Pathema Annotation

Pathema Focus

Data types

- Annotation
  - ORF management
  - Gene features/GO assignments
  - Biochemical pathways
  - Subsystems/properties
  - CV-terms
- Comparative analysis
- SNPs and Assemblies for all category A-C organisms

Pathema Focus

- Collaborators
- Instructional Classes
  - Annotation/SOPs
  - Software
  - Web usage
- Open Source Software Support
  - Testing
  - Packaging
  - DB-angnostic
  - Often API-based

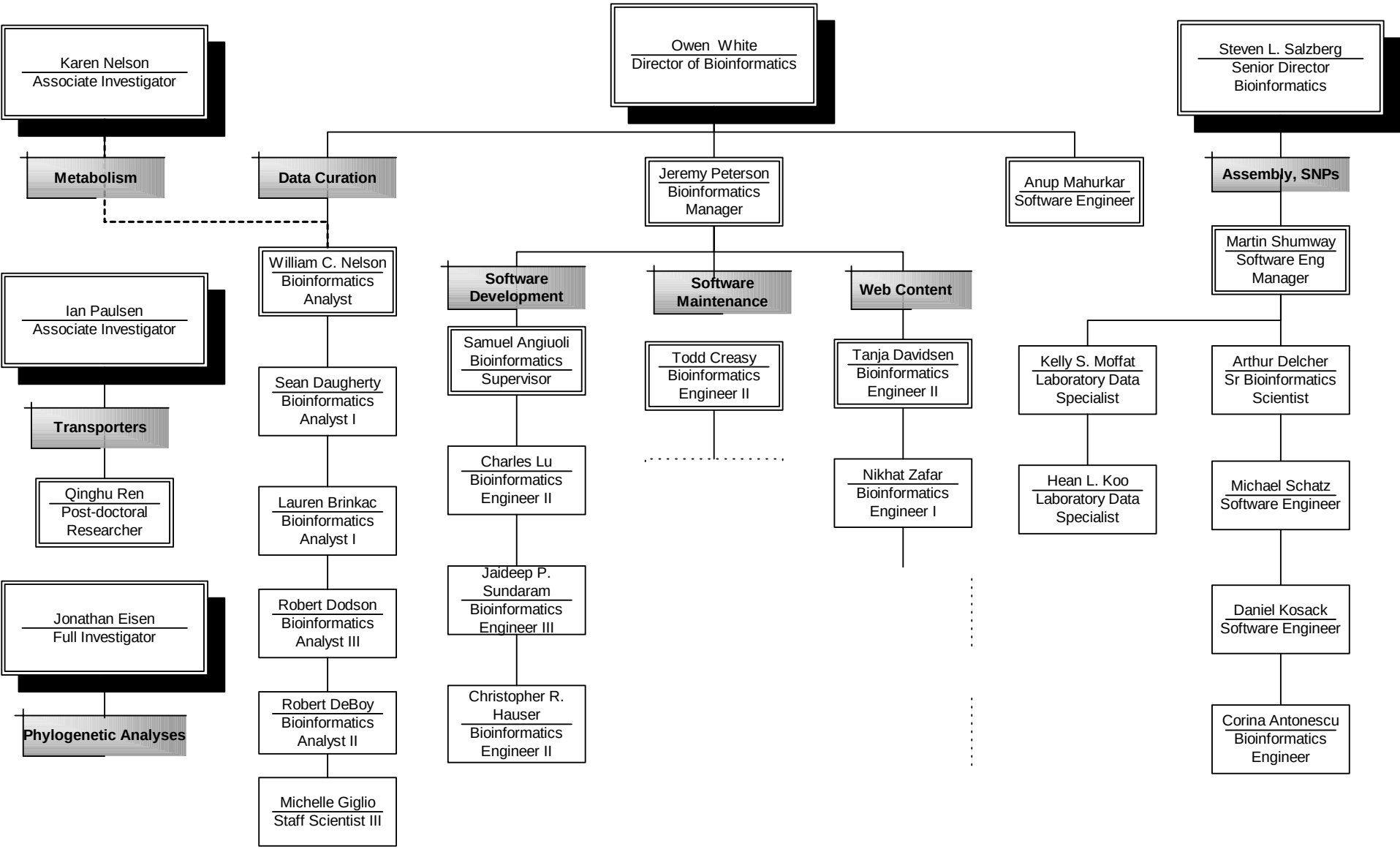
~~Datatypes~~

Biodefense-relevant genes
aka “system”

- transcriptionally altered genes from expression studies
- vaccine candidates
- outer membrane proteins
- pathogenicity plasmids
- extracellular proteases

TIGR/BRC Organisms

- *Bacillus anthracis*
- *Burkholderia mallei*
- *Burkholderia pseudomallei*
- *Clostridium botulinum*
- *Clostridium perfringens*
- *Entamoeba histolytica*


Owen White
Director of Bioinformatics
Karen Nelson
Associate Investigator
Metabolism
Ian Paulsen
Associate Investigator
Transporters
Qinghu Ren
Post-doctoral Researcher
Jonathan Eisen
Full Investigator
Phylogenetic Analyses
Data Curation
William C. Nelson
Bioinformatics Analyst
Sean Daugherty
Bioinformatics Analyst I
Lauren Brinkac
Bioinformatics Analyst I
Robert Dodson
Bioinformatics Analyst III
Robert DeBoy
Bioinformatics Analyst II
Michelle Giglio
Staff Scientist III
Jeremy Peterson
Bioinformatics Manager
Software Development
Samuel Angiuoli
Bioinformatics Supervisor
Charles Lu
Bioinformatics Engineer II
Jaideep P. Sundaram
Bioinformatics Engineer III
Christopher R. Hauser
Bioinformatics Engineer II
Software Maintenance
Todd Creasy
Bioinformatics Engineer II
Web Content
Tanja Davidsen
Bioinformatics Engineer II
Nikhat Zafar
Bioinformatics Engineer I
Anup Mahurkar
Software Engineer
Steven L. Salzberg
Senior Director Bioinformatics
Assembly, SNPs
Martin Shumway
Software Eng Manager
Kelly S. Moffat
Laboratory Data Specialist
Hean L. Koo
Laboratory Data Specialist
Arthur Delcher
Sr Bioinformatics Scientist
Michael Schatz
Software Engineer
Daniel Kosack
Software Engineer
Corina Antonescu
Bioinformatics Engineer

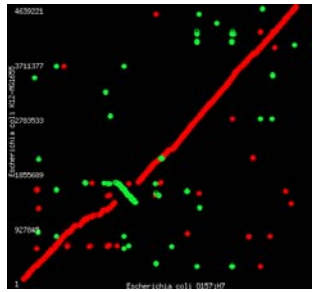
Software & Services

- Mummer - whole genome alignment. E,B,V
- Glimmer - gene finding system. B
- TIGRFams – orthologous families B
- Chado database – web presence and comparative analysis E,B,V
- Manatee - manual annotation tool. E,B,V
- Workflow - custom pipelines. E,B,V
- Sybil - comparative analysis system. E,B,V
- Assembly Service E,B,V

Help Wanted

- Better position effect/linear context
- Moving Genbank annotation to another sequence
- Lit. curation
- Sub-systems

Mummer 3.0



- MUMs: Maximal Unique Matches
  - Algorithm finds all matches
  - String them together and align gaps
- Suffix trees
  - Fast alignment of long DNA sequences
  - Linear time and space requirements
  - Streaming algorithm
- Memory maximization
  - 2 year dev time optimizing suffix tree impl.

Mummer Performance

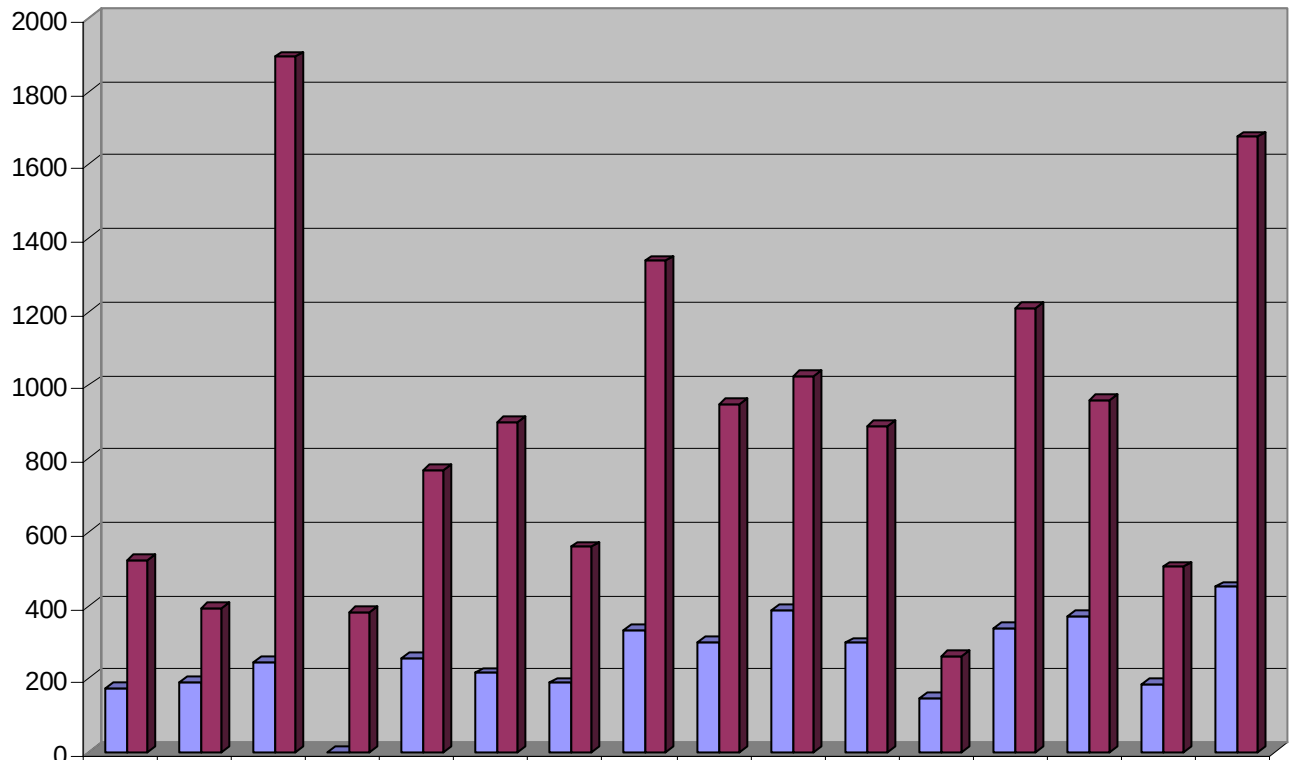
On a 2.4 GHz Pentium PC running Linux:

	3.0		2.0	
	Mb	Time	Mb	Time
E. coli K12 vs. E. coli O157	78	13s	102	14s
P. falciparum all chromosomes vs. P. yoelii	552	16:15	752	18:29
D. melanogaster arm 2L vs. D. pseudoobscura all contigs	467	13:55	465	14:39

TIGRFams

- Heavily curated multiple alignments based on protein families of the same function.
- Proposed “cure” for transitive annotation.
- Based on Hidden Markov Models (HMMs).
- >1,000 families.
- Complete assignments to GO.
- Cutoff scores for each family.
  - Trusted (automated name assignment)
  - Noise (manual inspection required)
- Downloadable. Fully integrated into Interpro.

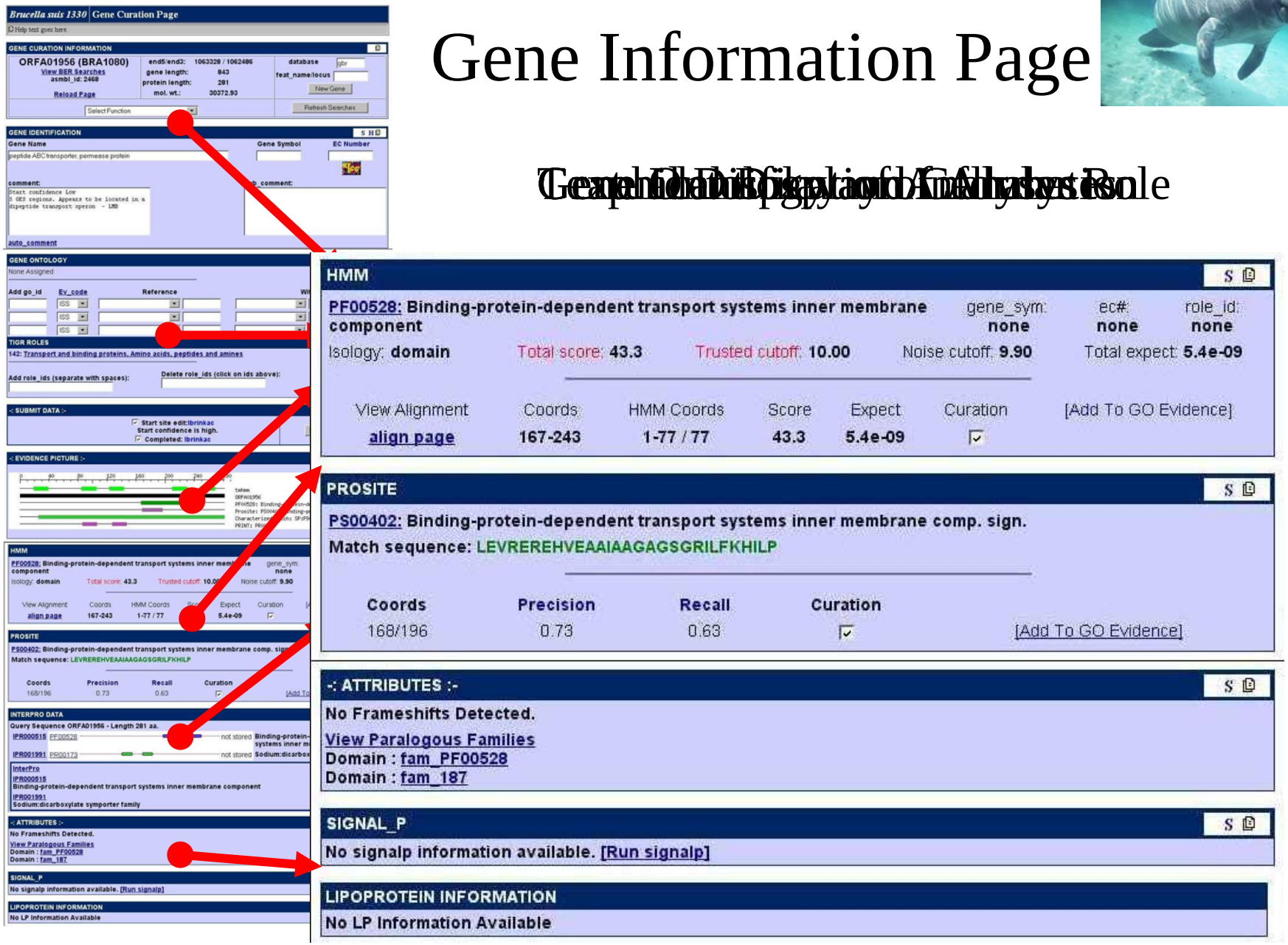
TIGRFAMs: **Orthologous Families**



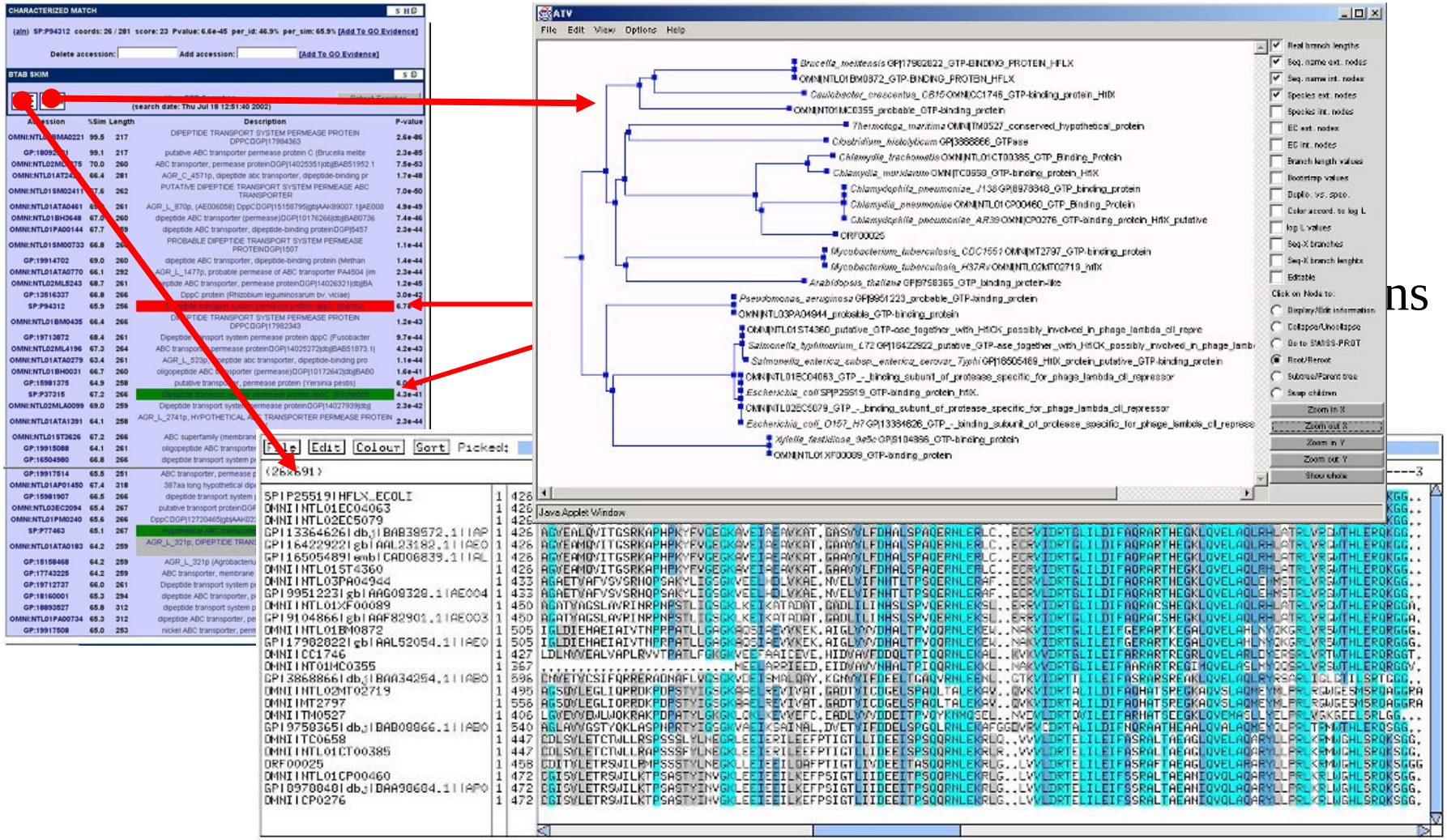
	M. j.	Chl tr	Strep	Chl tr	Therm	A. f.	B. b.	Caul	D. r.	H. i.	H.p.	M. g.	TB	Neis.	Trep	Vib.
TIGRFAMS	174	190	247	0	256	216	189	332	302	386	298	148	338	370	186	450
Identified genes	522	393	1894	381	768	899	559	1338	947	1024	889	263	1210	961	504	1676

Gene Information Page

Gene Identification
Gene Ontology
Evidence Picture
Submit Data
Gene Curation Information
Gene Curation Page
Interpro Data
Attributes
Signal_P
Lipoprotein Information
Tigr Roles
Hmm
Prosite
Role



Pair-wise Alignment Summary



ns

Workflow System

Viewing systems exist (e.g., DAS, gbrowse, Apollo, Manatee, Artemis), but how to *create* data?

- Implemented in Java, ~5000 lines
- Describes workflow as a directed acyclic graph
- Supports serial and parallel processes
- Executes onto Condor/LSF
- Two main files:
  - Human-readable config files
  - XML templates
- config file + XML template → XML “instance”

Workflow System

- XML instance, contents:
  - Complete description of pipeline
  - Contains status of pipeline
  - Allows monitoring:
    - resumption of failed instances
    - straightforward tracking of multiple instances

Workflow System

- My scripts work fine, why use your workflow system?
  - Allows monitoring.
  - Supports sharing between centers.
  - Uncouples: DB ←→ Computing
    - Remote computing
    - Substitution of new systems.

Annotation Engine

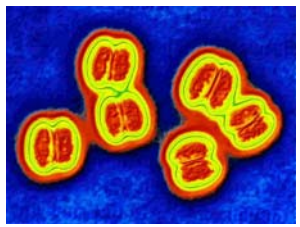
- Send bacterial sequence to el tigre
- Run through our pipeline
- Return search results, etc
- View in Manatee
- NOT an automated annotation system

bing badda boom.

Bacterial Projects

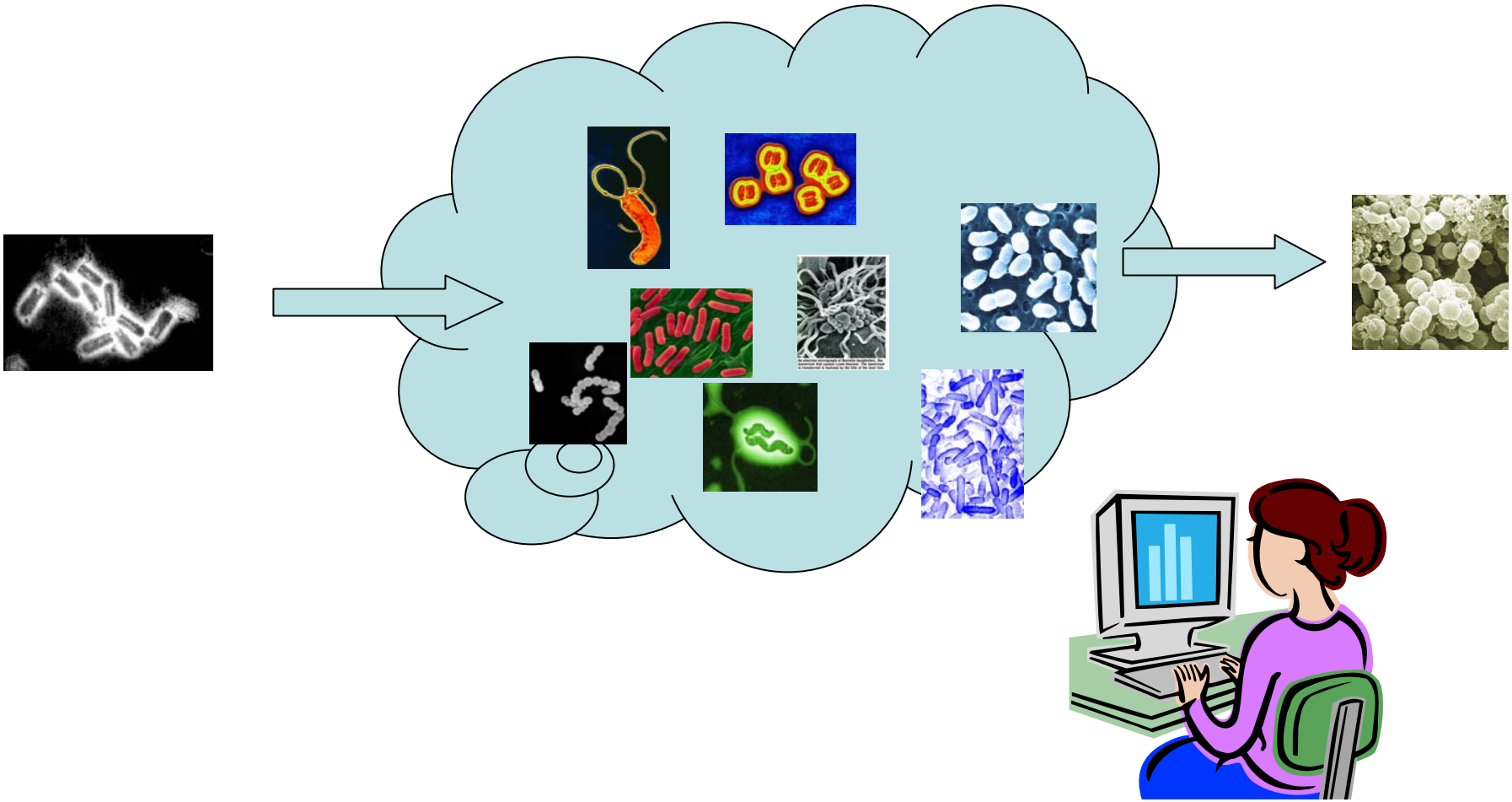
Exp Date	Status	Organism	PI
5/31/2003	annotation	Brucella ovis ATCC25840	Ian Paulsen
6/30/2004	annotation	Clostridium perfringens ATCC13124	Ian Paulsen
6/30/2004	annotation	Clostridium perfringens SM101	Ian Paulsen
7/31/2004	annotation	Ehrlichia chaffeensis Arkansas	Herve Tettelin
7/31/2004	annotation	Neorickettsia sennetsu Miyayama	Herve Tettelin
7/31/2004	annotation	Anaplasma phagocytophilum HZ	Herve Tettelin
8/31/2004	annotation	Myxococcus xanthus DK 1622	William Nierman
9/30/2004	annotation	Fibrobacter succinogenes S85	Karen Nelson
3/31/2006	annotation	Silicibacter pomeroyi DSS-3	Naomi Ward
9/30/2004	pre-annotation	Dichelobacter nodosus VCS1703A	Ian Paulsen
9/30/2004	pre-annotation	Pseudomonas fluorescens Pf-5	Ian Paulsen
3/14/2005	pre-annotation	Prevotella ruminicola 23	Karen Nelson
9/29/2004	pre-annotation	Streptococcus agalactiae A909	Herve Tettelin
8/31/2004	pre-annotation	Hyphomonas neptunium ATCC 15444	Naomi Ward
8/31/2004	pre-annotation	Sulfurihydrogenibium azorense	John Heidelberg
9/30/2004	pre-annotation	Acidobacterium capsulatum ATCC 51196	Naomi Ward
8/31/2004			John Heidelberg
unknown	pre-annotation	Mycoplasma capricolum ATCC27343	Bill Nierman
9/30/2004	closure	Ruminococcus albus 8	Karen E. Nelson
9/30/2004	closure	Verrucomicrobium spinosum DSM 4136	Naomi Ward
8/31/2004	closure	Burkholderia thailandensis E264	William Nierman
8/31/2004	closure	Pseudomonas syringae pv phaseolicola	Robin Buell
8/31/2004	closure	Arthrobacter aurescens	Karen Nelson
9/30/2005	closure	Tree of Life: Coprothermobacter proteolyticus	Jonathan Eisen
11/30/2004	closure	Streptococcus pneumoniae 670	Herve Tettelin
8/31/2005	closure	Salinibacter ruber	Karen Nelson
8/31/2005	closure	Yellowstone Cyanobacteria: A-Prime	John Heidelberg
8/31/2005	closure	Yellowstone Cyanobacteria: B-Prime	John Heidelberg
9/30/2005	closure	Aeromonas hydrophila	John Heidelberg
9/30/2004	random	Epulopiscium spp.	Karen E. Nelson
9/30/2005	random	Tree of Life: Chrysiogenes arsenatis	Jonathan Eisen
9/30/2005	random	Tree of LIfe: Dictyoglomus thermophilum	Jonathan Eisen
9/30/2005	random	Tree of LIfe: Thermodesulfobacterium commune	Jonathan Eisen
9/30/2005	random	Tree of Life: Thermomicrobium roseum	Jonathan Eisen
9/30/2005	random	Tree of Life: Thermodesulfovibrio yellowstonii	Jonathan Eisen
11/9/2005	random	Haemophilus influenzae type b (HK707)	Herve Tettelin
12/31/2007	random	Chlamydophila pecorum	Steven Gill
12/31/2007	random	Chlamydia pneumoniae-Koala	Steven Gill
12/31/2007	random	Chlamydophila psittaci	Steven Gill
12/31/2007	random	Chlamydia suis-S45	Steven Gill
12/31/2007	random	Simkania negevensis-Z	Steven Gill
12/31/2007	random	Waddlia chondrophila	Steven Gill
11/30/2004	library	Mycoplasma bovis Donetta PG45	Barbara Methe

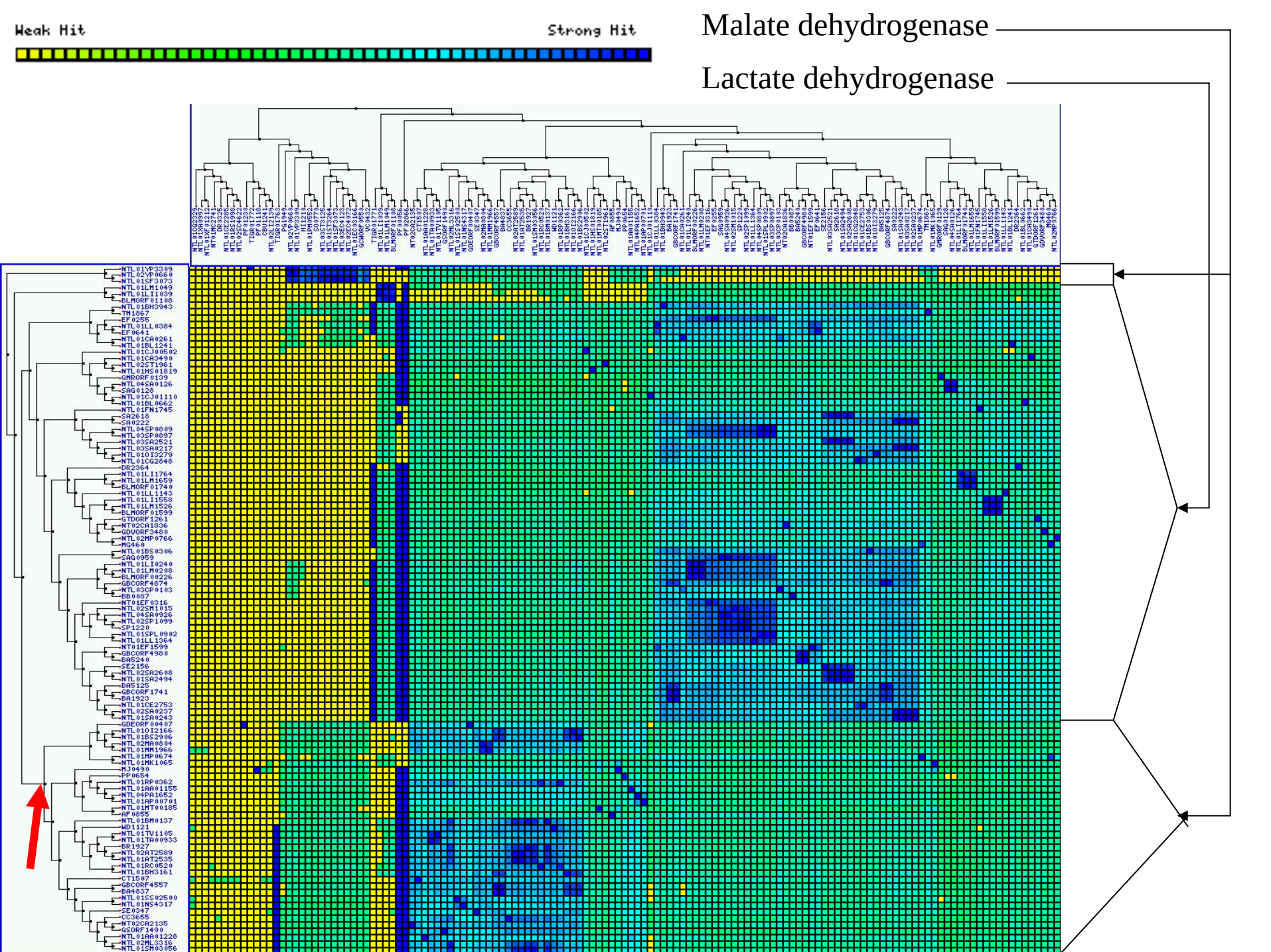
Serial Annotation



1, 2, 3, 4, 5, 6….

Streaming Annotation




Weak Hit
Strong Hit
Malate dehydrogenase
Lactate dehydrogenase

Ident Information					
Gene	com_name	ec_num	gene_sym	role_id	mainrole
GDEORF00407	malate dehydrogenase	1.1.1.37	mdh	147	Transport and binding proteins
NTL01OI2166	malate dehydrogenase	1.1.1.37	citH		
NTL01BS2906	malate dehydrogenase	1.1.1.37	citH		
NTL02MA0804	malate dehydrogenase				
NTL01MM1966	Malate dehydrogenase	1.1.1.37			
NTL01MP0674	L-lactate dehydrogenase		ldh		
NTL01MK1065	Lactate dehydrogenase		Ldh		
MJ0490	malate dehydrogenase	1.1.1.27		114	Energy metabolism
PP0654	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NTL01RP0362	MALATE DEHYDROGENASE (mdh)				
NTL01AA01155	malate dehydrogenase		mdh2		
NTL04PA1652	malate dehydrogenase				
NTL01AP00701	313aa long hypothetical malate dehydrogenase	NULL			
NTL01MT00185	lactate dehydrogenase				
AF0855	L-malate dehydrogenase, NAD+-dependent	1.1.1.27	mdhA	114	Energy metabolism
NTL01BM0137	MALATE DEHYDROGENASE	1.1.1.37			
WD1121	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NTL01TV1105	Malate dehydrogenase				
NTL01TA00933	probable malate dehydrogenase				
BR1927	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NTL02AT2589	malate dehydrogenase		mdh		
NTL01AT2535	AGR_C_4782p, MALATE DEHYDROGENASE				
NTL01RC0520	malate dehydrogenase [EC:1.1.1.37]	1.1.1.37	mdh		
NTL01BH3161	malate dehydrogenase	1.1.1.37	citH		
CT1507	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
GBCORF4557	malate dehydrogenase, NAD-dependent	1.1.1.37	mdh	120	Energy metabolism
BA4837	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NTL01SS02500	2-ketoacid dehydrogenase (malate dehydrogenase, lactate dehydrogenase)		citH		
NTL01NS4317	malate dehydrogenase				
SE0347	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
CC3655	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NT02CA2135	malate dehydrogenase, NAD-dependent	1.1.1.37	mdh	120	Energy metabolism
GSORF1490	malate dehydrogenase	1.1.1.37	mdh	120	Energy metabolism
NTL01AA01228	malate dehydrogenase	NULL	mdh1		
NTL02ML3316	malate dehydrogenase				
NTL01SM03056	PROBABLE MALATE DEHYDROGENASE PROTEIN	1.1.1.37	mdh		

gene models

feat_name	database	locus/pub_locus	gene product_name	pub_comment	comment	gene structure curated	gene annotation curated
ORF00536	gde	/				☑	☑
MJ0490	arg	MJ0490/MJ0490	malate dehydrogenase		name change due to experimental evidence by Madern, D and Zaccai, G. Info received via email to website. See Michelle.	☑	☑
ORF07491	gpp	PP0654/PP0654	malate dehydrogenase		GBF she, vds - protein starts probably with MDVQGELAQGKA (own experimental data) Start confidence Low	☑	☑
ORF01378	gaf	AF0855/AF0855	L-malate dehydrogenase, NAD+-dependent		novel Archaeoglobus gene not in our database! Got OLAP from BLAST, GP:2065007	☑	☑
ORF00221	dmg	WD1121/WD1121	malate dehydrogenase			☑	☑
ORF01389	gbr	BR1927/BR1927	malate dehydrogenase			☑	☑
ORF01856	gct	CT1507/CT1507	malate dehydrogenase		Involved in reverse TCA cycle? -JFK	☑	☑
ORFB05120	gbc	GBCORF4557/GBCORF4557	malate dehydrogenase, NAD-dependent			☑	☑
ORF02792	gba	BA4837/BA4837	malate dehydrogenase			☑	☑
ORF01085	gse	SE0347/SE0347	malate dehydrogenase		Start confidence Med; HMM and BTAB evidnece. mjb	☑	☑
ORF02087	gcc	CC3655/CC3655	malate dehydrogenase		BER2multi groups this ORF with unpublished proteins and genomes. RTD Best candidate for malate DHase in the genome	☑	☑
ORF02498	ggs	GSORF1490/GSORF1490	malate dehydrogenase		Best candidate for malate dehydrogenase in the genome. mjb	☑	☑

common annotation

gene product names:
malate dehydrogenase
other:

[Multiple Assignments]

pub_comment:
☐ overwrite pub_comments

comment:
☐ overwrite comments
Annotated in batch should still get a quick review

go_ids *(Additional evidence can be added after submission)*

- ☑ GO:0004470 (F) malic enzyme activity
- ☐ GO:0006099 (P) tricarboxylic acid cycle
- ☑ GO:0016615 (F) malate dehydrogenase activity

best GO guesses shared *(Additional evidence can be added after submission)*
☐ **delete selected GO ids** *(Check here, then click Submit)*

other go ids: *(Seperated by spaces)*
go_id

EC#s 1.1.1.37
other:

gene_sym mdh

bing badda boom.