Zeldenrust,M.E., Rahamat-Langendoen,J.C., Postma,M.J. and van Vliet,J.A. (2008) The value of ProMED-mail for the Early Warning Committee in the Netherlands: more specific approach recommended. *Euro Surveill.*, **13**, 8033.
5. Mungall,C.J., Emmert,D.B. and FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, 337–346.
6. Ecker,D.J., Sampath,R., Willett,P., Wyatt,J.R., Samant,V., Massire,C., Hall,T.A., Hari,K., McNeil,J.A., Büchen-Osmond,C. *et al.* (2005) The Microbial Rosetta Stone database: a compilation of global and emerging infectious microorganisms and bioterrorist threat agents. *BMC Microbiol.*, **5**, 1–19.
7. Hari,K.L., Goates,A.T., Jain,R., Towers,A., Harpin,V.S., Robertson,J.M., Wilson,M.R., Samant,V.S., Ecker,D.J., McNeil,J.A. *et al.* (2009) The Microbial Rosetta Stone: a database system for tracking infectious microorganisms. *Int. J. Legal Med.*, **123**, 65–69.
8. 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.
9. Smith,B., Ceusters,W., Klagges,B., Köhler,J., Kumar,A., Lomax,J., Mungall,C., Neuhaus,F., Rector,A.L. and Rosse,C. (2005) Relations in biomedical ontologies. *Genome Biol.*, **6**, R46.
10. U.S. Department of Health and Human Services Public Health Service, CDC. (1992) Principles of Epidemiology. *An introduction to Applied Epidemiology and Biostatistics*, 2nd edn. Atlanta.