

# 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.
  - [Analysis and Curation of Short Gene Models](#): Describes the process used by curators for evaluating potential false positives 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](#)
  - [Functional Automated Annotation](#): 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.
  - [Start Site Curation](#): Describes the criteria used by curators when evaluating and editing the initiation codon for each gene model.
  - [Frameshift Edit and Analysis](#): 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](#)

**Gene Model  
Curation**

**Functional  
Curation**

**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 sequenced organism, name assignments are usually based on sequence similarity. Therefore, all TIGR name assignments should be regarded 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.

### 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 Sanger (Pfam). 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 TmHMM. For enzymes, curators review active site information from matches in SwissProt, MEROPS, 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 Genome Properties 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.

## Evidence Types

## Levels of Database Match

### Levels of Database Match

Descriptors for annotation include at minimum a common name, role category, and Gene Ontology (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 are 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 identity/similarity (typically >35% identity) 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 'homolog' in the name

- If one or 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

**Pathema - Entamoeba**  
Bioinformatics Resource Center

Search  Annotation Search for  Go

Home Searches Genome Tools Comparative Tools Lists Downloads Carts

Entamoeba Manual | Home > Genome Tools > Gene Page EHI\_142000

**JCVI Annotation Display: EHI\_142000**  [Download](#)

Primary Locus: None | JCVI Locus: EHI\_142000 | SWISS-PROT/TrEMBL AC: None | GenBank ID: None  
Function: histone acetyltransferase, putative

Locus Name	EHI_142000
Putative identification	histone acetyltransferase, putative
Coordinates	79753-78509
DNA Molecule Name	1101698137486 135674 bp 6 Entamoeba histolytica HM-1:IMSS
Gene length	1244 nt
Protein length	415 aa
Molecular Weight	48490.29
pI	7.2305
Gene Ontology (GO) Role Category	GO:0003723: molecular_function, RNA binding
Gene Ontology (GO) Role Category	GO:0004402: molecular_function, histone acetyltransferase activity
Curation Status	JCVI manual curation complete
Community Annotation	<a href="#">Click here to submit new annotation for this gene</a>

182258748.1

Protein vs. All Alignment  
Primer Search  
Gene Codon Count  
GC Content Display  
Third Position GC Skew  
Genome Region Comparison  
Region View

enter™  
U T E

Evidence

Current Annotation

Curation Status

# Web Community Submission

Registration



**Pathema - Entamoeba**  
Bioinformatics Resource Center

Search  Annotation Search  for

Home Searches Genome Tools Comparative Tools Lists Downloads Carts

Entamoeba Manual | Home > Community Annotation Organism Men

To enter community annotation you must first [register](#) with Pathema. Once registered please enter your email address below:

Your email address:

Please update the annotation below:

Locus: EHI\_142000

Protein Function:

Gene Symbol:

E.C. Number:

5' End:

3' End:

Frameshift:  Yes  No

Please give us your rationale for the annotation change:

Annotation Inferred From:

PubMed Reference:

Other Evidence:

Please provide us with any other necessary information on this annotation:

Comment:

Update Annotation

Provide Inference and Evidence

Curator  
Direct Assay  
Expression Pattern  
Genetic Interaction  
Mutant Phenotype  
Physical Interaction

Other Information




# Community Annotation

**Pathema - Entamoeba**  
Bioinformatics Resource Center

Search  Annotation Search for  Go

Home Searches Genome Tools Comparative Tools Lists Downloads Carts

Entamoeba Manual | Home > Genome Tools > Gene Page EHI\_142000

**JCVI Annotation Display: EHI\_142000**  [Download](#)

Primary Locus: None | JCVI Locus: EHI\_142000 | SWISS-PROT/TrEMBL AC: None | GenBank ID: None  
Function: histone acetyltransferase, putative

Locus Name	EHI_142000
Putative identification	histone acetyltransferase, putative
Coordinates	79753-78509
DNA Molecule Name	1101698137486 135674 bp 6 Entamoeba histolytica HM-1:IMSS
Gene length	1244 nt
Protein length	415 aa
Molecular Weight	48490.29
pI	7.2305
Gene Ontology (GO) Role Category	GO:0003723: molecular_function, RNA binding
Gene Ontology (GO) Role Category	GO:0004402: molecular_function, histone acetyltransferase activity
Curation Status	JCVI manual curation complete

[Click here to submit new annotation for this gene](#)

**Community Annotation**

External Curator Information:  
Name: **Girija Ramakrishnan**  
Institute: University of Virginia

Proposed Gene Information:  
5 Prime End: 79753  
3 Prime End: 78509  
Protein Name: histone acetyltransferase, putative  
Gene Symbol: EhMYST  
EC Number: 2.3.1.48

Evidence Information:  
Inferred From: Inferred From Direct Assay (IDA)  
Pubmed Reference: 15555732  
Other Evidence: Transcription analysis by RT-PCR  
Comment: Enzymatic activity was determined of recombinant protein expressed in E. coli.

**Curator Information**

**Proposed Gene Information**

**Evidence Information**

T U T E



# 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

Pathema - Entamoeba  
Bioinformatics Resource Center

Search Annotation Search for  Go

Home Searches Genome Tools Comparative Tools Lists Downloads Carts

Entamoeba Manual | Home > Community Annotation List Organism Menu

### Pathema Community Annotation List Download

Gene	External Curator	External Curator's Organization	Proposed Curation	Evidence For Curation
<a href="#">EHI_074170</a>	Elisa Azuara	UACM	5' End: 48896 3' End: 50590 Protein Name: hypothetical protein	Inferred From: Inferred From Direct Assay (IDA) Other Evidence: We clone this gene and the intron is smaller Comment: We have sequenced this gene and the intron is smaller than the one is reported.
<a href="#">EHI_142000</a>	Girija Ramakrishnan	University of Virginia	5 Prime End: 79753 3 Prime End: 78509 Protein Name: histone acetyltransferase, putative Gene Symbol: EhMYST EC Number: <a href="#">2.3.1.48</a>	Inferred From: Inferred From Direct Assay (IDA) Pubmed Reference: <a href="#">15555732</a> Other Evidence: Transcription analysis by RT-PCR Comment: Enzymatic activity was determined of recombinant protein expressed in E. coli.
<a href="#">EHI_060740</a>	Carol Gilchrist	University of Virginia	5' End: 1544 3' End: 2251 Protein Name: EF-hand calcium-binding domain containing protein Gene Symbol: URE3-BP	Inferred From: Inferred From Direct Assay (IDA) Pubmed Reference: <a href="#">11278344</a> Other Evidence: <a href="#">12466263</a>
<a href="#">EHI_198670</a>	Carol Gilchrist	University of Virginia	5' End: 26158 3' End: 25719 Protein Name: ferredoxin Gene Symbol: fdx1	Inferred From: Inferred From Direct Assay (IDA) Pubmed Reference: <a href="#">2903444</a>

[Contact Us](#) | ©1999-2008 The J. Craig Venter Institute

## Totals To Date:

4 Annotation Updates

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

# 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*

# Manatee: Manual Annotation Tool

[manatee.sourceforge.net](http://manatee.sourceforge.net)

BER SKIM submit |

[View BER Searches](#) search date: Wed Oct 23 12:59:20 2002

accession	%sim	length	description	p-value
OMNI-SO2740	100.0	349	biotin synthase (Shewanella oneidensis MR-1)	1.5e-176
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Serratia	2.5e-119
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Escherich	7.2e-120
			biotin synthetase (Escherichia coli)	1.5e-119
			ynthase BioB {uncultured bacterium pCosHE2}	1.5e-119
			etase (Escherichia coli O157:H7 VT2-Sakai)CGP13	5.1e-119
			se {Yersinia pestis CO92}COMNIINTL02YP2986 biot	8.3e-119
			-like protein {uncultured bacterium pCosFS1}	9.5e-118
			ntesis, sulfur insertion? (Escherichia coli O157:H	2.2e-118
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Erwinia h	3.6e-118
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Salmonell	5.1e-119
			ase (Vibrio cholerae El Tor N16961)CGP9655583lg	5.1e-119
			etase (Salmonella enterica serovar Typhi CT18)CG	1.1e-118
			se (Pseudomonas aeruginosa PAO1)CGP9946364lgbI	7.7e-116
			ynthase BioB {uncultured bacterium pCosAS1}	9.1e-113
			ase (Xanthomonas campestris pv. campestris ATCC3	2.8e-111
			ase (Xanthomonas axonopodis pv. citri 306)CGP21	6.6e-110
			se (Buchnera aphidicola Sg)CGP21623185lgbIAAM6	1.4e-109
			ase (Xylella fastidiosa 9a5c)CGP9104834lgbIAAF8	8.4e-110
			TIN SYNTHASE PROTEIN (Ralstonia solanacearum GMI	4.7e-109
			ynthase (EC 2.8.1.6) (Biotin synthetase), (Buchnera	1.1e-107
			otin synthase (Acinetobacter calcoaceticus)	1.6e-106
			ase (Caulobacter crescentus CB15)CGP13425251lgb	3.0e-105
			IASE (Brucella melitensis 16M)CGP17984969lgbAAL	6.3e-105

### GENE CURATION INFORMATION

**ORF04813 (SO2740)**

[View BER Searches](#)  
asmbI\_id: 7974

[Reload Page](#)

Select Display

end5/end3: 2856763 / 2855711  
gene length: 1053  
protein length: 350  
molecular wt: 38790.13

database:   
feat\_name / locus:

### GENE IDENTIFICATION

submit | history |

gene name:

gene\_sym:

EC number(s):

EC GO suggestions: [GO:0004076](#)  biotin synthase activity (F)

private comment:

public comment:

[nt\\_comment](#)

### EVIDENCE PICTURE

submit |

sec structure: Coil(-), Strand(blue), Helix(yellow)

S02740

TIGR00433: biotin synthase

PF06968: Biotin and Thiamin Synthesis associated domain

PF04055: radical SAM domain protein

COG0502 (p-value: none)

Characterized match: SP:P12996

# 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

Multiple Gene Annotation Tool Logged into [gbp668] as [lbrinkae](#)

[Home](#)
[View Tree](#)
[Alignment](#)
[Non-Redundant Btab](#)
[Refresh Searches](#)

Do update		Show Alignment		Gene	GO Information	Gene Name	Symbol	EC#	Length	Role	BER match	GenProp	HMM
Ref	<input type="checkbox"/>	C	Start Edit		<a href="#">See all</a>							<a href="#">See all</a>	<a href="#">See all</a>
<input checked="" type="radio"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	auto	<a href="#">ORFA02139</a> gbp668 <b>completed</b> <a href="#">Comment</a>	<a href="#">GO:0015161</a> <a href="#">GO:0015774</a> <a href="#">GO:0045227</a>	capsular polysaccharide biosynthesis/export protein			429aa	90,147	GB:ABD78147.1 <input type="text"/> <a href="#">SKIM</a>   <a href="#">Align</a>	<a href="#">GenProp0467</a>	<a href="#">PF02563</a>
<input type="radio"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	auto	<a href="#">ORFC00300</a> gbp406e <b>completed</b> <a href="#">Comment</a>	<a href="#">GO:0015161</a> <a href="#">GO:0015774</a> <a href="#">GO:0045227</a>	capsular polysaccharide biosynthesis/export protein			429aa	90,147	GB:ABD78147.1 <input type="text"/> <a href="#">SKIM</a>   <a href="#">Align</a>	<a href="#">GenProp0467</a>	<a href="#">PF02563</a>
<input type="radio"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	auto	<a href="#">ORFA00547</a> gbt <b>annotation</b> <a href="#">Comment</a>	<a href="#">GO:0015161</a> <a href="#">GO:0015774</a> <a href="#">GO:0045227</a>	capsular polysaccharide biosynthesis/export protein					GB:ABD78147.1 <input type="text"/> <a href="#">SKIM</a>   <a href="#">Align</a>		<a href="#">PF02563</a>
<input type="radio"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	auto	<a href="#">ORFA03907</a> gbp1106a	<a href="#">GO:0015161</a> <a href="#">GO:0015774</a>	capsular polysaccharide biosynthesis/export protein					GB:ABD78147.1 <input type="text"/> <a href="#">SKIM</a>   <a href="#">Align</a>		

[Move Back in Region](#)
[Move Forward in Region](#)
[Zoom in](#)

Move in the region by:  bp

**gbp668\_231\_ORFA02139** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp406e\_373\_ORFC00300** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbt\_494\_ORFA00547** RPQRVLSRRFAAANAPQSASTTTTLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp1106a\_791\_ORFA03907** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp513\_277\_ORFG00489** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp1710b\_450\_ORFF00866** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp1710a\_890\_ORF00938** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbpa\_334\_ORFE00533** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**ntbp01\_5\_ORF00622** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp1106b\_386\_ORF07156** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY  
**gbp1655\_256\_ORFE00507** RPQRILSRRFAAANARHSFSTITLPRRNDKFKPLAVASAAAALASGCALAPGPALDSSRMNDNLSAPT DSTVY

[ReDoalignment](#)



# Paralogous Family Annotation

[Change display options](#)

gene name	GO id	Select Action:	Sort options: By aa length Intron options: Full length	HMM Show all	Para domains Show all
acetyltransferase, GNAT family	GO:0008152 (ISS)	<input type="checkbox"/> 3766.m00052 EHI_133980 [ GC   ED   GO ] SC: N AC: Y CM: N		<input type="checkbox"/> Show No HMMs PF00583 : acetyltransferase, GNAT family [R]   [S]	2848 : PF00583.fasta.msf [A]
acetyltransferase, GNAT family	GO:0008080 (ISS) GO:0008152 (ISS)	<input type="checkbox"/> 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)	<input type="checkbox"/> 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)	<input type="checkbox"/> 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)	<input type="checkbox"/> 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)	<input type="checkbox"/> 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)	<input type="checkbox"/> 4150.m00007 EHI_187730 [ 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)	<input type="checkbox"/> 3218.m00010 EHI_065290 [ GC   ED   GO ] SC: N AC: Y CM: N		PF00583 : acetyltransferase, GNAT family [R]   [S]	2848 : PF00583.fasta.msf [A]

# Annotation Jamboree Results

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

# 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

**Functional Names**

#### GO Terms

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

**Go Terms**

#### Gene Structural Curation

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


**Structural Curation**

#### Literature References

Functional Assignments with associated experimental references	11
--	----

**Literature References**

# Gene Page

 **Pathema - Burkholderia**  
Bioinformatics Resource Center
Search  Annotation Search for

Burkholderia Manual | Home > Genome Tools > Gene Page BURPS1106B\_A1276

Annotation

- JCVI Annotation
- JCVI Sequences
- PubMed Search Results
- Terminator Information
- Evidence
- Gene Graphic
- TmHMM Information
- TIGRFAM & Pfam Matches
- Evidence Alignment
- Gene Ontology (GO) Display
- COG Display
- Clusters >>
- Analysis >>**
- Protein vs. All Alignment
- Primer Search
- Gene Codon Count
- GC Content Display
- Third Position

## JCVI Annotation Display: BURPS1106B\_A1276 i

**Annotation**

JCVI Locus Name      BURPS1106B\_A1276

DNA Molecule Name    pseudochromo\_j Burkholderia pseudomallei 1106b

Putative identification    N-carbamoyl-L-amino acid hydrolase

Gene Symbol            amaB\_1

Enzyme Commission #    3.5.1.87

**Comment**            This functional annotation is a direct result of community participation in a Pathema-Burkholderia annotation jamboree.

**Curation Status**        JCVI manual curation complete

**Community Annotation**    [Click here to submit new annotation for this gene](#)

**Role Categories**

Cellular Role Category    Central intermediary metabolism: Other

Gene Ontology (GO) Role Category    GO:0008152: biological\_process, metabolic process

Gene Ontology (GO) Role Category    GO:0050538: molecular\_function, N-carbamoyl-L-amino-acid hydrolase activity

**Attributes**

Coordinates            1291014-1289734

Gene length            1281 nt

Protein length        426 aa

Molecular Weight    44996.38

pI                        5.8944

GO Term	Type	GO Term Name	GO Term Definition	Evidence Code	With	Reference
GO:0008152	biological_process	metabolic process	The chemical reactions and pathways, including anabolism and catabolism, by which living organisms transform chemical substances. Metabolic processes typically transform small molecules, but also include macromolecular processes such as DNA repair and replication, and protein synthesis and degradation.	inferred from sequence similarity	GENBANK:CAA69999.1	GO_REF:0000012
GO:0050538	molecular_function	N-carbamoyl-L-amino-acid hydrolase activity	Catalysis of the reaction: N-carbamoyl-L-2-amino acid (a 2-ureido carboxylate) + H2O = L-2-amino acid + NH3 + CO2.	inferred from sequence similarity	GENBANK:CAA69999.1	GO_REF:0000012



# 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. HHSN266200400038C**