Expert Assertions Through Community Annotation Jamborees

Lauren Brinkac
The J. Craig Venter Institute



What is Pathema?

- A NIAID Bioinformatics Resource Center designed to support bio-defense and infectious disease research.
- Pathema provides detailed curation and comparative analysis of six target pathogens:
 - Category A priority pathogens:

Bacillus anthracis, Clostridium botulinum

Category B priority pathogens:

Burkholderia mallei, Burkholderia pseudomallei, Clostridium perfringens, Entamoeba histolytica



Community Comments

Annotation. Currently we have "raw data" accumulating 100-fold faster than it is being annotated. GenBank and most other resources do not allow an "expert" on one gene to annotate that gene in the dozens of new complete genomes that come out each week, even if the expert was motivated to do so.

The static annotation in CMR is a real problem. It would be nice to have a constantly updated genome based on current research in the field. An example is what Fiona Brinkman has done with the Pseudomonas aeruginosa genome website. The B. mallei ATCC 23344 genome annotation has numerous errors that have been identified by researchers in the field and no one will update this information in the CMR or GenBank. If you want to do something that will actually push the field forward, you should work o

all display of gene annotations are easily link to pathway informations, pubmed publication literature report of each gene characterisation-validation, gene microarray result for transcription datas eg latest submission data by Dr Patrick Tan's data on gene expression in different growth period, proteomics-mass spec data for protein expression-immunoreactive proteins and localisation of protein if there is any reported immunohisto microscopy work



Annotation Protocols



Clostridium Protocols

Submit Your Own Protocol

Standard Operational Procedures Biosafety Protocols Detection Protocols Epitope Protocols Reagent Links Safety Requirements

Standard Operational Procedures

The following is completed documentation of many of JCVI's standard operating procedures (SOPs) for gene model and functional curation. Each SOP outlines the purpose and scope of each process or procedure, all requirements necessary to carry out each procedure, a detailed description of the process or analysis, and all measures utilized to ensure that data and data types generated are consistent, current, and maintain specific quality requirements.

- · Gene Model Curation
 - Gene Prediction: Outlines the process for locating and predicting coding and non-coding genome features.

Gene Model Curation

- Analysis and Curation of Short Gene Models: Describes the process used by curators for evaluating potential false ositives
 generated by Glimmer.
- Creation of a Pseudomolecule: Describes the process of creating a "pseudomolecule" for an unfinished genomic sequence.
- Annotation mapping: Describes the procedure of mapping gene models and annotation from a reference genome to a related molecule.
- Homology Searches: Outlines all of our pre-computed homology searches run to generate evidence for functional annotation.
- Functional Curation
 - <u>Functional Automated Annotation</u>: Defines our automated annotation program that generates putative functional annotation to each gene model.
 - Functional Manual Curation: Documents the layered evidence based approach used by curators for accurately assigning
 descriptive functional annotation to each gene model.

Functional Curation

- Start Site Curation: Describes the criteria used by curators when evaluating and editing the initiation codon for each gene
 model.
- <u>Frameshift Edit and Analysis</u>: Describes the procedure used by curators for evaluating and editing potential frame shifts, point
 mutations, and sequence ambiguities.
- Overlap Analysis and Curation: Defines the criteria used by curators when reviewing overlapping regions of gene models.
- · Supporting Documentation
 - Naming Convention Guidelines

Homology Searches

Naming Convention Guidelines



JCVI Gene Naming and Annotation Conventions

All genomes sequenced at TIGR are initially assigned annotation through an automated process. Names and functional annotation are then manually curated. As direct experimental evidence rarely exists for each gene in a sequence direct assignments should be reparted as provisional. We strive to annotate each gene with as much information as we can confidently impart, but are also wary of inferring too much from sequence similarity. We prefer to err on the side of caution and we have devised a nomenclature scheme that reflects our degree of confidence in a particular assignment.

We encourage feedback from the community to help identify errors or to provide suggestions to improve the annotation of our genes.

Evidence Types

Levels of

Database Match

Information used during manual curation

· Pairwise search results

Protein translations of all genes are searched vs. a non-redundant amino acid database to generate a file of pair-wise alignments. Matches to experimentally characterized proteins are given special consideration.

HMM matches

Protein translations of all genes are searched against Hidden Markov Models (HMMs) built at TIGR (TIGRFAMs) and at <u>Sanger</u> (Pfams). HMMs are statistical models built from multiple alignments of proteins which share sequence similarity. TIGR classifies HMMs into more than a dozen "isology" types, each of which represents a different degree of confidence about function.

Paralogous families

Each predicted protein translation is searched against the complete protein set for a genome to identify protein families found within the organism.

Biologically significant motifs and sites

Protein translations of all genes are searched vs. PROSITE for biologically significant patterns. Potential transmembrane domains are predicted by <u>TmHIMM</u>. For enzymes, curators review active site information from matches in <u>SwissProt</u>, <u>MEROPS</u>, and other databases.

Gene context

Physical location of a gene within a gene cluster or putative operon can be significant in some assignments - particularly transporters or enzymes involved in biosynthetic or metabolic pathways.

· Genome Properties

Each of TIGR's <u>Genome Properties</u> comprises a suite of a genes that function in a known metabolic pathway, cellular activity, or cellular structure. Genes are evaluated by HMM matches and context-based rules, and assigned to the appropriate Genome Property.

Levels of Database Match

Descriptors for annotation include at minimum a common name, role category, and <u>Gene Ontology</u> (GO) function and 'process' terms for each gene, and may also include a gene symbol, an Enzyme Commission number, and public comments. Each gene is assigned as many descriptors as a rare relevant. In the course of reviewing data we have developed the following criteria regarding assignments.

Specific function: indicated by a specific name and gene symbol.

- The protein translation has a good database match to a protein that whose function and process have been experimentally characterized. Both pairwise and multiple protein sequence alignments reveal a high degree of identify/similarity (typically >35% identify) along the entire length of the protein. There may be an essentially full-length match to a highly specific (i.e., 'equivalog' isology type) HMM. Active sites, substrate or cofactor binding sites, or motifs that are characteristic of a protein should be conserved. Strong conservation of gene context (e.g. operon structure) is also taken as evidence for certain function. For genes with a certain function we use the most widely-recognized name and gene symbol. Highly specific GO function and process terms are used if available. Enzymes of certain function are annotated with their full IUBMB number; the IUBMB enzyme name may also be used for clarity if it seems more informative than its common name.

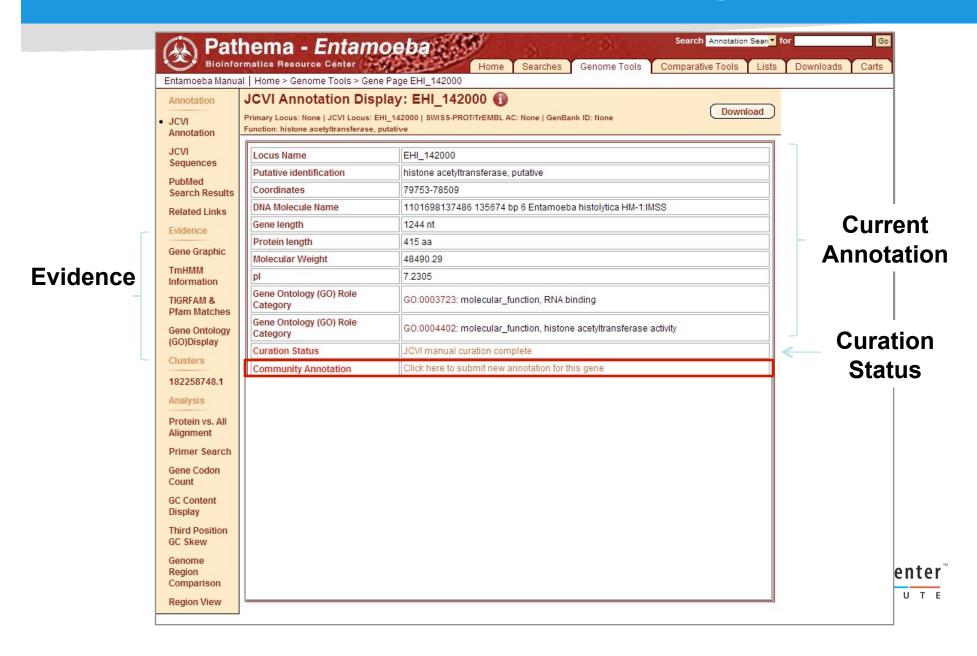
 Likely (or unlikely) function: indicated by 'putative' or 'putative
 - If one o'r more lines of evidence is weak, but most of the data agrees, we conclude the gene is likely performing the function the name implies, and the name is preceded with "putative". In such cases, the percent identity (e.g., 30-35% identity) or HMM score (score is between the trusted and noise cutoffs) is not quite high enough to impart certainty. GO terms may be more general than for gene models of certain function, and with few exceptions, gene symbols are not used. For an enzyme with a putative specific function partial IUBMB numbers may be used.
 - When there are strong lines of conflicting evidence, we consider the function indicated by the common name to be unlikely, and add 'homolog' to the common name. Such assignment can arise from two situations. In the first situation, sequence homology is very strong, but unlike a 'putative' match, we do NOT believe the query protein has the same function as the match. This might be because some critical piece of evidence is absent (e.g., non-conservation of catalytic residues in an enzyme), or because the function is not predicted to exist in this particular organism (e.g., photosynthetic enzyme matches in a non-photosynthetic organism). In the second situation, there is an essentially full-length match to a set of genes whose names are the same or similar, at least one of which has some experimental characterization, but because the sequence conservation (e.g., 25-30% identity) falls below even the 'putative' range, we consider functional conservation to be unlikely. Furthermore, there are no family or domain names available. In this case we use the matching proteins' name but add 'homolog' to it, and apply descriptors appropriate for a protein of unknown function.
 - Note that while using 'homolog' to denote non-conserved function in high-quality matches has been a long-time practice of TIGR annotators, using 'homolog' to retain the names of lower-quality matches that might otherwise be called 'conserved hypothetical proteins' is a relatively recent practice. Also, the criteria for 'putative' annotation has been made more rigorous. Therefore, it is likely that some older gene models that were called 'putative' or 'conserved hypothetical' would be called 'homolog' by the newer naming criteria.

Generic function: indicated by protein family name or domain name.

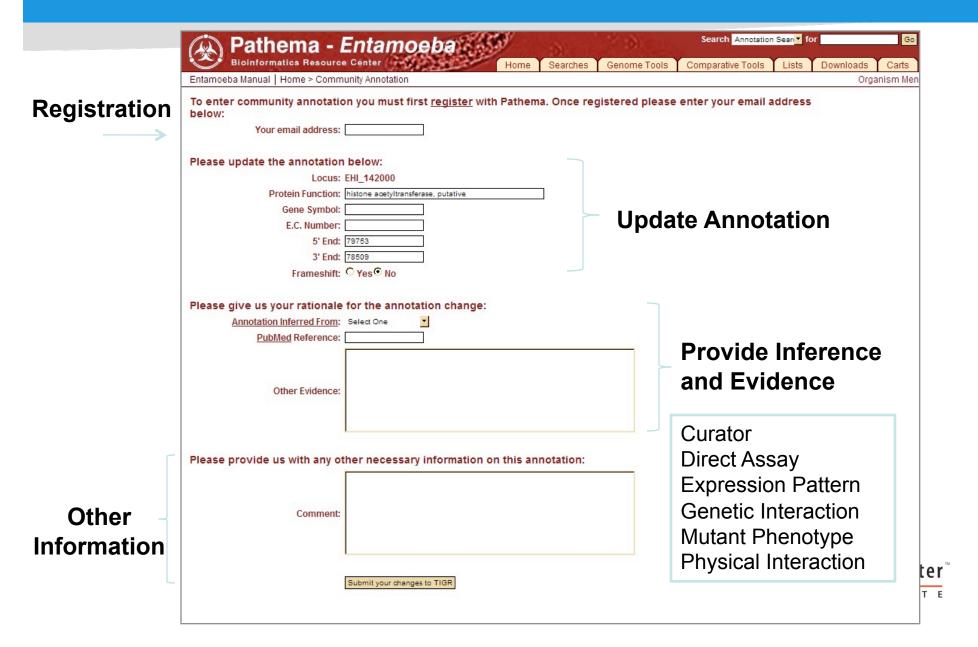
- When the best (or only) annotation evidence indicates membership in a defined family, but does not justify more specific naming, we use family names defined by a TIGR or Pfam HMM(s), curated databases such as SwissProt, or in the literature, e.g., "carbohydrate kinase, FGGY family".
- When the extent of sequence homology is limited to a defined protein domain (usually modelled as an HMM), rather than a defined family or full-length characterized protein, we may use the domain name, e.g., "ABC1 domain protein". Since domains are themselves often used to define a family in the literature, the distinction between family and domain based names is not rigid.



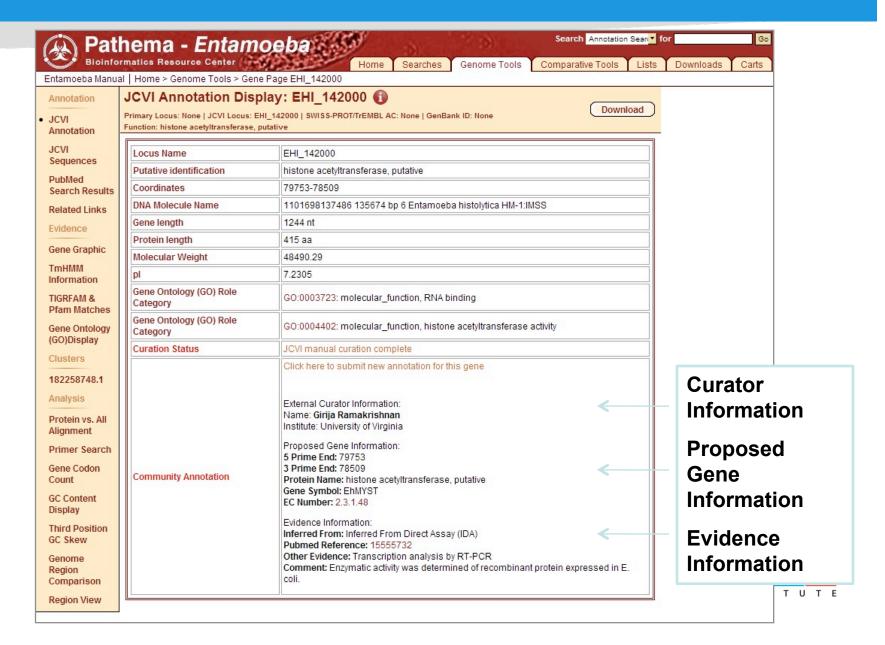
Pathema Gene Page



Web Community Submission



Community Annotation



Workshops and Training

On JCVI Campus

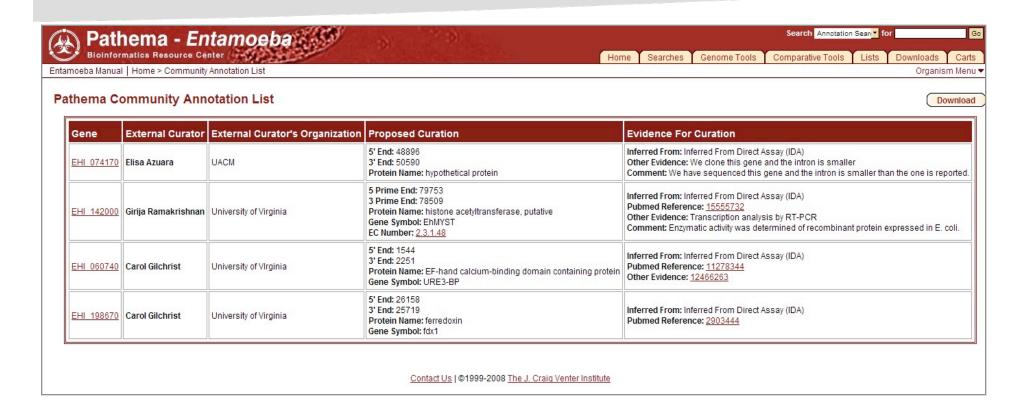
- 2 Days Hands On
- Live/Archived Webinar Broadcast
- 4 Workshops Conducted, one per clade

Organism Specific Meetings

- 4 Hours Hands On
- 5 Workshops Conducted
- 148 Researchers Attended
- 123 Community Registration



Web Community Submissions



Totals To Date:

4 Annotation Updates

- 1 gene structure 1 protein name
- 3 gene symbols 1 enzyme number
- 4 literature references



Advantages and Disadvantages

Advantages

- Value added to annotation
- Convenient & accessible for the researcher
- Minimal effort needed
- Low cost

Disadvantages

- Training not required
 - Lack of adherence to standards
- Incentive and intimidation



Annotation Jamboree Goals

- Incorporate expert annotation data into existing Pathema genome submissions.
- Update existing annotation based on expert assertions.
- Tag annotation with updated experimental references.
- Provide the community with the opportunity to become familiar with JCVI annotation procedures.

Annotation Jamboree Details

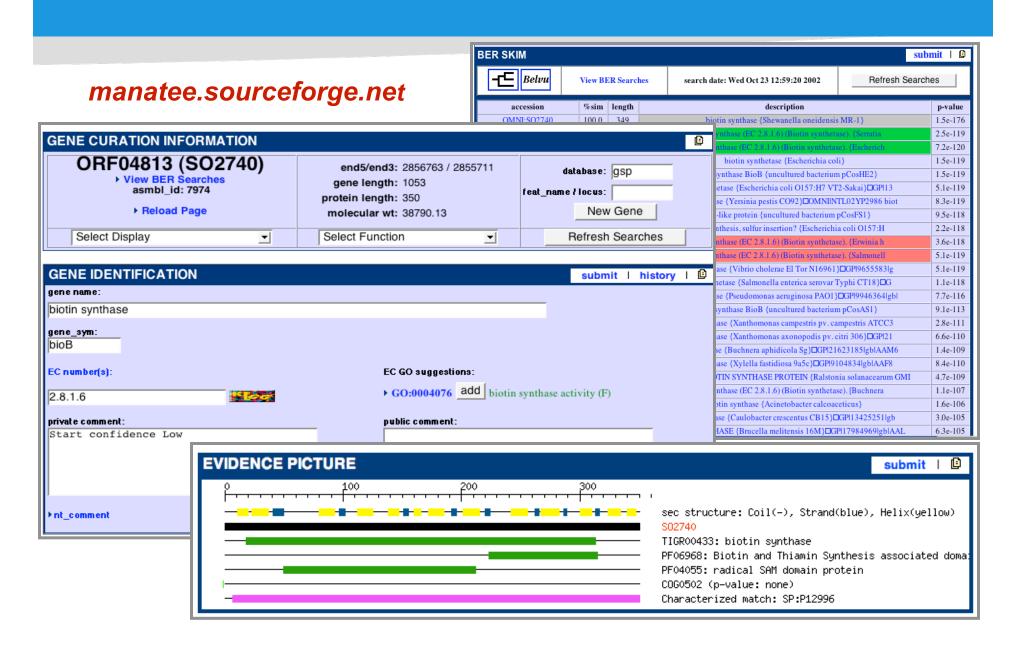
- 3 day Jamborees held at JCVI
- First day teaches our annotation methodologies and Manatee
- Follow up days allow the community to annotate with aid from our expert curators and present latest research on gene sets
- Data dissemination and participant recognition within Pathema and GenBank

Excellent participation and feedback from our

participants

J. Craig Venter

Manatee: Manual Annotation Tool



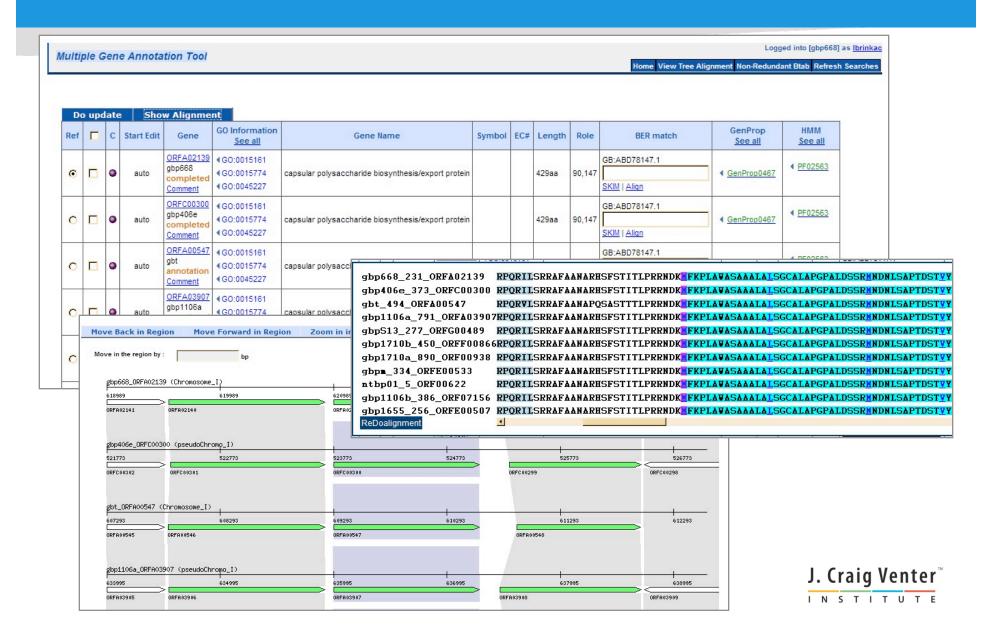
Evidence Standards

phosphomethylpyrimidine kinase

- Function was inferred from direct assay using the accession SP|P76422 with reference PMID: 10075431
- Process was inferred from direct assay using the accession SP|P76422 with reference PMID: 10075431
- putative hypoxanthine phosphoribosyltransferase
 - Function was inferred from sequence similarity using the accession TIGR01203
 - Process was inferred from mutant phenotype using the accession SP|Q02522 with reference PMID: 1465108



MGAT: Multi-Gene Annotation Tool



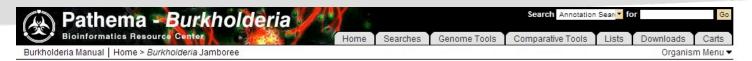
Paralogous Family Annotation

Change display options					
gene name	GO id	Select Action:	Sort options: By aa length Intron options: Full length 1	HMM Show all Show No HMMs	Para domains Show all
acetyltransferase, GNAT family	GO:0008152 (ISS)	3766.m00052 EHI_133980 [GC ED GO] SC: N AC: Y CM: N		PF00583 : acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	4148.m00011 EHI_186730 [GC ED GO] SC: N AC: Y CM: N		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	3806.m00002 EHI_141360 [GC ED GO] SC: N AC: Y CM: N		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	3677.m00003 EHI_123240 [GC ED GO] SC: N AC: Y CM: N		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	3878.m00025 EHI_151490 [GC ED GO] SC: N AC: Y CM: N		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	4213.m00019 EHI_197250 [GC ED GO] SC: N AC: Y CM: N		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	4150.m00007 EHI_187730 [GC ED GO]		PF00583: acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	3218.m00010 EHI_065290 [GC ED GO]		PF00583 : acetyltransferase, GNAT family [R] [S]	2848 : PF00583.fasta.msf [A]
		SC: N AC: Y CM: N			

Annotation Jamboree Results

Jamboree Statistics	Entamoeba	Burkholderia	Total
Number of Participants	14	9	
Community Researchers	6	6	23
JCVI Analysts	8	3	
Curated Functional Assignments	1022	543	
Assignments made by commuity researchers	595	41	1565
Assignments made by JCVI based on community contributions	n/a	404	1505
Assignments made by JCVI	427	98	
GO Terms Assigned	2019	1480	
Assignments made by commuity researchers	123	0	3499
Assignments made by JCVI based on community contributions	500	1198	0.100
Assignments made by JCVI	1396	282	
Paralogous Families Curated	120	n/a	120
Gene Structures Edited	43	86	
Assignments made by JCVI based on community contributions	43	69	129
Assignments made by JCVI	0	17	
Experimental References Added	285	11	296
Number of Genome Projects Updated	1	10	11
Number of Post Collaborations Established	1	3	4

Jamboree Statistics



Burkholderia Jamboree

A three-day, hands-on Burkholderia Annotation Jamboree was conducted in an effort to train community experts in the Burkholderia field on JCVI's internal tools for structural and functional annotation. The jamboree was held on the JCVI campus, Sept 24, 2008 - Sept 26, 2008. Training consisted of a series of lectures describing JCVI's prokaryotic annotation methodologies and functional annotation tool, Manatee. Additionally, community researchers participated in a series of instructional hands-on annotation training exercises. After completing the training portion of the jamboree, community researchers subsequently contributed to the manual curation of Burkholderia genome projects.

JAMBOREE STATISTICS

Total Jamboree Contributions

Participants	Functional Names Assigned	Go Terms Assigned	Gene Structural Curation	Literature References	Burkholderia genomes updated
9	502	1480	86	11	10

Total

Participants

Burkholderia Researchers 6 JCVI Analysts 3

Participants

Functional Names

Assignments made by Burkholderia Researchers 41
Assignments made by JCVI based on community contributions 404
Assignments made by JCVI 98

GO Terms

Assignments made by JCVI based on community contributions 1198
Assignments made by JCVI 282

Gene Structural Curation

Edited gene structures by JCVI based on community contributions 69
Edits made by JCVI 17

Literature References

Functional Assignments with associated experimental references 1

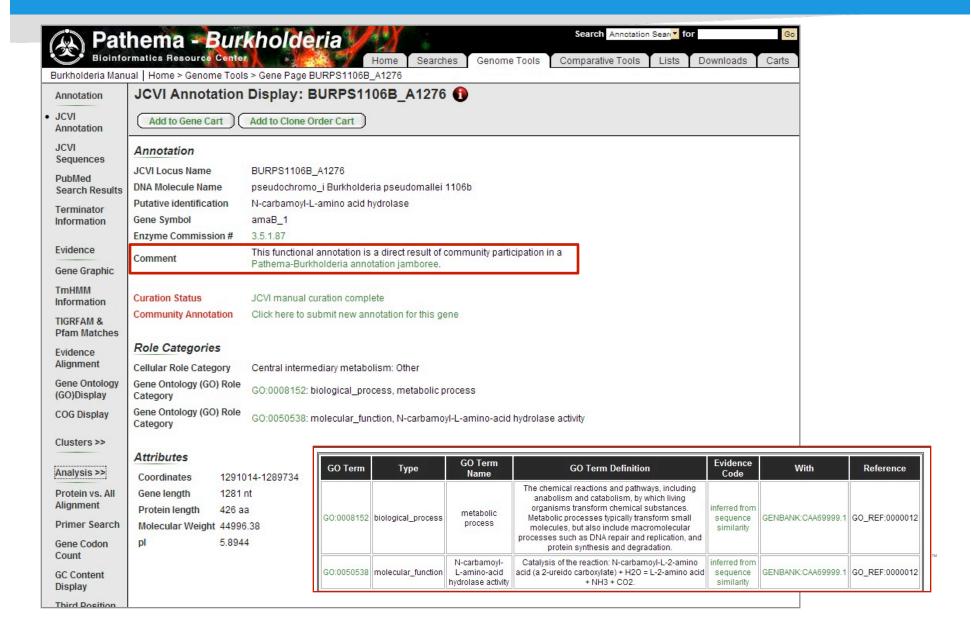
Functional Names

Go Terms

Structural Curation

Literature References

Gene Page



GenBank Updates

LOCUS	ABN89094 178 aa linear BCT 03-APR-2009					
DEFINITION	shock protein HslVU, ATP-dependent protease subunit HslV					
	[Burkholderia pseudomallei 1106a].					
ACCESSION	ABN89094					
VERSION	ABN89094.2 GI:210148305					
DBSOURCE	accession CP000572.1					
KEYWORDS						
SOURCE	Burkholderia pseudomallei 1106a					
ORGANISM	Burkholderia pseudomallei 1106a					
	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales;					
	Burkholderiaceae; Burkholderia; pseudomallei group.					
REFERENCE	1 (residues 1 to 178)					
AUTHORS	Harkins, D.M., Brinkac, L.M., Brown, K.A., Hung, G.C., Tuanyok, A.,					
	Zhang, B. and Nierman, W.C.					
TITLE	Revision of Burkholderia pseudomallei 1106a					
JOURNAL	Unpublished					
REFERENCE	2 (residues 1 to 178)					
AUTHORS	DeShazer, D., Woods, D.E. and Nierman, W.C.					
TITLE	Direct Submission					
JOURNAL	Submitted (13-FEB-2007) The Institute for Genomic Research, 9712					
	Medical Center Dr, Rockville, MD 20850, USA					
REFERENCE	3 (residues 1 to 178)					
AUTHORS	Harkins, D.M.					
TITLE	Direct Submission					
JOURNAL	Submitted (30-OCT-2008) The Institute for Genomic Research, 9712					
	Medical Center Dr, Rockville, MD 20850, USA					
REMARK	Protein update by submitter					
REFERENCE	4 (residues 1 to 178)					
AUTHORS	Harkins, D.M.					
TITLE	Direct Submission					
JOURNAL	Submitted (18-NOV-2008) The Institute for Genomic Research, 9712					
	Medical Center Dr, Rockville, MD 20850, USA					
REMARK	Protein update by submitter					
COMMENT	On Oct 31, 2008 this sequence version replaced gi:126225554.					
	Source DNA is available from BEIresources					
	(http://www.beiresources.org/).					
	This functional annotation is the result of community participation					
	in the Pathema-Burkholderia annotation jamboree held at the J.					
	Craig Venter Institute, September 24-26, 2008					
	(http://pathema.jcvi.org). Participating institutions for the					
	genome revision: The J. Craig Venter Institute, University of Texas					
	at Austin, Imperial College London, Armed Forces Institute of					
	Pathology, and Northern Arizona University.					

- Contributors become authors on submission updates
- Annotation jamboree & participating institutions acknowledged in comments
- db_xrefs to gene page established.



Long Term Benefits

- Continued collaborations
 - 4 Community Researchers
- Expert assertions
 - 20 additional assertions / 2 months
 - Curation of genomic islands and SSRs
- Dissemination of annotation standards
- Custom analysis and data generation



Advantages and Disadvantages

Advantages

- Value added to annotation
- In depth hands on training
 - Adherence to standards
- High incentive and lower intimidation
- Long term benefits

Disadvantages

- Large cost
- Time commitment



Next Annotation Jamboree

- Pathema Bacillus/Clostridium Annotation Jamboree
 - When: May 13-15, 2009
 - Cost: Free
 - Update annotation with expert assertions
 - 19 Bacillus anthracis
 - 11 Bacillus cereus
 - 12 Clostridium botulinum
 - 9 Clostridium perfringens
- More information: http://pathema.jcvi.org

Acknowledgments

PI: Granger Sutton

Informatics Engineers

Tanja Davidsen (manager)

Erin Beck

Kevin Galinsky

Jason Inman

Alex Richter

Bob Montgomery

Informatics Analysts

Scott Durkin (manager)

Ramana Madupu

Bob Dodson

Derek Harkins

Susmita Shrivastava

Lis Caler

All others past & present who contributed to the development and data analyses of this resource.

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.

J. Craig Venter HHSN266200400038C