D408–D414 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 20 October 2009*
doi:10.1093/nar/gkp850

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

Lauren M. Brinkac[1], Tanja Davidsen[1], Erin Beck[1], Anuradha Ganapathy[4], Elisabet Caler[1], Robert J. Dodson[1,2], A. Scott Durkin[1], Derek M. Harkins[1], Hernan Lorenzi[1], Ramana Madupu[1], Yinong Sebastian[1,3], Susmita Shrivastava[1], Mathangi Thiagarajan[1], Joshua Orvis[4], Jaideep P. Sundaram[1], Jonathon Crabtree[4], Kevin Galens[4], Yongmei Zhao[4,5], Jason M. Inman[1], Robert Montgomery[1], Seth Schobel[1], Kevin Galinsky[1], David M. Tanenbaum[1], Adam Resnick[1], Nikhat Zafar[1], Owen White[4] and Granger Sutton[1,]*

[1]J. Craig Venter Institute, Rockville, MD 20850, [2]Northwestern University, Chicago, IL 60208, [3]National Institutes of Health-NIAID, Bethesda, MD 20892, [4]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201 and [5]SAIC/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 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 1. Genomes and organisms supported by Pathema as of 1 August 2009

Pathema clade	Target NIAID pathogen	Organisms supported	Completed genomes	Draft genomes	NIAID category	Associated disease
Bacillus		40	21	19		
	Bacillus anthracis	19	6	13	A	Anthrax
Burkholderia		41	24	18		
	Burkholderia mallei	10	4	6	B	Glanders
	Burkholderia pseudomallei	12	4	8	B	Melioidosis
Clostridium		36	23	13		
	Clostridium botulinum	15	10	5	A	Botulism
	Clostridium perfringens	9	3	6	B	Enterotoxemia
Entamoeba		3	3	0		
	Entamoeba histolytica	1	1	0	B	Amebiasis
Total Pathema		120	71	50		

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

Table 2. Pathema curation assertions

Pathema clade	Total organisms	Predicted genes	Evidence types supporting manual curation					Curated specialty genes					Annotation data types			Curation assertions		
			Sequence similarity	Mutant phenotype	Expression pattern	Direct assay	Genome context	Epitopes	Virulence factors	Multidrug exporters	Protein interactions	Experimentally verified	Protein name (%)	Gene symbol (%)	EC number (%)	Molecular function (%)	Biological process (%)	Cellular component (%)
Bacillus	40	217 352	10 645	61	0	57	12	758	74	6473	163	343	69	20	14	91	94	36
Burkholderia	41	245 739	48 142	104	11	72	110	418	122	5448	52	714	70	19	15	82	80	40
Clostridium	36	131 359	28 803	3	1	32	55	345	17	2897	1	227	73	22	16	74	73	34
Entamoeba	3	28 560	2537	1	0	14	0	0	176	141	0	31	12	1	14	16	10	5
Total	120	623 010	90 127	169	12	175	177	1521	389	14 959	216	1315	68	19	15	80	80	36

Only a subset of annotation data types and curation assertions used by Pathema to describe predicted genes based on supporting evidence are included.

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

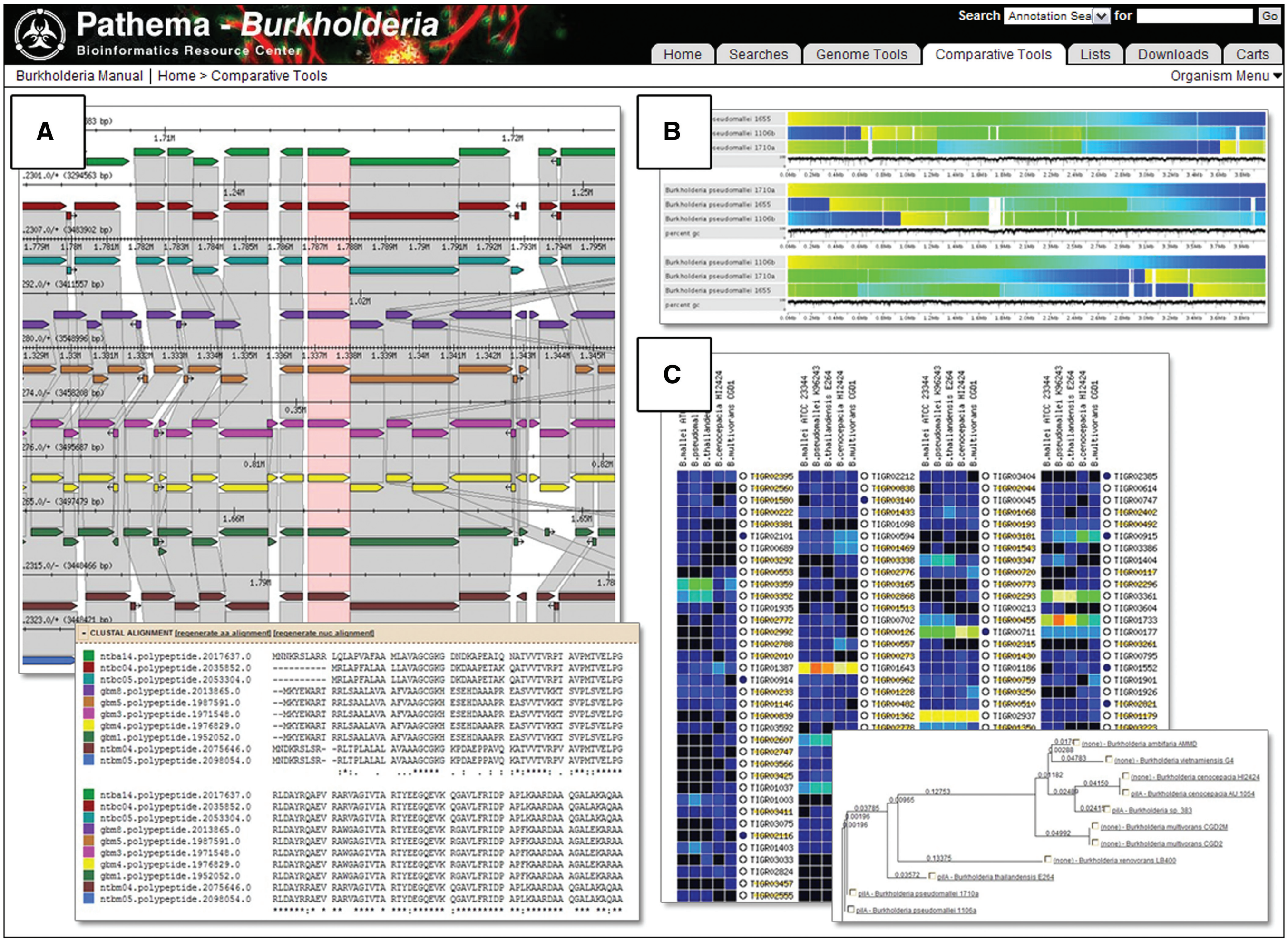
A semantic visualization tool, based on the National Library of Medicine's SemMed viewer (21), is integrated within Pathema. This tool provides access to biomedical literature archived in PubMed, through manually curated semantic condensate data records of relevant subjects for each Pathema clade. Records 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


Pathema - Burkholderia
Bioinformatics Resource Center
Search Annotation Sea for Go
Home
Searches
Genome Tools
Comparative Tools
Lists
Downloads
Carts
Burkholderia Manual | Home > Comparative Tools
Organism Menu
A
B
C
CLUSTAL ALIGNMENT [regenerate aa alignment] [regenerate nuc alignment]

```
ntba14.polypeptide.2017637.0   MNNKRSLARR LQLAPVAFAA MLAVAGCGKG DNDKAPEAIQ NATVVTVRPT AVPMTVELPG
ntbc04.polypeptide.2035852.0   ---------- MRLAPFALAA LLAVAGCGKG DKDAAPETAK QATVVTVRPT AVPMTVELPG
ntbc05.polypeptide.2053304.0   ---------- MRLAPFALAA LLAVAGCGKG DKDAAPETAK QATVVTVRPT AVPMTVELPG
gbm8.polypeptide.2013865.0     --MKYEWART RRLSAALAVA AFVAAGCGKH ESEHDAAAPR EASVVTVKKT SVPLSVELPG
gbm5.polypeptide.1987591.0     --MKYEWART RRLSAALAVA AFVAAGCGKH ESEHDAAAPR EASVVTVKKT SVPLSVELPG
gbm3.polypeptide.1971548.0     --MKYEWART RRLSAALAVA AFVAAGCGKH ESEHDAAAPR EASVVTVKKT SVPLSVELPG
gbm4.polypeptide.1976829.0     --MKYEWART RRLSAALAVA AFVAAGCGKH ESEHDAAAPR EASVVTVKKT SVPLSVELPG
gbm1.polypeptide.1952052.0     --MKYEWART RRLSAALAVA AFVAAGCGKH ESEHDAAAPR EASVVTVKKT SVPLSVELPG
ntbm04.polypeptide.2075646.0   MNDKRSLSR- -RLTPLALAL AVAAAGCGKG KPDAEPPAVQ KATVVTVRPV AVPMTVELPG
ntbm05.polypeptide.2098054.0   MNDKRSLSR- -RLTPLALAL AVAAAGCGKG KPDAEPPAVQ KATVVTVRPV AVPMTVELPG
                                           :*:.   .   ...*****  . :  . : : :*:****: . :**::*****

ntba14.polypeptide.2017637.0   RLDAYRQAPV RARVAGIVTA RTYEEGQEVK QGAVLFRIDP APLKAARDAA QGALAKAQAA
ntbc04.polypeptide.2035852.0   RLDAYRQAEV RARVAGIVIA RTYEEGQEVK QGAVLFRIDP APLKAARDAA QGALAKAQAA
ntbc05.polypeptide.2053304.0   RLDAYRQAEV RARVAGIVIA RTYEEGQEVK QGAVLFRIDP APLKAARDAA QGALAKAQAA
gbm8.polypeptide.2013865.0     RLDAYRQAEV RAWGAGIVTA RTYEEGQEVK RGAVLFRIDP APFKAARDAA AGALEKARAA
gbm5.polypeptide.1987591.0     RLDAYRQAEV RAWGAGIVTA RTYEEGQEVK RGAVLFRIDP APFKAARDAA AGALEKARAA
gbm3.polypeptide.1971548.0     RLDAYRQAEV RAWGAGIVTA RTYEEGQEVK RGAVLFRIDP APFKAARDAA AGALEKARAA
gbm4.polypeptide.1976829.0     RLDAYRQAEV RAWGAGIVTA RTYEEGQEVK RGAVLFRIDP APFKAARDAA AGALEKARAA
gbm1.polypeptide.1952052.0     RLDAYRQAEV RAWGAGIVTA RTYEEGQEVK RGAVLFRIDP APFKAARDAA AGALEKARAA
ntbm04.polypeptide.2075646.0   RLDAYRRAEV RARVAGIVTA RTYDEGQEVK QGAVLFRIDP APLKAARDAA QGALAKAQAA
ntbm05.polypeptide.2098054.0   RLDAYRRAEV RARVAGIVTA RTYDEGQEVK QGAVLFRIDP APLKAARDAA QGALAKAQAA
                               ******:* * **  **** * ***:****** :********* **:****** *** **:**
```

Figure 1. Pathema-*Burkholderia* Comparative Tools. This figure shows some of the comparative tools available on Pathema for the *Burkholderia* clade. (**A**) Protein orthologous cluster: *Burkholderia* multidrug efflux pump AmrA region and Clustal alignments; (**B**) Comparative genomic region: *Burkholderia* whole genomes aligned to a reference; (**C**) Evidence comparison: differences in evidence occurrence across multiple *Burkholderia* genomes and phylogeny of selected proteins.

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 Angiuoli, Sean Daugherty, Michelle Gwinn Giglio, Heather Huot, Anup Mahurkar and Jennifer Wortman. The authors would also like to thank Tom Rindflesch and Dongwook 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 contract HHSN266200400038C. Funding for open access charge: NIAID.

Conflict of interest statement. None declared.

REFERENCES

1. Greene,J.M., Collins,F., Lefkowitz,E.J., Roos,D., Scheuermann,R.H., Sobral,B., Stevens,R., White,O. and Di Francesco,V. (2007) National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect. Immun.*, **75**, 3212–3219.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
3. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
4. Darling,R.G., Catlett,C.L., Huebner,K.D. and Jarrett,D.G. (2002) Threats in bioterrorism. I: CDC category A agents. *Emerg. Med. Clin. North Am.*, **20**, 273–309.
5. Moran,G.J. (2002) Threats in bioterrorism. II: CDC category B and C agents. *Emerg. Med. Clin. North Am.*, **20**, 311–330.
6. Gilad,J., Harary,I., Dushnitsky,T., Schwartz,D. and Amsalem,Y. (2007) Burkholderia mallei and Burkholderia pseudomallei as bioterrorism agents: national aspects of emergency preparedness. *Isr. Med. Assoc. J.*, **9**, 499–503.
7. Roffey,R., Tegnell,A. and Elgh,F. (2002) Biological warfare in a historical perspective. *Clin. Microbiol. Infect.*, **8**, 450–454.
8. Upcroft,P. and Upcroft,J.A. (2001) Drug targets and mechanisms of resistance in the anaerobic protozoa. *Clin. Microbiol. Rev.*, **14**, 150–164.
9. Urwin,R. and Maiden,M.C.J. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.*, **11**, 479–487.
10. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
11. O'Connor,B.D., Day,A., Cain,S., Arnaiz,O., Sperling,L. and Stein,L.D. (2008) GMODWeb: a web framework for the Generic Model Organism Database. *Genome Biol.*, **9**, R102.
12. Webb,E.C. (1992) *Enzyme Nomenclature*. Academic Press, San Diego, California.
13. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R., Lund,O. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.*, **3**, e91.
14. Goll,J., Rajagopala,S.V., Shiau,S.C., Wu,H., Lamb,B.T. and Uetz,P. (2008) MPIDB: the microbial protein interaction database. *Bioinformatics*, **24**, 1743–1744.
15. Busch,W. and Saier,M.H. (2002) The Transporter Classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol.*, **37**, 287–337.
16. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
17. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
18. Altschul,S., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
19. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
20. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., De Castro,E., Langendijk-Genevaux,P.S., Pagni,M. and Sigrist,C.J. (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
21. Kilicoglu,H., Fiszman,M., Rodriguez,A., Shin,D., Ripple,A.M. and Rindflesch,T.C. (2008) Semantic MEDLINE: a web application to manage the results of PubMed searches. *Proceedings of the Third International Symposium for Semantic Mining in Biomedicine (SMBM), Turku Finland*, Sep. 1–3;69–76.
22. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
23. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
24. Haft,D.H., Selengut,J.D., Brinkac,L.M., Zafar,N. and White,O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, **21**, 293–306.
25. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–S232.

26. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
27. Crabtree,J., Angiuoli,S.V., Wortman,J.R. and White,O.R. (2007) Sybil: methods and software for multiple genome comparison and visualization. *Methods Mol. Biol.*, **408**, 93–108.
28. Abhyankar,M.M., Hochreiter,A.E., Connell,S.K., Gilchrist,C.A., Mann,B.J. and Petri,W.A. Jr (2009) Development of the Gateway system for cloning and expressing genes in *Entamoeba histolytica*. *Parasitol. Int.*, **58**, 95–97.
29. Cer,R.Z., Mudunuri,U., Stephens,R. and Lebeda,F.J. (2009) IC50-to-Ki: a web-based tool for converting IC50 to Ki values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Res.*, **37**, W441–W445.
30. Janvilisri,T., Scaria,J., Thompson,A.D., Nicholson,A., Limbago,B.M., Arroyo,L.G., Songer,J.G., Grohn,Y.T. and Chang,Y.F. (2009) Microarray identification of *Clostridium difficile* core components and divergent regions associated with host origin. *J. Bacteriol.*, **191**, 3881–3891.
31. Cruz-Castaneda,A., Hernandez-Sanchez,J. and Olivares-Trejo,J.J. (2009) Cloning and identification of a gene coding for a 26-kDa hemoglobin-binding protein from *Entamoeba histolytica*. *Biochimie*, **91**, 383–389.
32. Melendez-Hernandez,M.G., Barrios,M.L., Orozco,E. and Luna-Arias,J.P. (2008) The vacuolar ATPase from *Entamoeba histolytica*: molecular cloning of the gene encoding for the B subunit and subcellular localization of the protein. *BMC Microbiol.*, **8**, 235.
33. Zhang,H., Ehrenkaufer,G.M., Pompey,J.M., Hackney,J.A. and Singh,U. (2008) Small RNAs with 5′-polyphosphate termini associate with a Piwi-related protein and regulate gene expression in the single-celled eukaryote *Entamoeba histolytica*. *PLoS Pathog.*, **4**, e1000219.
34. Marchat,L.A., Orozco,E., Guillen,N., Weber,C. and Lopez-Camarillo,C. (2008) Putative DEAD and DExH-box RNA helicases families in *Entamoeba histolytica*. *Gene*, **424**, 1–10.
35. Abhyankar,M.M., Hochreiter,A.E., Hershey,J., Evans,C., Zhang,Y., Crasta,O., Sobral,B.W., Mann,B.J., Petri,W.A. Jr. and Gilchrist,C.A. (2008) Characterization of an *Entamoeba histolytica* high-mobility-group box protein induced during intestinal infection. *Eukaryot. Cell*, **7**, 1565–1572.
36. Gilchrist,C.A., Baba,D.J., Zhang,Y., Crasta,O., Evans,C., Caler,E., Sobral,B.W., Bousquet,C.B., Leo,M., Hochreiter,A. *et al.* (2008) Targets of the *Entamoeba histolytica* transcription factor URE3-BP. *PLoS Negl. Trop. Dis.*, **2**, e282.
37. Duerkop,B.A., Herman,J.P., Ulrich,R.L., Churchill,M.E. and Greenberg,E.P. (2008) The Burkholderia mallei BmaR3-BmaI3 quorum-sensing system produces and responds to N-3-hydroxy-octanoyl homoserine lactone. *J. Bacteriol.*, **190**, 5137–5141.
38. Majumder,S. and Lohia,A. (2008) *Entamoeba histolytica* encodes unique formins, a subset of which regulates DNA content and cell division. *Infect. Immunity*, **76**, 2368–2378.
39. Lopez-Camarillo,C., de la Luz Garcia-Hernandez,M., Marchat,L.A., Luna-Arias,J.P., Hernandez de la Cruz,O., Mendoza,L. and Orozco,E. (2008) *Entamoeba histolytica* EhDEAD1 is a conserved DEAD-box RNA helicase with ATPase and ATP-dependent RNA unwinding activities. *Gene*, **414**, 19–31.
40. Li,J. and McClane,B.A. (2008) A novel small acid soluble protein variant is important for spore resistance of most Clostridium perfringens food poisoning isolates. *PLoS Pathog.*, **4**, e1000056.
41. Lopez-Casamichana,M., Orozco,E., Marchat,L.A. and Lopez-Camarillo,C. (2008) Transcriptional profile of the homologous recombination machinery and characterization of the EhRAD51 recombinase in response to DNA damage in *Entamoeba histolytica*. *BMC Mol. Biol.*, **9**.
42. Jhingran,A., Padmanabhan,P.K., Singh,S., Anamika,K., Bakre,A.A., Bhattacharya,S., Bhattacharya,A., Srinivasan,N. and Madhubala,R. (2008) Characterization of the *Entamoeba histolytica* Ornithine Decarboxylase-Like Enzyme. *PLoS Negl. Trop. Dis.*, **2**, e115.
43. Whitlock,G.C., Estes,D.M. and Torres,A.G. (2007) Glanders: off to the races with Burkholderia mallei. *Fems Microbiol. Lett.*, **277**, 115–122.
44. Sun,J., Tuncay,K., Haidar,A.A., Ensman,L., Stanley,F., Trelinski,M. and Ortoleva,P. (2007) Transcriptional regulatory network discovery via multiple method integration: application to *E. coli* K12. *Algorithms Mol. Biol.*, **2**, 2.
45. Tiyawisutsri,R., Holden,M.T.G., Tumapa,S., Rengpipat,S., Clarke,S.R., Foster,S.J., Nierman,W.C., Day,N.P.J. and Peacock,S.J. (2007) Burkholderia Hep_Hap autotransporter (BuHA) proteins elicit a strong antibody response during experimental glanders but not human melioidosis. *BMC Microbiol.*, **7**, 19.
46. Vidal,J.E., Chen,J., Li,J. and McClane,B.A. (2009) Use of an EZ-Tn5-based random mutagenesis system to identify a novel toxin regulatory locus in *Clostridium perfringens* strain 13. *PLoS ONE*, **4**, e6232.