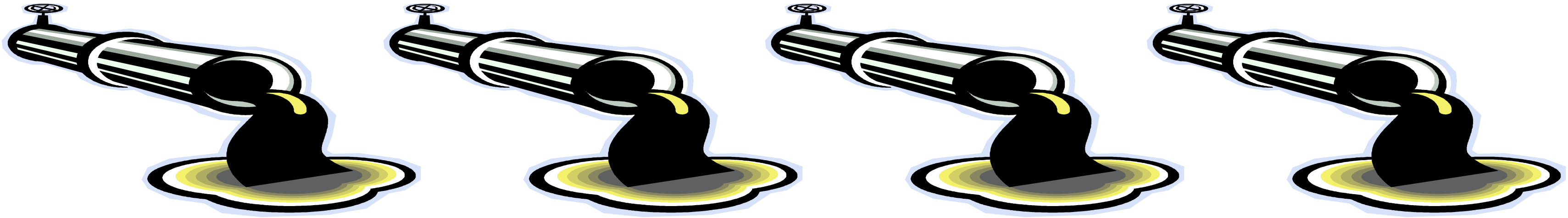
BrainGrab:

Capturing Curator Expertise into Reusable Annotation Rules

Daniel Haft
2009

J. Craig Venter
INSTITUTE

There are many well-developed annotation pipelines ….

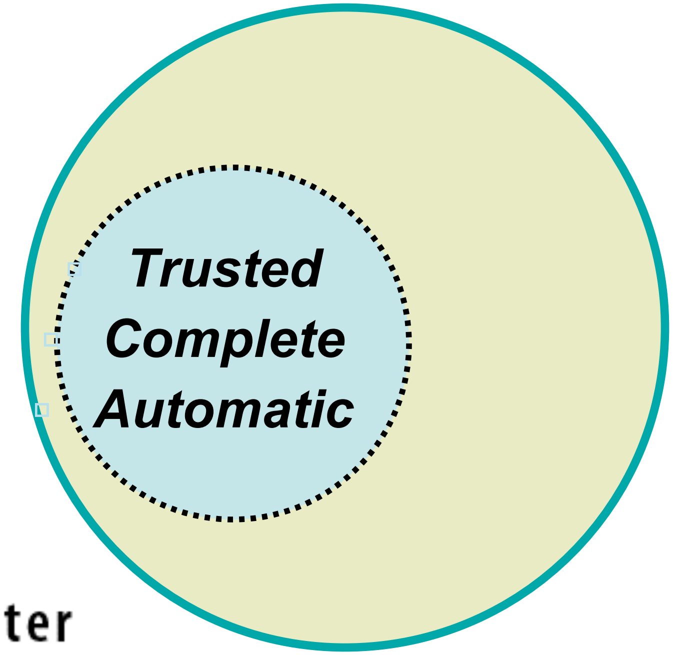


…. that may be hard to unify into a shared community resource,

but **FACTS** and **RULES** can become common currency.

Annotation Proceeds from …

- **Outside looking in**
  - Search tool + cutoff + implications = **annotation rule**
  - Achieves partial coverage



Trusted
Complete
Automatic

e.g. TIGRFAMs

TIGRFAMs as annotation rules

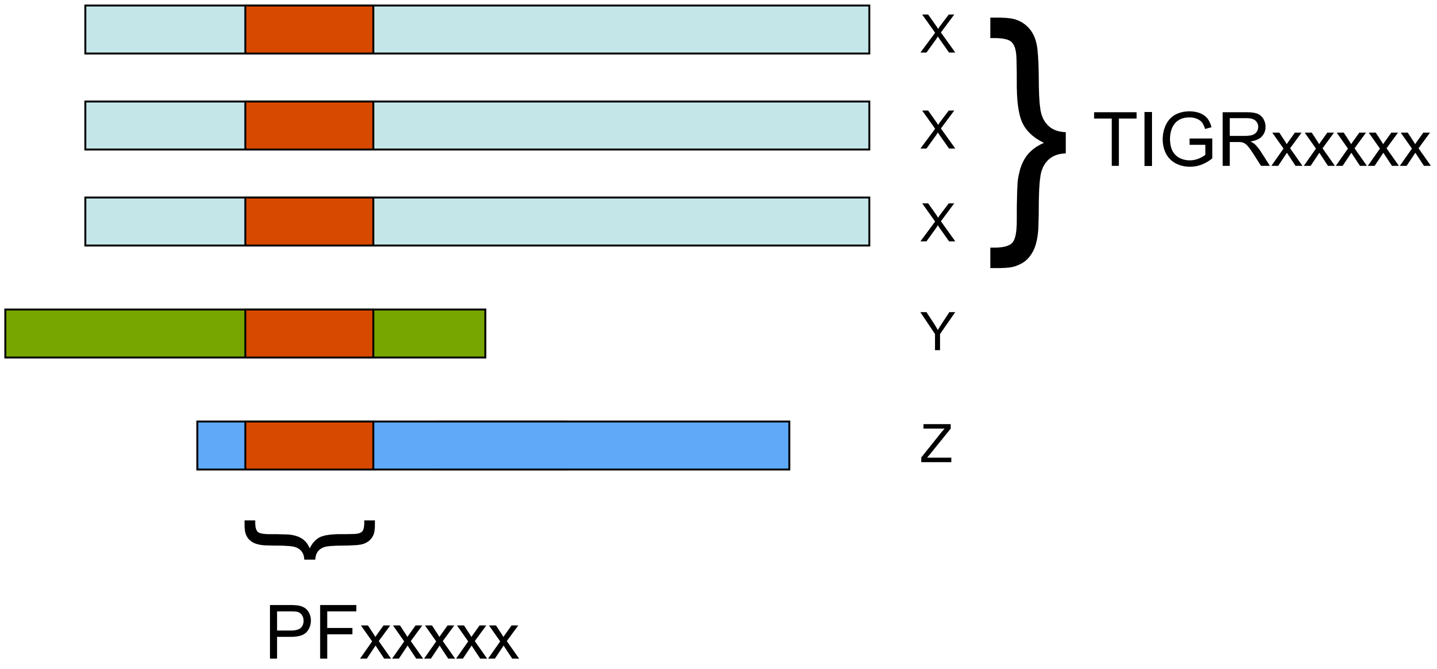
EC number	computable !
GO term	computable !
HMM hit	computable !!
protein name	computable ?

Gene Finder → HMM hits → Functions → Pathways → Phylogenetic Profile → ?

Specificity rankings for annotation rules

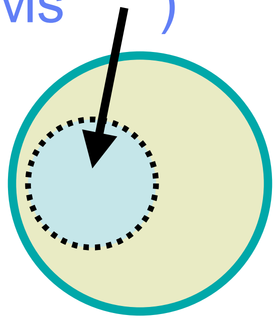
- **EXCEPTION** additional info, *e.g.* “vegetative”
- **EQUIVALOG** A protein that shares its main function with another by means of conservation from a common ancestral protein.
- **SUBFAMILY** can name a whole class
- **DOMAIN** class name for a protein region

TIGRFAMs equivalogs *vs.* Pfam domains

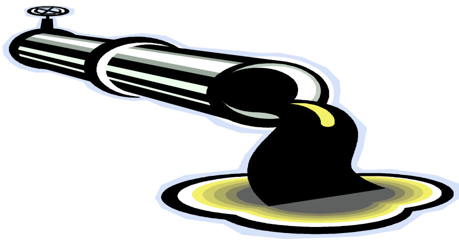


Annotation Proceeds from …

- outside looking in (e.g. TIGRFAMs)

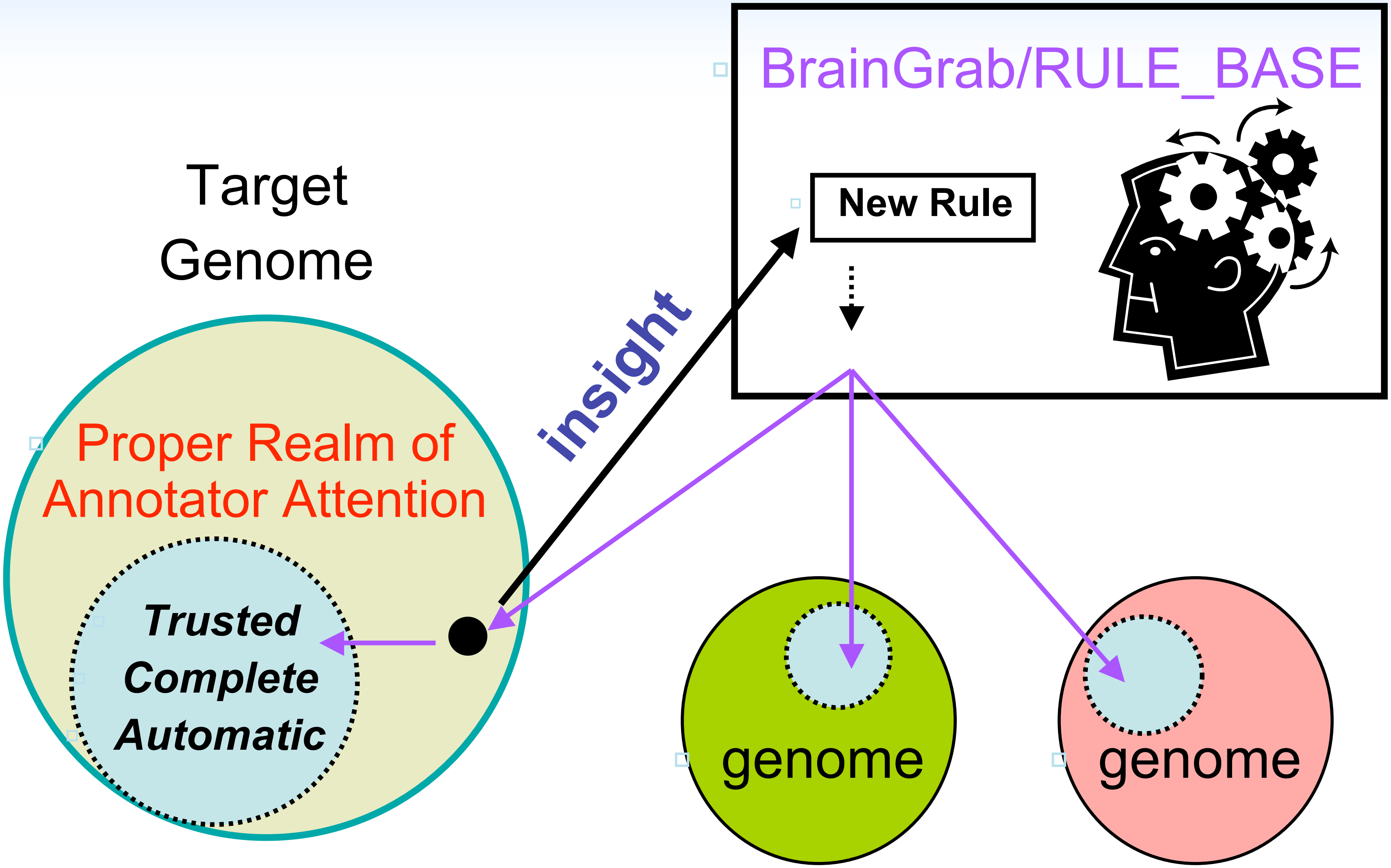


- inside looking out
  - Every gene gets annotated
  - Mixed evidence types
  - Heuristic best-guess annotation



Annotation Proceeds from …

- **Outside --> in** (e.g. TIGRFAMs): for every model
  - Search tool + cutoff + standards = **annotation rule**
  - Achieves partial coverage
- **Inside --> out**
  - Mixed evidence types
  - Every gene gets annotated
- **Hybrid** (BrainGrab) for every **unfinished** protein
  - Look for means to annotate: blastp, synteny, hole-filling, etc.
  - Capture annotator logic as a new rule
  - Add to library of rules/models for all future genomes



BrainGrab/RULE_BASE

New Rule

Target Genome

insight

Proper Realm of Annotator Attention

Trusted
Complete
Automatic

genome

genome

MANATEE Biocuration in a Dual-Use Pipeline

- Multiple types of **stored evidence**
  - Persistent & Flexibly Interleaved
  - Supports selective re-annotation
  - Features **annotation-driving** databases
    - CHAR
    - TIGRFAMs
    - Genome Properties
    - BrainGrab Rules

- Evidence used by **Machine** and by **Experts**
  - **MANATEE** interface for annotators
  - Capture new rules with **BrainGrab**

A Wealth of Pre-computes

- **BLAST** searches
- **HMM** searches: TIGRFAMs & Pfam
- Other InterPro classifiers
- SignalP, TmHMM, OMP, misc. motifs
- Boutique databases (*e.g.* TransportDB)
- Taxonomy, Phenotypic, Genome Properties

if protein = “SelD” and “genome contains 2-selenouridine synthase”

THIS_HMM_HIT [TIGR00476] && **GENOME_HMM_HIT** [TIGR03167]

then protein gets GO process term “tRNA seleno-modification” (but don’t remove other GO terms)

Field	GO ids
Contents	GO:0070329
Mode	append

if protein belongs to PF00281 and is found in a bacterial genome

THIS_HMM_HIT [PF00281] && **GEN_STATE** [GenProp0006, ”Bacteria”]

then apply proper terms for protein name, gene symbol, common name, etc.

Field	GO ids
Contents	GO:0003735 GO:0006412 GO:0022625
Mode	replace

Field	com_name
Contents	**50S ribosomal protein L5**
Mode	replace

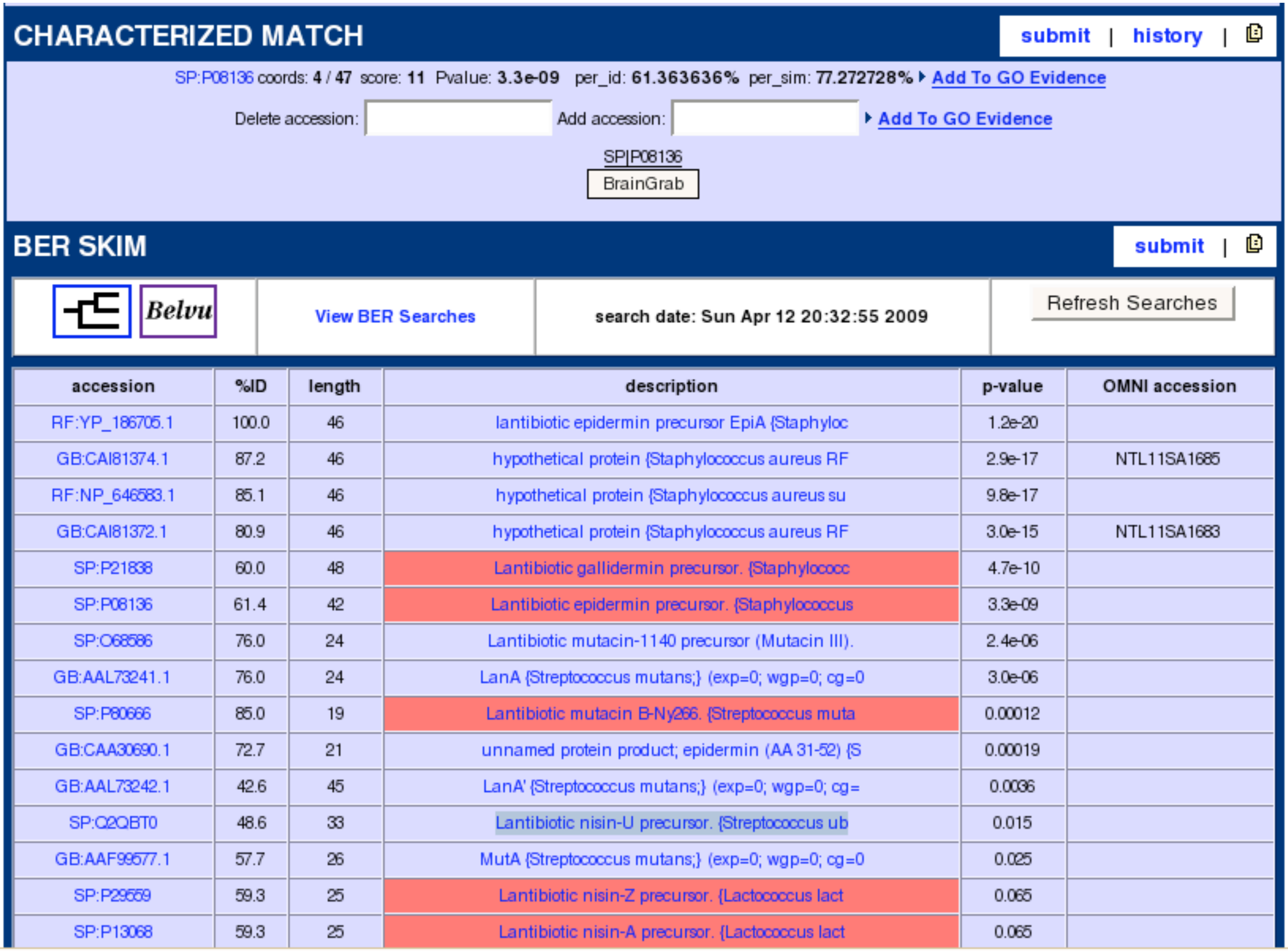
Would be wrong in the Archaea !!

etc.

Some Predicate Types:

```
THIS_HMM_HIT [accession]
NEAR_HMM_HIT [distance, accession]
GENOME_HMM_HIT [accession]
GEN_PROP [property,value]
DEFAULT_METHOD [accession]
THIS_BLAST_HIT [seven parameters]
```

And we will happily add YOUR evidence type …
haft@jcvi.org



CHARACTERIZED MATCH

submit | history |

SP:P08136 coords: **4 / 47** score: **11** Pvalue: **3.3e-09** per_id: **61.363636%** per_sim: **77.272728%** ▸ Add To GO Evidence

Delete accession: ______ Add accession: ______ ▸ Add To GO Evidence

SP|P08136

BrainGrab

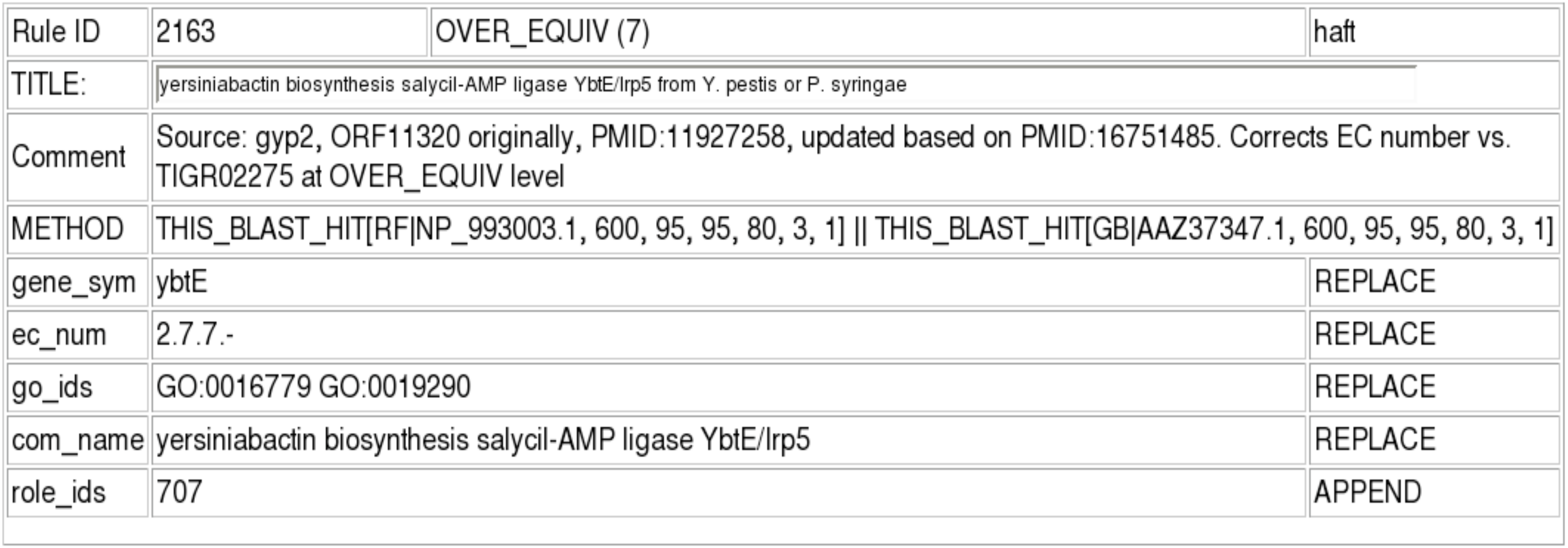
BER SKIM

submit |

Belvu | View BER Searches | **search date: Sun Apr 12 20:32:55 2009** | Refresh Searches

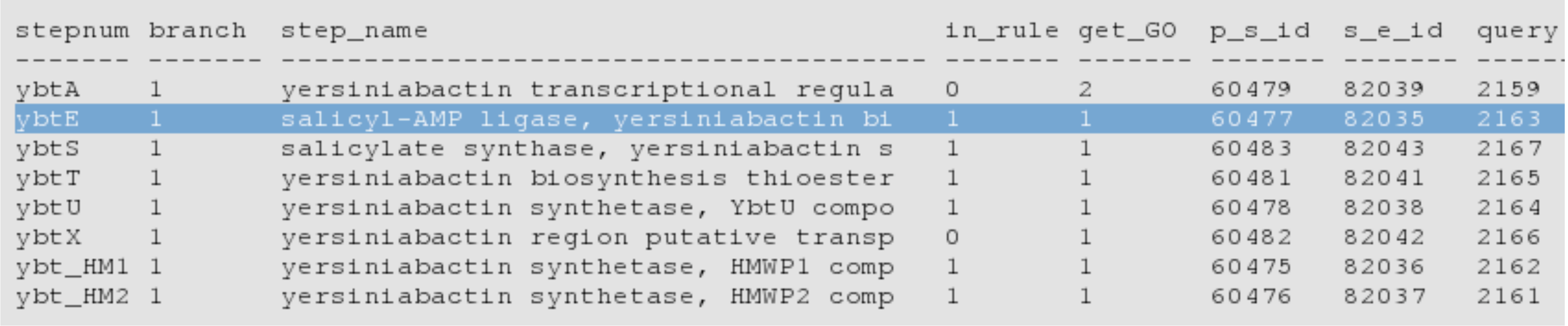
accession	%ID	length	description	p-value	OMNI accession
RF:YP_186705.1	100.0	46	lantibiotic epidermin precursor EpiA {Staphyloc	1.2e-20	
GB:CAI81374.1	87.2	46	hypothetical protein {Staphylococcus aureus RF	2.9e-17	NTL11SA1685
RF:NP_646583.1	85.1	46	hypothetical protein {Staphylococcus aureus su	9.8e-17	
GB:CAI81372.1	80.9	46	hypothetical protein {Staphylococcus aureus RF	3.0e-15	NTL11SA1683
SP:P21838	60.0	48	Lantibiotic gallidermin precursor. {Staphylococc	4.7e-10	
SP:P08136	61.4	42	Lantibiotic epidermin precursor. {Staphylococcus	3.3e-09	
SP:O68586	76.0	24	Lantibiotic mutacin-1140 precursor (Mutacin III).	2.4e-06	
GB:AAL73241.1	76.0	24	LanA {Streptococcus mutans;} (exp=0; wgp=0; cg=0	3.0e-06	
SP:P80666	85.0	19	Lantibiotic mutacin B-Ny266. {Streptococcus muta	0.00012	
GB:CAA30690.1	72.7	21	unnamed protein product; epidermin (AA 31-52) {S	0.00019	
GB:AAL73242.1	42.6	45	LanA' {Streptococcus mutans;} (exp=0; wgp=0; cg=	0.0036	
SP:Q2QBT0	48.6	33	Lantibiotic nisin-U precursor. {Streptococcus ub	0.015	
GB:AAF99577.1	57.7	26	MutA {Streptococcus mutans;} (exp=0; wgp=0; cg=0	0.025	
SP:P29559	59.3	25	Lantibiotic nisin-Z precursor. {Lactococcus lact	0.065	
SP:P13068	59.3	25	Lantibiotic nisin-A precursor. {Lactococcus lact	0.065	

Sample BrainGrab rule (acting like a TIGRFAMs HMM)



Rule ID	2163	OVER_EQUIV (7)	haft
TITLE:	yersiniabactin biosynthesis salycil-AMP ligase YbtE/Irp5 from Y. pestis or P. syringae		
Comment	Source: gyp2, ORF11320 originally, PMID:11927258, updated based on PMID:16751485. Corrects EC number vs. TIGR02275 at OVER_EQUIV level		
METHOD	THIS_BLAST_HIT[RF|NP_993003.1, 600, 95, 95, 80, 3, 1] || THIS_BLAST_HIT[GB|AAZ37347.1, 600, 95, 95, 80, 3, 1]		
gene_sym	ybtE		REPLACE
ec_num	2.7.7.-		REPLACE
go_ids	GO:0016779 GO:0019290		REPLACE
com_name	yersiniabactin biosynthesis salycil-AMP ligase YbtE/Irp5		REPLACE
role_ids	707		APPEND

BrainGrab/RULE_BASE evidence as computable objects (for Genome Properties)



```
stepnum branch   step_name                                 in_rule get_GO   p_s_id   s_e_id   query
------- -------  ----------------------------------------  ------- -------  -------  -------  ------
ybtA     1       yersiniabactin transcriptional regula     0       2        60479    82039    2159
ybtE     1       salicyl-AMP ligase, yersiniabactin bi     1       1        60477    82035    2163
ybtS     1       salicylate synthase, yersiniabactin s     1       1        60483    82043    2167
ybtT     1       yersiniabactin biosynthesis thioester     1       1        60481    82041    2165
ybtU     1       yersiniabactin synthetase, YbtU compo     1       1        60478    82038    2164
ybtX     1       yersiniabactin region putative transp     0       1        60482    82042    2166
ybt_HM1  1       yersiniabactin synthetase, HMWP1 comp     1       1        60475    82036    2162
ybt_HM2  1       yersiniabactin synthetase, HMWP2 comp     1       1        60476    82037    2161
```

J. Craig Venter
INSTITUTE

Acknowlegements

Malay Basu

Alex Richter

Ramana Madupu

Kevin Galens

Jeremy Selengut

JCVI microbial annotation