Pathema: a clade-specific bioinformatics resource center for pathogen research

Lauren M. Brinkac1, Tanja Davidsen1, Erin Beck1, Anuradha Ganapathy4, Elisabet Caler1, Robert J. Dodson1,2, A. Scott Durkin1, Derek M. Harkins1, Hernan Lorenzi1, Ramana Madupu1, Yinong Sebastian1,3, Susmita Shrivastava1, Mathangi Thiagarajan1, Joshua Orvis4, Jae Deep P. Sundaram1, Jonathon Crabtree4, Kevin Galens4, Yongmei Zhao4,5, Jason M. Inman1, Robert Montgomery1, Seth Schobel1, Kevin Galinsky1, David M. Tanenbaum1, Adam Resnick1, Nikhat Zafar1, Owen White4 and Granger Sutton1,*

1J. Craig Venter Institute, Rockville, MD 20850, 2Northwestern University, Chicago, IL 60208, 3National Institutes of Health-NIAID, Bethesda, MD 20892, 4Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201 and 5SAIC/National Cancer Institute, Gaithersburg, MD 20877, USA

Received August 14, 2009; Revised September 17, 2009; Accepted September 18, 2009

ABSTRACT
Pathema (http://pathema.jcvi.org) is one of the eight Bioinformatics Resource Centers (BRCs) funded by the National Institute of Allergy and Infectious Disease (NIAID) designed to serve as a core resource for the bio-defense and infectious disease research community. Pathema strives to support basic research and accelerate scientific progress for understanding, detecting, diagnosing and treating an established set of six target NIAID Category A–C pathogens: Category A priority pathogens; Bacillus anthracis and Clostridium botulinum, and Category B priority pathogens; Burkholderia mallei, Burkholderia pseudomallei, Clostridium perfringens and Entamoeba histolytica. Each target pathogen is represented in one of four distinct clade-specific Pathema web resources and underlying databases developed to target the specific data and analysis needs of each scientific community. All publicly available complete genome projects of phylogenetically related organisms are also represented, providing a comprehensive collection of organisms for comparative analyses. Pathema facilitates the scientific exploration of genomic and related data through its integration with web-based analysis tools, customized to obtain, display, and compute results relevant to ongoing pathogen research. Pathema serves the bio-defense and infectious disease research community by disseminating data resulting from pathogen genome sequencing projects and providing access to the results of inter-genomic comparisons for these organisms.

INTRODUCTION
Pathema is a community driven bioinformatics resource that provides access to genomic data integrated with analysis tools designed to aid researchers in identifying potential targets for novel therapeutics, vaccines, and diagnostics for six selected National Institute of Allergy and Infectious Disease (NIAID) priority pathogens (1). Organisms classified by NIAID as priority pathogens are selected based on their association as agents or potential agents of bioterrorism. The priority pathogens Pathema supports includes five prokaryotes Bacillus anthracis, Burkholderia mallei, Burkholderia pseudomallei, Clostridium botulinum, and Clostridium perfringens and one eukaryote Entamoeba histolytica. To provide researchers with a comprehensive collection of organisms for comparative analyses, 66 unique strains of priority pathogens are supported, to include 54 phylogenetically related species. Organisms are grouped taxonomically by genus, with associated data stored in four distinct databases, each accessible through four different clade web interfaces. Each Pathema clade resource, linked from one central Pathema gateway interface, is tailored to address the specific data and analysis needs of each scientific community; feedback gathered through outreach activities. Pathema disseminates high-quality, up-to-date data to

*To whom correspondence should be addressed. Tel: +1 301 795 7293; Fax: +1 301 294 3142; Email: gsutton@jcvi.org

© The Author(s) 2009. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
include genome sequence, annotation data types and curation assertions, and specialty datasets as they relate to ongoing pathogen and infectious disease research. The most current data generated is displayed throughout the resource and Pathema deposits all relevant data in public repositories such as the Pathogen Portal (http://www.pathogenportal.org/), GenBank (2) and the GO repository (3). Integrated with this data is a suite of sophisticated bioinformatics software and over 50 analysis tools customized to retrieve, display and compute results relevant to the research of each Pathema target pathogen community. Bioinformatics tools for cross-genome comparisons and identification of metabolic pathways are also integrated to facilitate the identification of potential targets for vaccine development, therapeutics, and diagnostics. In addition, clade-specific training courses, detailed tutorials, standard operating procedures are offered to provide instruction and documentation on the use of this system and underlying databases.

**PATHEMA ORGANISMS**

Pathema supports sequence and detailed curation of six NIAID target priority pathogens and related species grouped taxonomically by genus into four clades: *Bacillus*, *Burkholderia*, *Clostridium* and *Entamoeba* (Table 1). These pathogens are included among two of three high-priority categories (Categories A, B and C) classified by NIAID based on their relative capabilities for causing morbidity or mortality from disease in case of biowarfare (http://www3.niaid.nih.gov/topics/BiodefenseRelated/Biodefense/research/CatA.htm). The inclusion of closely related species provides researchers with a comprehensive collection of organisms for comparative analyses.

The *Bacillus* clade supports 40 prokaryotic organisms including the target pathogen *B. anthracis* (Category A), as well as the pathogens *B. cereus* and *B. thuringiensis*. Long regarded as one of the preferred biological warfare agents, *B. anthracis* is the causative agent of anthrax. Its potential for use as a bioweapon was demonstrated by the autumn 2001 anthrax letter attacks in the US. Its lethality, combined with ease of laboratory production and ability to disseminate anthrax spores in aerosol form, accounts for its interest as a biowarfare agent (4).

Included among the 41 prokaryotes supported by the *Burkholderia* clade are the target pathogens *B. mallei* and *B. pseudomallei* (Category B), as well as the pathogen *B. cepacia*. *B. mallei* is responsible for glanders, a disease that occurs mostly in horses and related animals. Glanders has been associated with war for centuries, to include the use of *B. mallei* as a bioweapon in World War I, World War II, and anecdotal evidence supports its use in Afghanistan. Its ease of transmission and severity of disease makes *B. mallei* of interest as an agent for bioterrorism (5). *Burkholderia pseudomallei*, a human and animal pathogen, is the causative agent of melioidosis, an infectious disease endemic to Southeast Asia and northern Australia, and may occur in other tropical and subtropical regions. Its severe course of infection, aerosol infectivity and worldwide availability resulted in its inclusion as a potential agent of biological warfare or bioterrorism (6).

The *Clostridium* clade supports 36 prokaryotic organisms encompassing the four main species responsible for disease in humans. These include the target pathogens *C. botulinum* (Category A), *C. perfringens* (Category B), as well as the pathogens *C. difficile* and *C. tetani*. Different strains of *C. botulinum* produce different types of toxins apart from the well-known botulinum neurotoxin, the causative agent of the disease botulism in humans and animals (4). The botulism toxin, considered the most lethal naturally occurring substance, was linked for use as a bioweapon during World War II and the Persian Gulf War (7). *C. perfringens* is known to be the most widely distributed pathogen in nature. It is shown to be a causative agent of human diseases such as gas gangrene, food poisoning, and enteritis necroticans, as well as various animal diseases (5).

Included in the *Entamoeba* clade are three parasitic protists: *E. histolytica*, *E. dispar* and *E. invadens*. The target pathogen *E. histolytica* (Category B), is the causative agent of the most common diarrheal disease, amebiasis. Amebiasis accounts for between 40,000 and 100,000 deaths annually, and is predominantly seen in developing countries where a high prevalence of infection

<table>
<thead>
<tr>
<th>Pathema clade</th>
<th>Target NIAID pathogen</th>
<th>Organisms supported</th>
<th>Completed genomes</th>
<th>Draft genomes</th>
<th>NIAID category</th>
<th>Associated disease</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacillus</strong></td>
<td><em>Bacillus anthracis</em></td>
<td>40</td>
<td>21</td>
<td>19</td>
<td>A</td>
<td>Anthrax</td>
</tr>
<tr>
<td></td>
<td><em>Bacillus cereus</em></td>
<td>19</td>
<td>6</td>
<td>13</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Bacillus thuringiensis</em></td>
<td>41</td>
<td>24</td>
<td>18</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td><strong>Burkholderia</strong></td>
<td><em>Burkholderia mallei</em></td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>B</td>
<td>Glanders</td>
</tr>
<tr>
<td></td>
<td><em>Burkholderia pseudomallei</em></td>
<td>12</td>
<td>4</td>
<td>8</td>
<td>B</td>
<td>Melioidosis</td>
</tr>
<tr>
<td><strong>Clostridium</strong></td>
<td><em>Clostridium botulinum</em></td>
<td>36</td>
<td>23</td>
<td>13</td>
<td>A</td>
<td>Botulism</td>
</tr>
<tr>
<td></td>
<td><em>Clostridium perfringens</em></td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>A</td>
<td>Enterotoxemia</td>
</tr>
<tr>
<td><strong>Entamoeba</strong></td>
<td><em>Entamoeba histolytica</em></td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>B</td>
<td>Amebiasis</td>
</tr>
<tr>
<td><strong>Total Pathema</strong></td>
<td></td>
<td>120</td>
<td>71</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A complete list of supported organisms is included in Supplementary Table S1.
is due to fecal contamination of food and water supply, factors that cannot be immediately remedied due to limited financial resources in these countries (8). Its interest as a potential biothreat organism is its low infectious dose and potential for dissemination through compromised food and water supplies.

To assist researchers in identifying correlations between patient phenotype and geography, symptoms/outcome and pathogen sequence variation, and to gain an understanding of the impact of pathogen genomic variations on drug resistance or vaccine efficacy, Pathema integrates epidemiological and clinical data. Where available, this data is obtained from the research community for each organism and includes: the original source location of each organism strain, detailed clinical information (e.g. date isolated, isolation source, historical background), genotype numbering based on Multi Locus Sequence Typing (9), and source contact information for obtaining the DNA.

**INTERFACE DESIGN AND DATABASE DESCRIPTION**

The main Pathema gateway interface serves as the central entry point to access Pathema’s target pathogens and related species through one of four distinct clade-specific web resources: *Bacillus, Clostridium, Burkholderia* and *Entamoeba*. This gateway provides general information, news and highlights, planned data updates, and tutorials relevant to the entire Pathema resource, with links to each of the four clade sites supporting clade-specific data and analysis tools. Based on feedback gathered through community outreach, Pathema’s four clade resources aim to target the individual research needs of each community by integrating the specific datasets and analysis tools requested by organism experts. Through the customized development of clade resources, Pathema serves as a core resource supporting scientific investigation and hypothesis generation of its supported target organisms.

The Pathema web interface uses the Coati (Collaborative Open Applications Tool Initiative) architecture framework. Coati is an open source project housed at SourceForge (http://sourceforge.net/projects/coati-api/). Each clade-specific web interface interacts with one of four separate Chado (10) relational database schemas that house Pathema clade sequence and annotation data, and comparative computes. Chado underlies many Generic Model Organism Database (GMOD) (11) installations and is a general schema used to share genomic data, annotations and analyses.

**CURATION DATA TYPES**

Pathema generates and continuously updates gene model and functional annotation data for 120 supported genome projects, disseminating data of over 600,000 predicted genes with common data types (Table 2). Common data types are assigned using an automated pipeline to process the genomic sequences of all Pathema organisms. This pipeline consists of several algorithms for the prediction of gene models and genome features (e.g. RNAs, terminators, repeats), and employs a hierarchical evidence ranking scheme to assign functional annotation [e.g. protein name, gene symbol, Enzyme Commission (EC) number (12), Gene Ontology (GO) terms]. By assigning common data types using one standardized pipeline across all organisms, comparative analyses become easier and more meaningful to the researcher. Additionally, based on the use of common data types, a rich set of curation assertions with supporting evidence are generated. These curation assertions are based on the Gene Ontology Consortium and attempt to describe the complete profile (i.e. molecular function, biological process, cellular location) of proteins in biologically meaningful ways, those that cannot be captured by individual data types alone. Standardized evidence types represent a diverse range of specific forms of evidence (i.e. direct assay, mutant phenotype) used to support each curation assertion. The use of standardized evidence types facilitates a mechanism to easily assess the level of confidence supporting each assertion, ultimately validating hypotheses derived from the profile analysis of individual proteins, orthologs and pathway data.

Common annotation data types and curation assertions with supporting evidence are computationally generated for all Pathema organisms. With the goal of providing the scientific community with the most accurate annotation, automated predictions are manually curated for each of Pathema’s six target pathogens. Established naming conventions and evidence interpretation guidelines are adhered to during this manual process. Additionally, the genomic annotation of these organisms reflects in-depth manual literature curation of biodefense and infectious disease related datasets. These datasets include clade-specific virulence factors, epitopes (13), protein–protein interactions (14), multidrug exporters (15) and experimentally characterized proteins. Inclusion of these datasets enrich existing genome annotation, thereby facilitating the identification of potential new targets of pathogen research interest.

Although Pathema’s six target pathogens are the primary focus of manual effort, Pathema strives to provide the same level of high-quality annotation across all organisms supported by the Pathema resource. To achieve this, a homology mapping strategy is employed. This strategy uses the MUMmer (16) whole genome alignment program to identify close protein homologs, with subsequent propagation of high-quality manually curated data from each target organism to all closely related Pathema clade members.

All annotation standard operating procedures, Pathema’s Gene Naming and Annotation Guidelines, and all other related annotation documentation is obtainable throughout the Pathema resource (http://pathema.jcvi.org/protocols).

**GENOME AND COMPARATIVE ANALYSIS TOOLS**

Pathema supports over 50 web-based data mining, single gene, whole-genome and multi-genome comparative tools.
to facilitate analyses of genomic sequence and annotation data across Pathema organisms. Tools are designed to facilitate scientific exploration in the areas of functional curation, pathogenicity, therapeutics, comparative analysis and functional genomics. While every tool has several applications, taken together they provide numerous opportunities for discovery and hypothesis generation (Supplementary Table S2).

Data mining
Pathema incorporates over 25 different search capabilities that enable data mining and retrieval of all data types stored in the Pathema database. Search tools query genes, genomes, sequences or text, matching user-defined strings across gene loci, gene symbols and protein product names. Virulence factors, epitopes, experimentally characterized proteins and protein interaction data can be retrieved using Pathema search tools across user-selected organisms. Other queries include EC#, GenBank, SwissProt (17) and GO id searches, and common sequence search methods such as BLAST (18), Hidden Markov Model (19) and protein motif searches (20) are also available.

Literature mining
A semantic visualization tool, based on the National Library of Medicine’s SemMed viewer (21), is integrated within Pathema. This tool provides access to biomedical literature archived in PubMed, through manually curated semantic condensate data records of relevant subjects for each Pathema clade. Records can be displayed in both graphical and word cloud format, and include links to external data sites containing relevant information, such as genetic databases, Unified Medical Language System (UMLS) entries and the original Medline reference.

Single gene analysis
Individual gene pages highlight annotation data and associated evidence, as well as provide access to single gene analysis tools for every gene available on Pathema. Annotation data displayed and downloadable includes protein product name, gene symbol, EC#, GO ids, functional role category assignment, and DNA and protein sequences. Literature references are provided for all proteins that are identified virulence factors, are associated with an epitope(s), interact with another protein(s), or have experimental characterizations. Calculating the transmembrane HMM profile (22), secondary structure and third position GC-Skew are just a few types of analyses that can be performed. Links to other relevant resources such as UniProt, GenBank, Prosite, Pfam (23), etc. are also available.

Whole-genome analysis
Over 20 different displays and analyses of whole-genome data are included in Pathema. These analysis tools enable the display and analysis of individual genomic data using a variety of different methods. Whole-genome data can
be displayed graphically as a linear representation of genes on regions of a chromosome or as a complete circle for an entire chromosome. Data can be investigated through biochemical pathways (24–26), codon usage tables, percent GC plots, computer generated 2D and restriction digest gels, and summary information such as average gene size or numbers of coding regions can be retrieved as viewable and downloadable tables and lists.

**Comparative analysis**

Integrated into Pathema are over 15 different comparative analysis tools for multi-genome comparisons among Pathema clade organisms (Figure 1). The basis for Pathema’s current comparative tools is either pre-generated Jaccard orthologous protein clusters or All versus All blastp searches. Incorporated, are the most popular tools of the publicly available Sybil comparative analysis suite (27). Sybil uses Pathema’s pre-generated protein clusters as the underlying data for its synteny gradient and comparative genomic displays.

Sybil protein cluster ortholog, paralog and singleton data are also available.

**COMMUNITY OUTREACH**

Pathema launched a community outreach strategic plan to assess the scientific and informatic needs of the pathogen research community. This community consists of over 950 identified researchers who study the six Pathema target pathogens, with over 25% participating in Pathema community outreach efforts. These efforts were designed to gather feedback during the initial phases of resource development and testing, with feedback continuously gathered during various training and other outreach activities. Pathema provides detailed training in the form of clade-specific annotation jamborees and hands-on Pathema resource workshops conducted both on site and in conjunction with major organism specific conferences. In-depth resource tutorials and manuals that describe Pathema tools and data are also available. Currently 20
scientific publications reference the use of Pathema and its underlying data sets (28–46).

AVAILABILITY
Pathema is maintained at the J. Craig Venter Institute and can be accessible through a web browser at http://pathema.jcvi.org. There are no license restrictions for user access to any of the data supported by Pathema, and all source code is managed under an open-source collaborative development paradigm. Web scripts and data maintenance programs are located at SourceForge under the Pathema project (http://sourceforge.net/projects/pathema). Pathema sequence and annotation data formatted GFF3 files can be obtained from the Pathema FTP download site (ftp://ftp.pathogenportal.org/gff3/Pathema/); retrievable from the ‘downloads’ tab off the main resource header or linked directly from each organism homepage. Additionally results obtained from complex searches or genomic comparisons are available in tab-delimited format throughout Pathema on each respective results page.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS
The authors would like to thank the J. Craig Venter Institute Information Technology and Bioinformatics Departments for their ongoing technical, engineering and scientific support to include Michael Heaney, Darnell Edwards, Tom Emmel, Dan H. Haft, Roland Richter and Jeremy Selengut as well as the support received from the Institute for Genomic Sciences to include Sam Angiuli, Sean Daugherty, Michelle Gwinn Giglio, Heather Huot, Anup Mahrkar and Jennifer Wortman. The authors would also like to thank Tom Rindflesch and Dongsu Kim Shin from the Lister Hill National Center for Biomedical Communications for providing the version of SemMed that was used in Pathema development activities.

FUNDING
National Institute of Allergy and Infectious Disease contact HHSN266200400038C. Funding for open access charge: NIAID.

Conflict of interest statement. None declared.

REFERENCES


