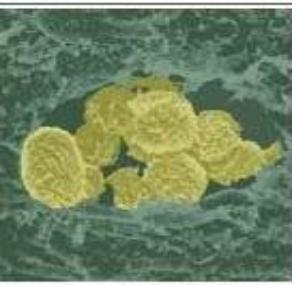
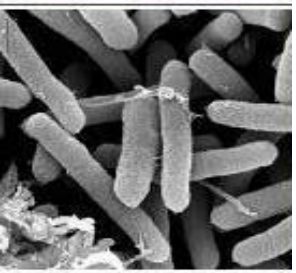
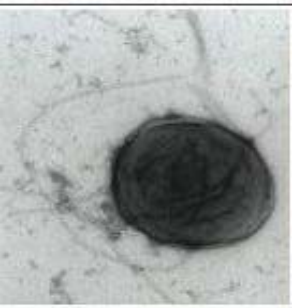
JGI DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

img.jgi.doe.gov

img INTEGRATED MICROBIAL GENOMES

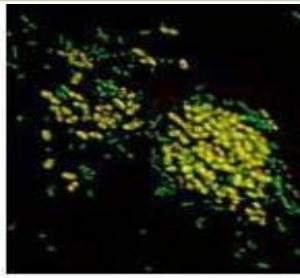
The (IMG) Systems for Comparative Analysis of Microbial Genomes & Metagenomes

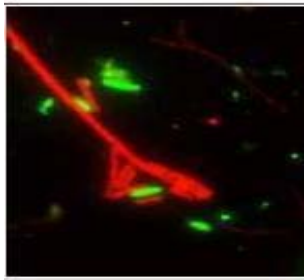
Victor M. Markowitz *

Nikos C. Kyrpides **

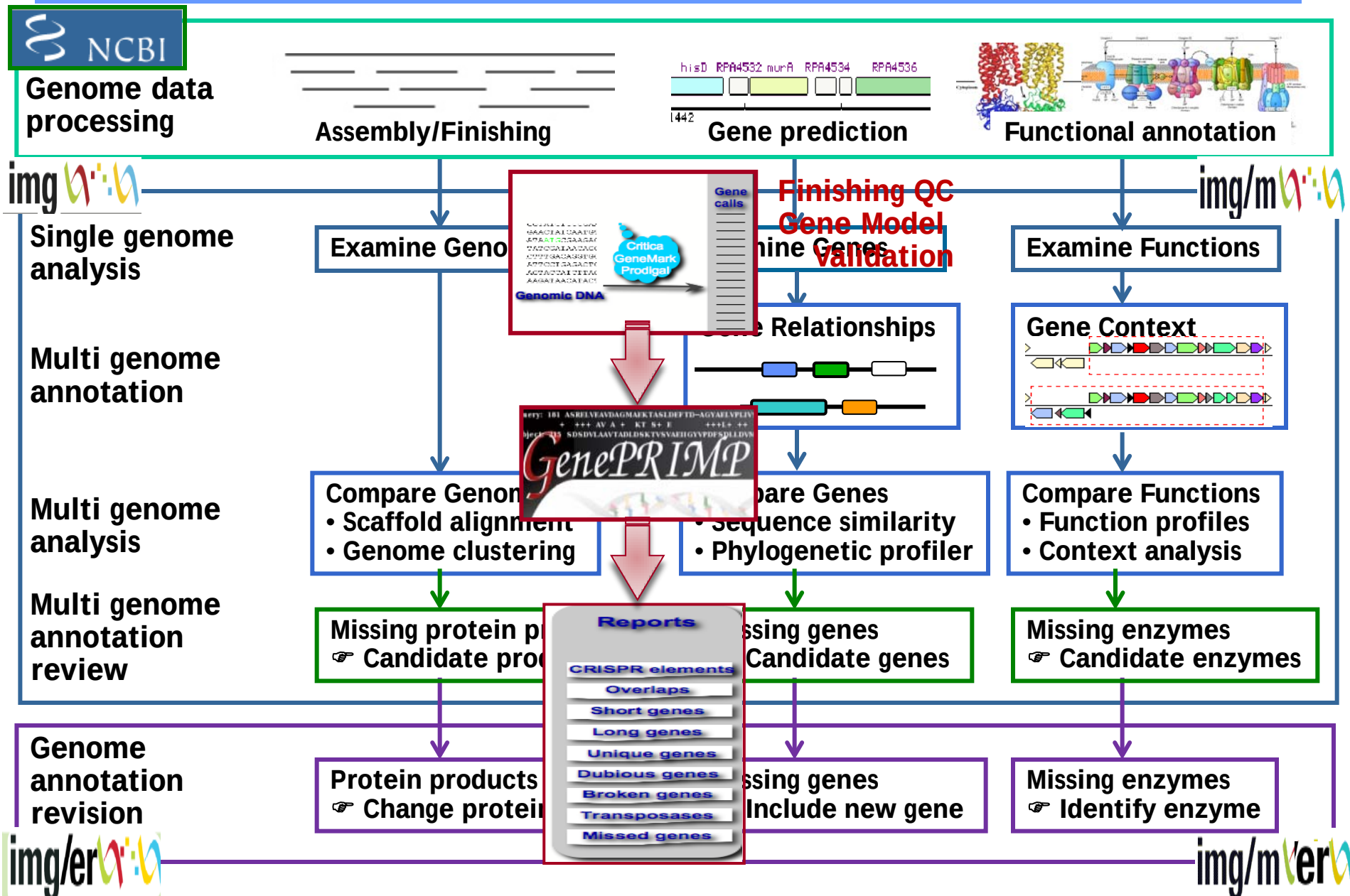
*** Lawrence Berkeley National Lab**

**** Genome Biology Program, Joint Genome Institute**

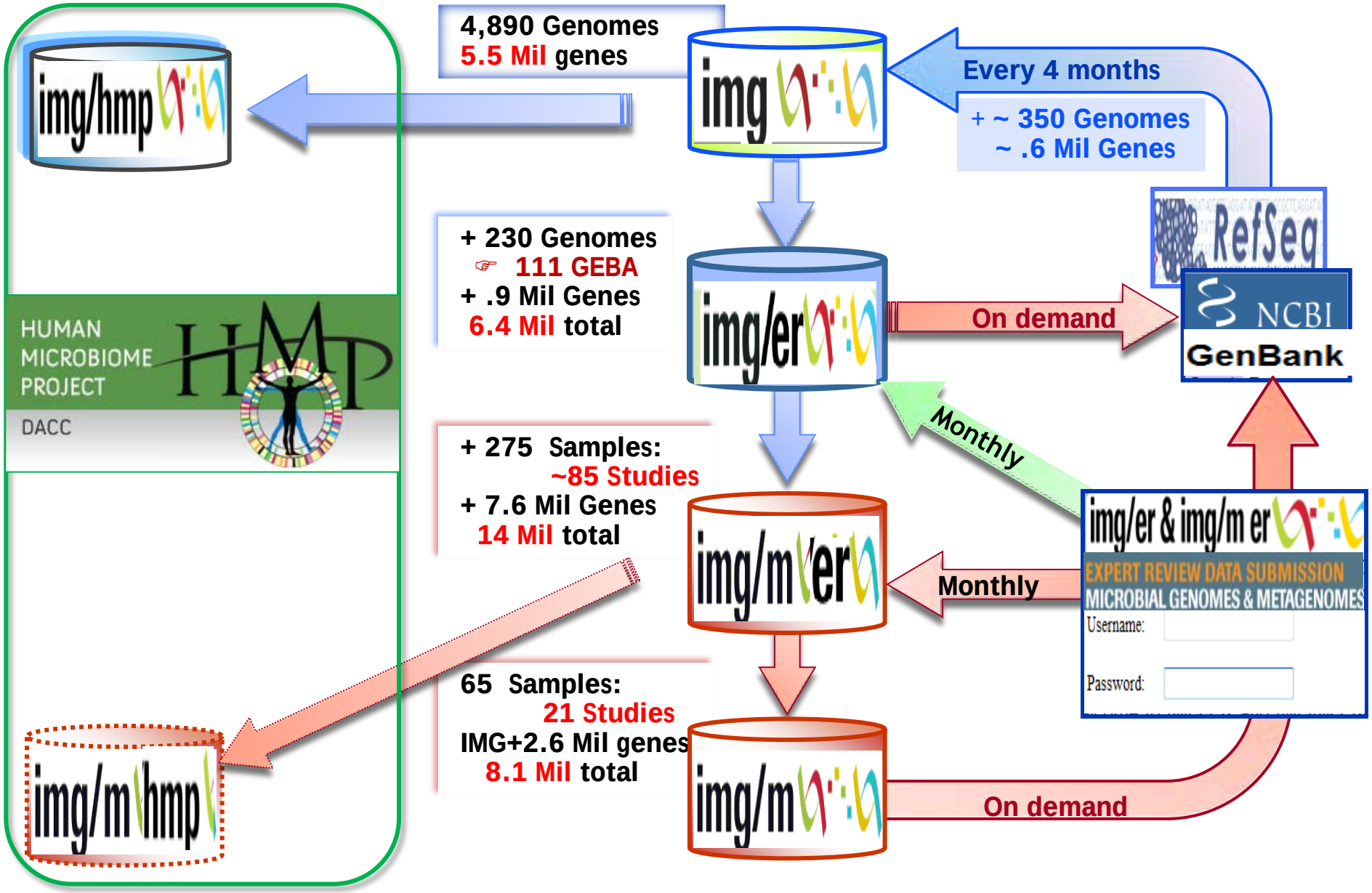




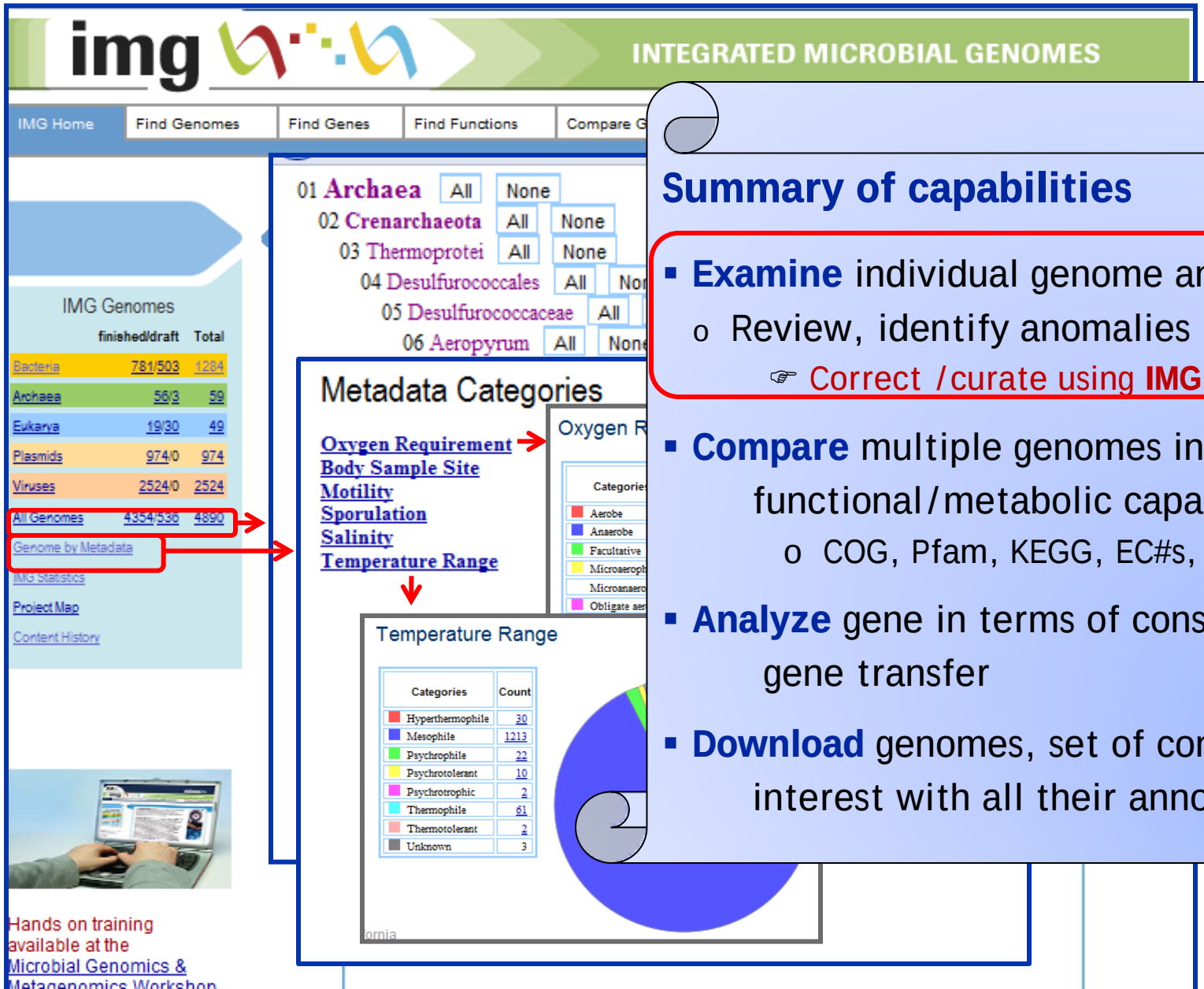

Genome sequence data processing & analysis



IMG systems: genome/metagenome data flow



IMG Synopsys



Summary of capabilities

- **Examine** individual genome annotations
  - o Review, identify anomalies
    - ☞ Correct /curate using **IMG ER**
- **Compare** multiple genomes in terms of functional/metabolic capabilities
  - o COG, Pfam, KEGG, EC#s, …
- **Analyze** gene in terms of conservation, gene transfer
- **Download** genomes, set of contigs of interest with all their annotations

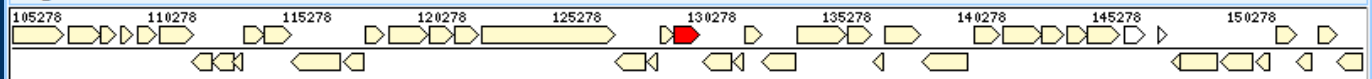
IMG: reasoning about annotations

Gene Information

Gene Information	
Gene Object ID	638180295
Gene Symbol	Ta0133
Locus Tag	Ta0133
Product Name	pirin related protein
SwissProt Protein Product	Pirin-like protein Ta0133
SEED	
IMG Term	
Description	similarity to known prote
Genome	Thermoplasma acidophilu
DNA Coordinates	129819..130733 (+)(915b
Scaffold Source	Thermoplasma acidophilu
GC Content	0.47
Accession	CAC11280
External Links	GI:10639278; UniProt/Sw
Fused Gene	No
Fusion Component	No
Protein Information	
Amino Acid Sequence Length	304aa
COG	COG1741 - Pirin-related
Families	- IPR003829 Pirin, N-te - IPR008778 Pirin, C-te - IPR011051 Cupin, Rn - IPR012093 Pirin
Transmembrane Helices	No
Signal Peptide	No
Statistics	peptide
Pfam	pfam02678 - Pirin pfam05726 - Pirin_C
Pathway Information	
Enzymes	
KEGG Orthology (KO) Term	KO:K06911 K06911
KEGG Pathway	
KEGG Orthology (KO) Modules	
MetaCyc Pathway	

Evidence For Function Prediction

Neighborhood



red = Current Gene
white = Pseudo Gene
green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene
||||||| CRISPR array
Sequence Viewer For Alternate ORF Search
Chromosome Viewer colored by COG
Chromosome Viewer colored by GC percentage

Conserved Neighborhood

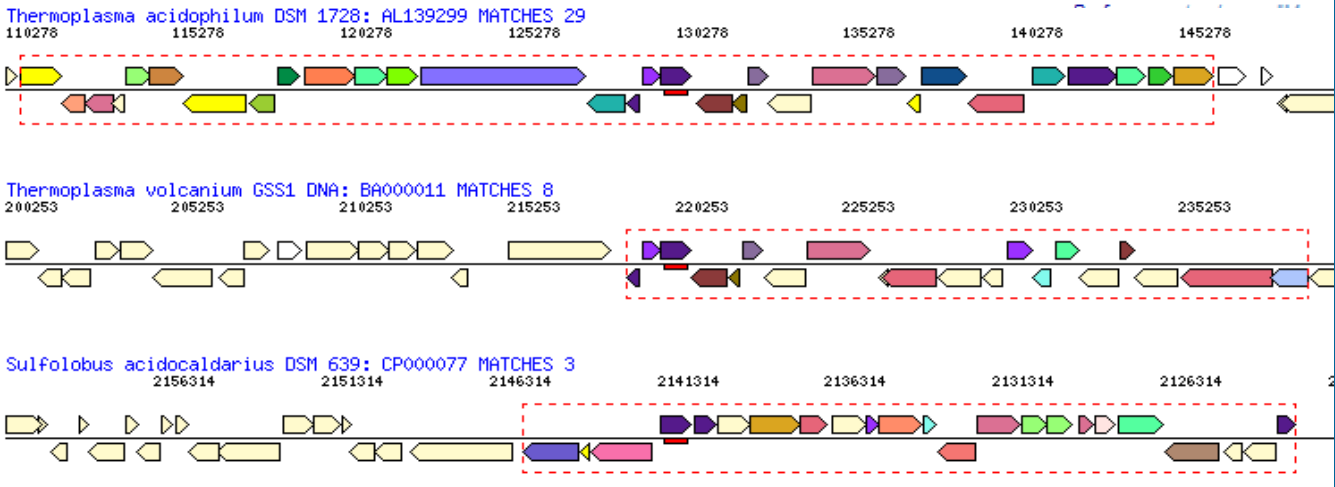
Ortholog Neighborhood Viewer
Chromosomal Cassette Viewer By -- Select Protein Cluster --

COG

COG ID	Consensus Sequence Length	Description	Percent Identity	Alignment On Query	E-value	Bit Score

Chromosomal Cassette By COG

Results limited to 15 neighborhoods. (Go to



IMG: Protein product review

Organism Information

Organism Name	Thermoplasma acidophil
Taxon Object ID	638154521
NCBI Taxon ID	273075
NCBI Project ID	110
GOLD ID in IMG Database	Gc0003

Genome Statistics

Compare Gene Annotation

Compare Gene Annotations

Compare

View annotati

Select filter *

Gene Object ID
638180166

Gene Detail

Gene Information
Gene Object ID
Gene Symbol
Locus Tag
Product Name
SwissProt Protein Product
IMG Term
Description

Chromosomal Cassette By COG

Thermoplasma acidophilum DSM 1728: AL139299 MATCHES 24

-10736 -5736 -736 4264 9264 14264 19264

Ferroplasma acidarmanus Fer1, unfinished sequence: NZ_AABC04000005 MATCHES 2

31765 36765 41765 46765 51765 56765 61765

FaciDRAFT_1021 : SNO glutamine amidotransferase(EC:2.6.-) 51477 .. 52049 (190 aa) (COG0311)

Evidence For Function Prediction

Conserved Neighborhood	
Ortholog Neighborhood Viewer	
Chromosomal Cassette Viewer By	COG

pyridoxine biosynthesis

20 top hi

al protein

Candidate Product Names for Query Gene (OID: 638180166)

Genome Name	Thermoplasma acidophilum DSM 1728
Gene Product Name	hypothetical protein
IMG Term(s)	pyridoxal phosphate synthase yaaE subunit
COG	Predicted glutamine amidotransferase involved in pyridoxine biosynthesis
Pfam	SNO glutamine amidotransferase family
KEGG Orthology (KO)	glutamine amidotransferase [EC:2.6.-.-]

Select	Homolog Gene	Original Product Name	IMG Term OID	IMG Term	D	C	Genome	Percent Identity	Alignment On Candidate	Alignment On Homolog	E-valu
◉	638394069	SNO glutamine amidotransferase	05446	pyridoxal phosphate synthase yaaE subunit	A	D	Ferroplasma acidarmanus Fer1	54.45			6.00e-

Add to MyIMG Annotation

MyIMG Annotation for Selected Genes

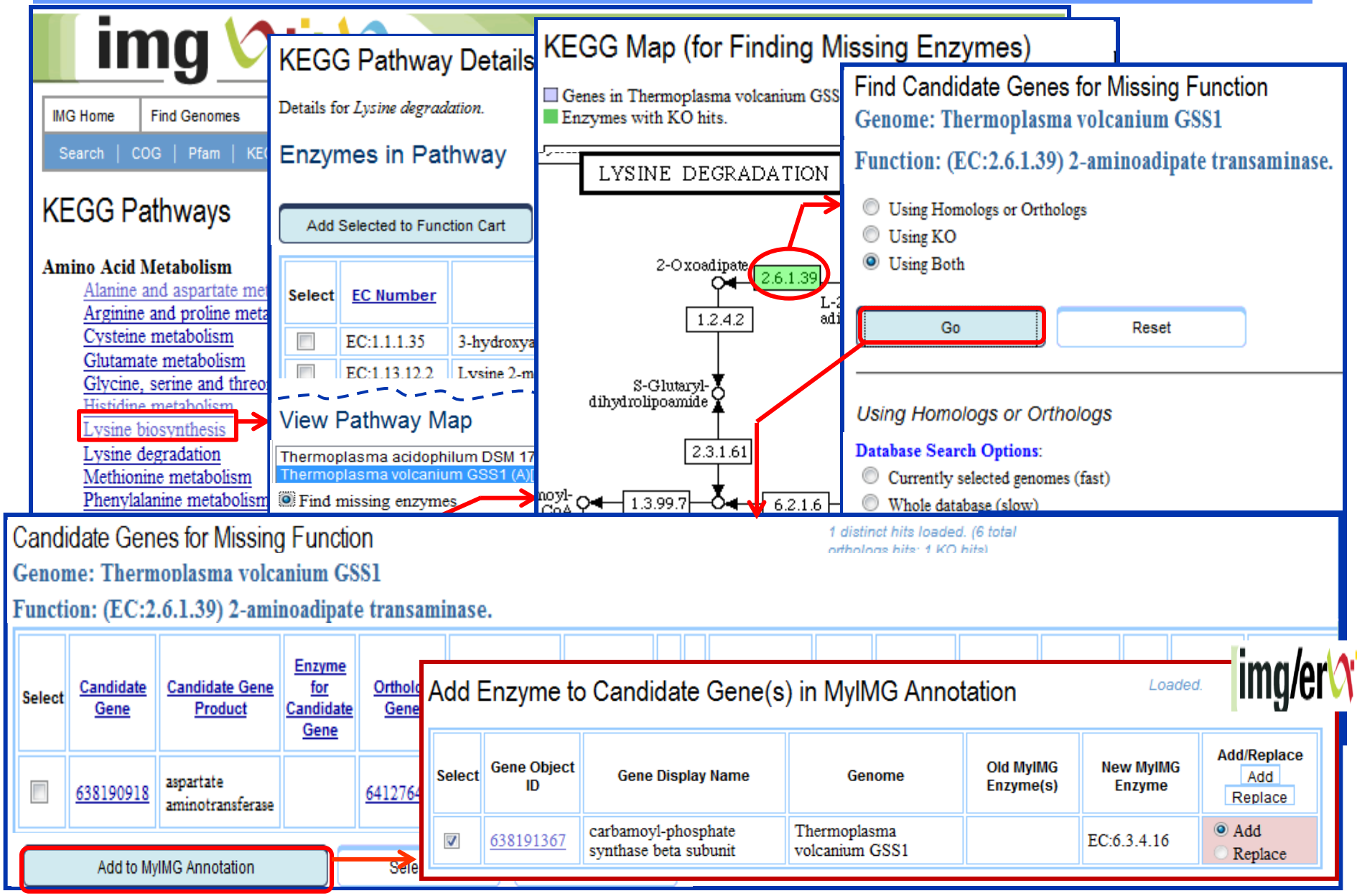
Select	Gene Object ID	Original Product Name	Annotated Product Name	Prot Description	EC Number
☑	638180166	hypothetical protein	SNO glutamine amidotransferase		

MyIMG Annotation

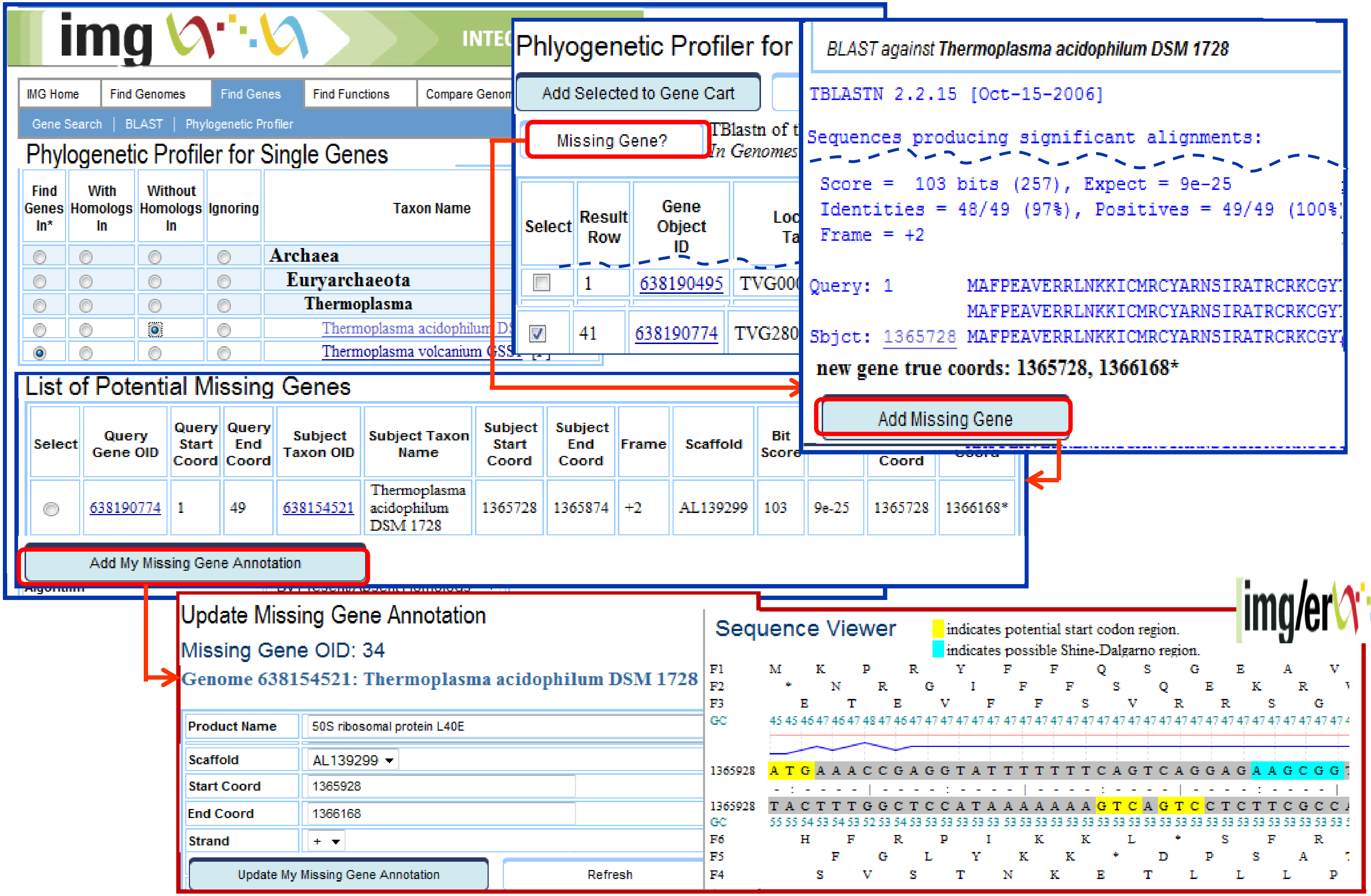
Product Name	SNO glutamine amidotransferase
Prot Description	
EC Number	

Update Annotation | Delete Annotation

IMG: Metabolic capability review- missing enzymes



IMG: Gene model review- missing genes



Phylogenetic Profiler for Single Genes

Find Genes In*	With Homologs In	Without Homologs In	Ignoring	Taxon Name
○	○	○	○	**Archaea**
○	○	○	○	**Euryarchaeota**
○	○	○	○	**Thermoplasma**
○	○	●	○	Thermoplasma acidophilum DS
●	○	○	○	Thermoplasma volcanium GSS1 [+]

Phlyogenetic Profiler for

TBlastn of t
In Genomes

Select	Result Row	Gene Object ID	Loc Ta
☐	1	638190495	TVG00
☑	41	638190774	TVG280

BLAST against ***Thermoplasma acidophilum DSM 1728***

```
TBLASTN 2.2.15 [Oct-15-2006]

Sequences producing significant alignments:

 Score =  103 bits (257), Expect = 9e-25
 Identities = 48/49 (97%), Positives = 49/49 (100%)
 Frame = +2

Query: 1       MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
                MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
Sbjct: 1365728 MAFPEAVERRLNKKICMRCYARNSIRATRCRKCGY
```

new gene true coords: 1365728, 1366168*

List of Potential Missing Genes

Select	Query Gene OID	Query Start Coord	Query End Coord	Subject Taxon OID	Subject Taxon Name	Subject Start Coord	Subject End Coord	Frame	Scaffold	Bit Score		Coord	Coord
○	638190774	1	49	638154521	Thermoplasma acidophilum DSM 1728	1365728	1365874	+2	AL139299	103	9e-25	1365728	1366168*

Update Missing Gene Annotation

Missing Gene OID: 34

Genome 638154521: Thermoplasma acidophilum DSM 1728

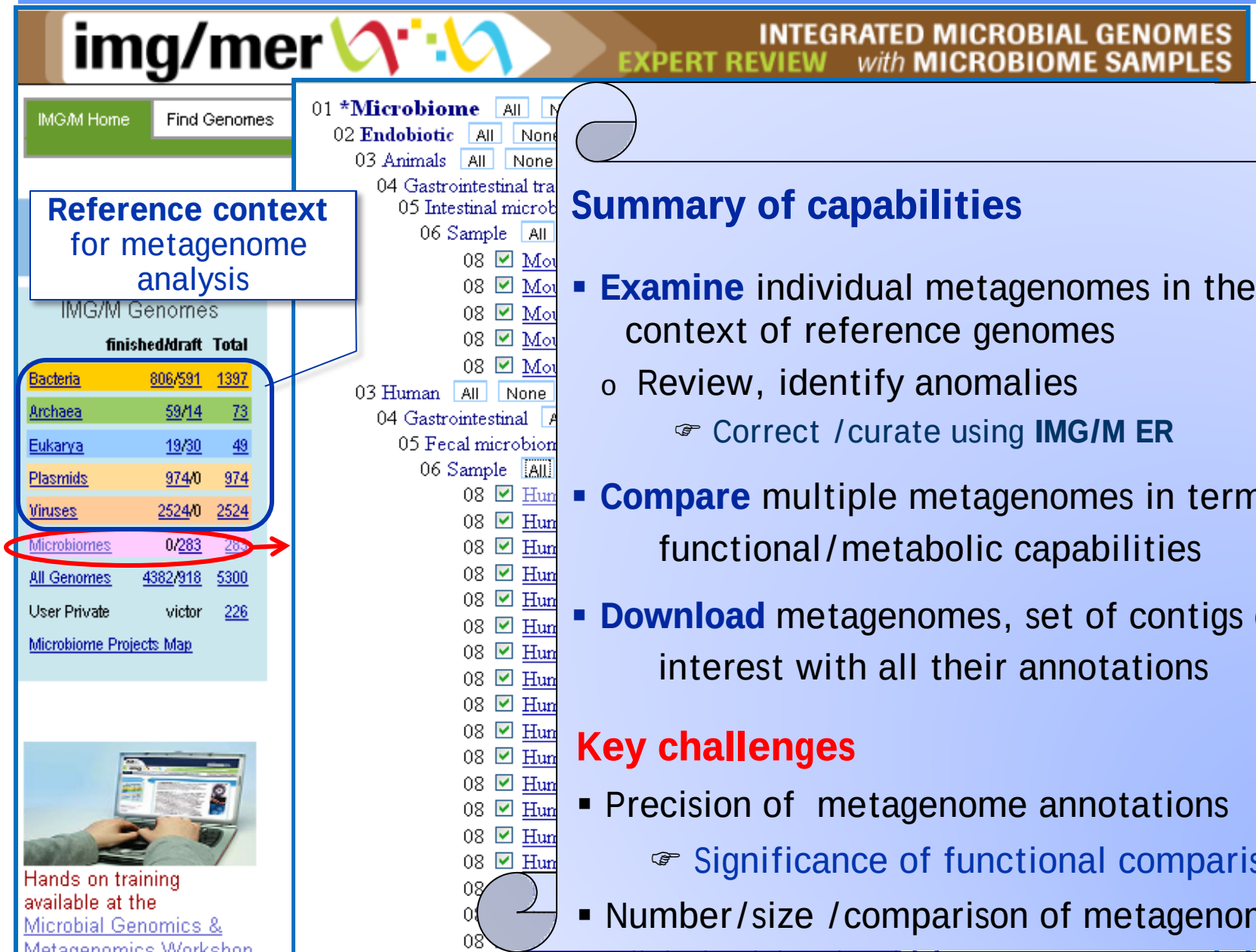
Product Name	50S ribosomal protein L40E
Scaffold	AL139299 ▾
Start Coord	1365928
End Coord	1366168
Strand	+ ▾

Sequence Viewer

🟨 indicates potential start codon region.
🟦 indicates possible Shine-Dalgarno region.

```
F1       M     K     P     R     Y     F     F     Q     S     G     E     A     V
F2          *     N     R     G     I     F     F     S     Q     E     K     R
F3          E     T     E     V     F     F     S     V     R     R     S     G
GC       45 45 46 47 46 47 48 47 46 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47 47
1365928  A T G A A A C C G A G G T A T T T T T T T C A G T C A G G A G A A G C G G
         - : - - - - - | - - - - - : - - - - - | - - - - - : - - - - - | - - - - - : - - - - - |
1365928  T A C T T T G G C T C C A T A A A A A A A G T C A G T C C T C T T C G C C
GC       53 53 54 53 54 53 52 53 54 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53
F6          H     F     R     P     I     K     K     L     *     S     F     R
F5             F     G     L     Y     K     K     *     D     P     S     A
F4          S     V     S     T     N     K     E     T     L     L     L     P
```

IMG/M Synopsis



Summary of capabilities

- **Examine** individual metagenomes in the context of reference genomes
  - o Review, identify anomalies
    - ☞ Correct /curate using **IMG/M ER**
- **Compare** multiple metagenomes in terms functional/metabolic capabilities
- **Download** metagenomes, set of contigs of interest with all their annotations

Key challenges

- Precision of metagenome annotations
  - ☞ Significance of functional comparisons
- Number/size /comparison of metagenomes

http://www.jgi.doe.gov/meetings/mgm/index.html

DOE Joint Genome Institute

Microbial Genomics & Metagenomics (MGM) Workshop

IMG Tutorial (Genome annotation and analysis)

IMG - Genes and Genomes — Athanasios Lykidis

Microbial genome data analysis in IMG is set in the comparative context of multiple microbial genomes. IMG allows navigating the microbial genome data space along three key dimensions: genomes (organisms), functions (terms and pathways) and genes. In this section, IMG-based comparative analysis of gene families and genomes will be

Hands-on IMG (exercises)

Exercise solutions

IMG - Functions and Pathways — Iain Anderson

IMG has several ways for users to interact with protein functions and pathways, including Clusters of Orthologous Groups (COGs) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. In addition, JGI is developing a controlled vocabulary for the representation of functions

MyIMG — Iain Anderson

The functional annotation for individual genes can be modified using the MyIMG Annotations features of MyIMG. In addition to curation of functional annotations, MyIMG provides support for uploading user genome selections that have been saved earlier from the Genome Browser or Genome Statistics and for setting systemwide user

Gene Context Analysis in IMG — Kostas Mavrommatis

Hands-on IMG (exercises) — Users

Exercise solutions — Iain Anderson

IMG/M Tutorial (metagenome analysis)

IMG/M — Natalia Ivanova

Introduction to Metagenome analysis

A snapshot of microbial community structure can be derived from analysis of metagenomic data. IMG/M methods and tools for establishing the taxonomic identity of community members will be presented along with tools for determining the fine population structure, genetic variation and genome dynamics of the dominant populations. Methods for assessing the diversity and abundance of microbial communities will be discussed.

Statistical analysis of metagenomic datasets — Amrita Pati

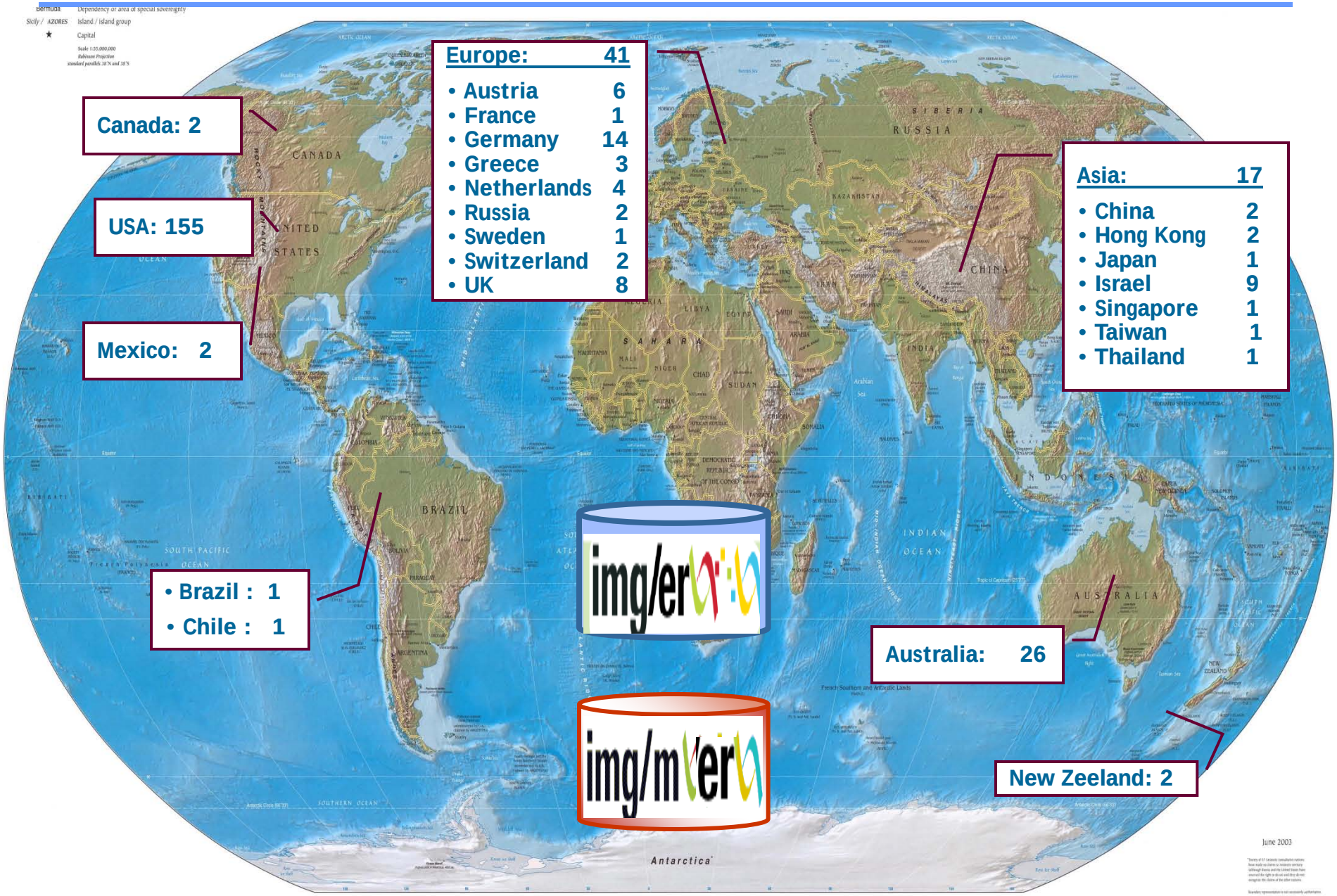
The systematic evaluation of the relative abundances of individual as well as sets of protein functions across various metagenomic datasets, can yield statistically significant deductions about over- and under-representation of protein function(s) and biological pathways in these communities. We can derive statistical methods for comparing the relative abundances of both individual as well as sets of protein families in 2 given metagenomic datasets. Statistical models for modeling individual abundances and methods for identifying protein families whose difference in abundances are statistically significant, will be presented.

Hands-on IMG (exercises) — Users

A MetaGenome Analysis test case — Athanasios Lykidis

The methodology and steps to analyze a metagenome in IMG/M will be presented with a user case

IMG ER systems: community users



Acknowledgements