Metabolic Reconstructions from Global Ocean Sampling (GOS) Marine Metagenome

Mathangi Thiagarajan

J. Craig Venter Institute

Pathways Tools Workshop 2010

- Metagenomics
- The Global Ocean Sampling (GOS) Project
- GOS - Community Makeup
- High Throughput Data Processing
- Metabolic Reconstruction – Mapping to MetaCyc and KEGG
- Metarep (Visualization) – Integrating with MetaCyc and KEGG
- Pathways Tools for GOS & metagenomic projects
- Conclusion
- Acknowledgements

Metagenomics

- Examining genomic content of organisms in community/environment to better understand
  - Diversity of organisms
  - Their roles and interactions in the ecosystem

- Cultivation independent approach to study microbial communities
  - DNA directly isolated from environmental sample and sequenced

Global Ocean Sampling Expedition

Investigate the fundamental microbial contributions from the Ocean waters to energy and nutrient cycling by analyzing its

a) biogeochemical cycling

b) community structure and function

c) microbial diversity

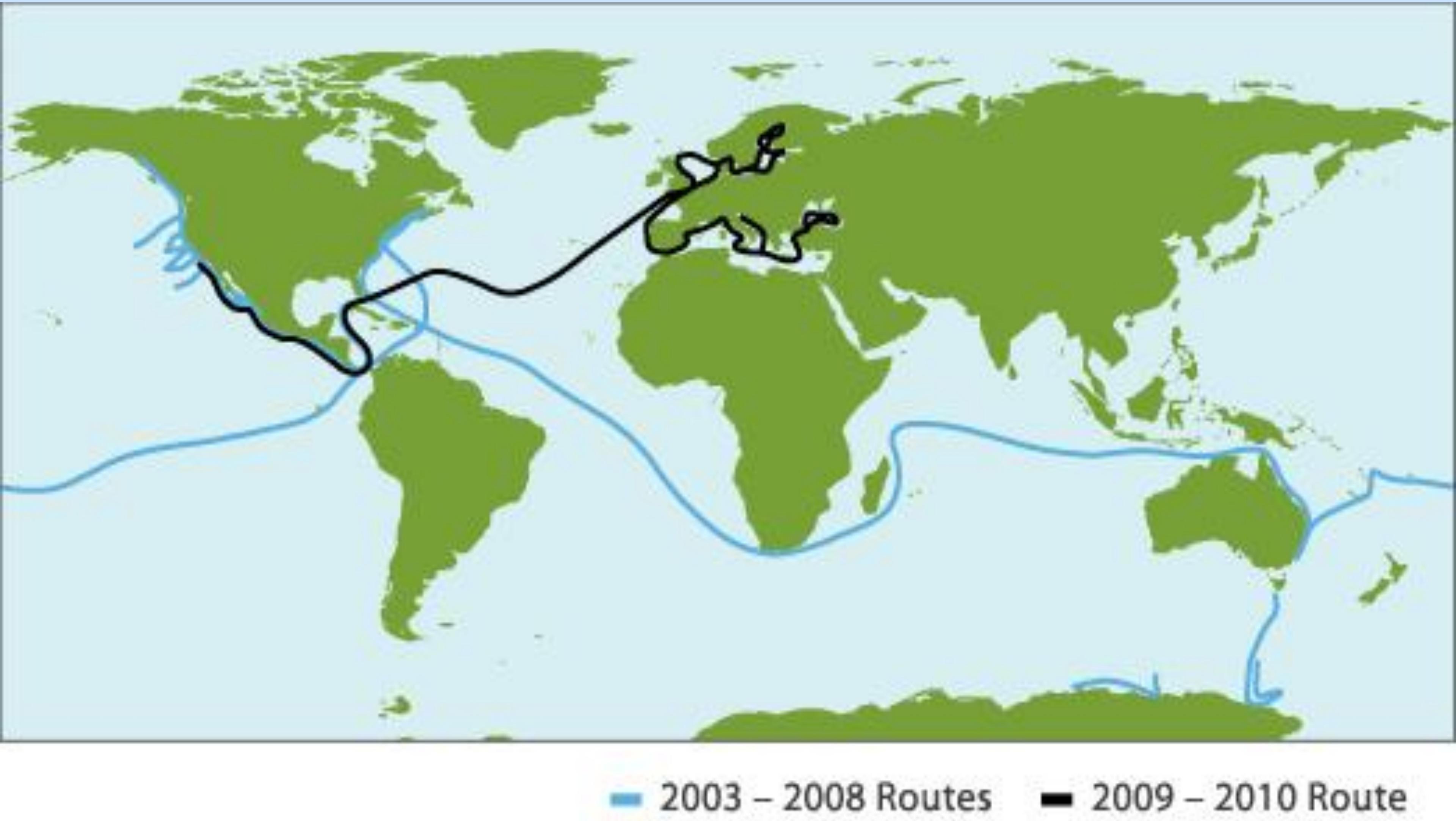
d) adaptation and evolution

GOS Phase I - Published in PLOS Biology 2007

GOS Circumnavigation - Analysis Phase

J. Craig Venter
INSTITUTE

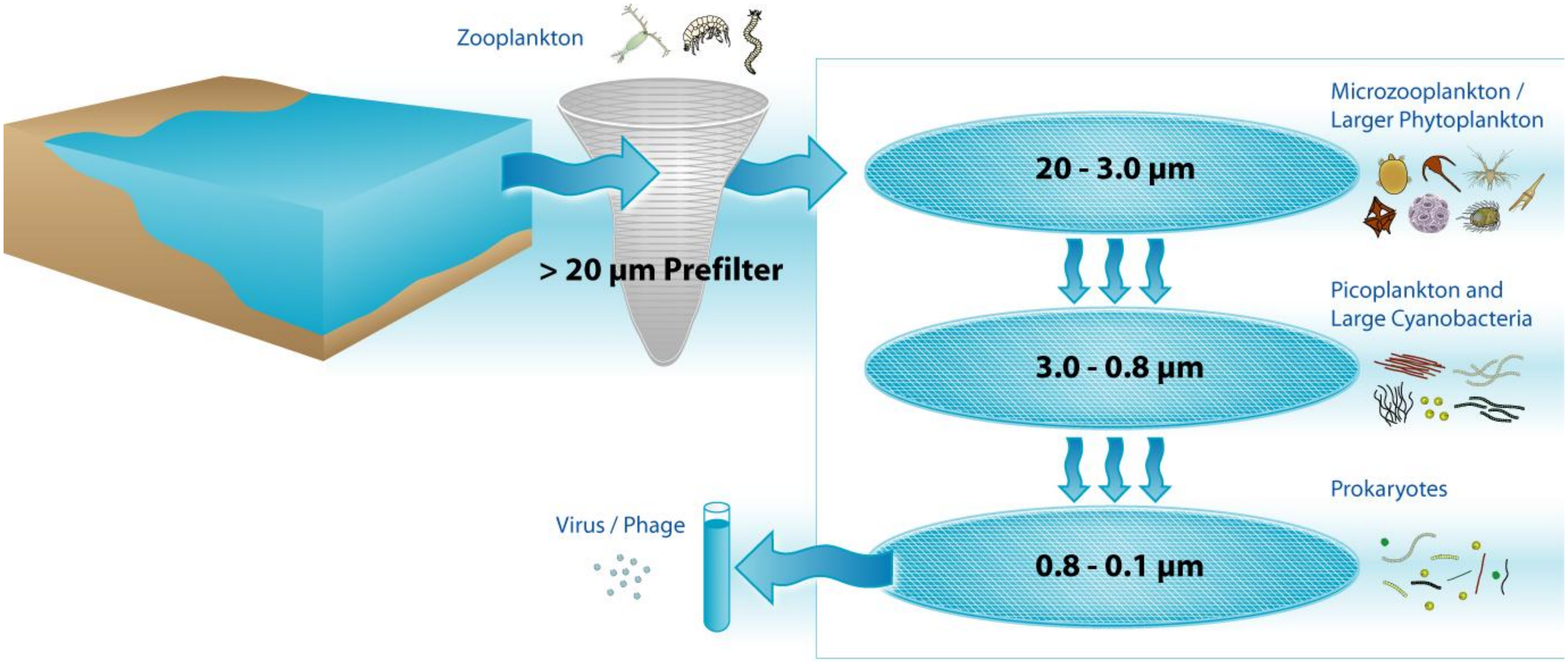
Global Ocean Sampling Expedition Route



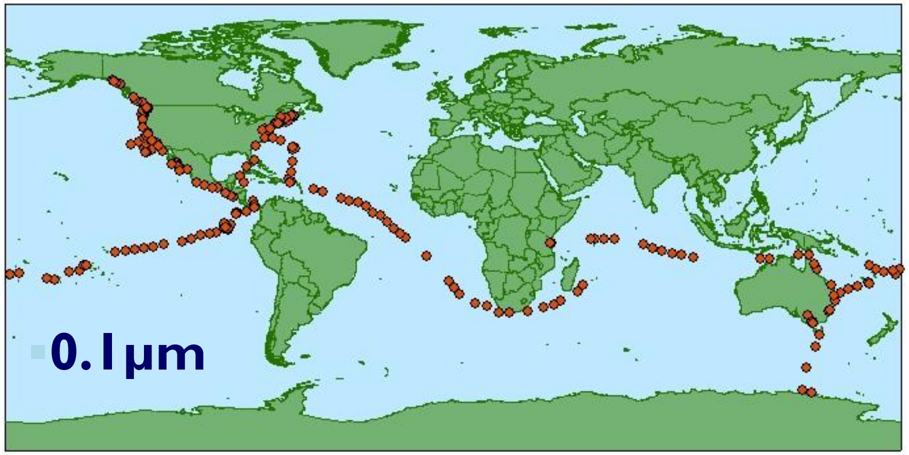
J. Craig Venter
INSTITUTE

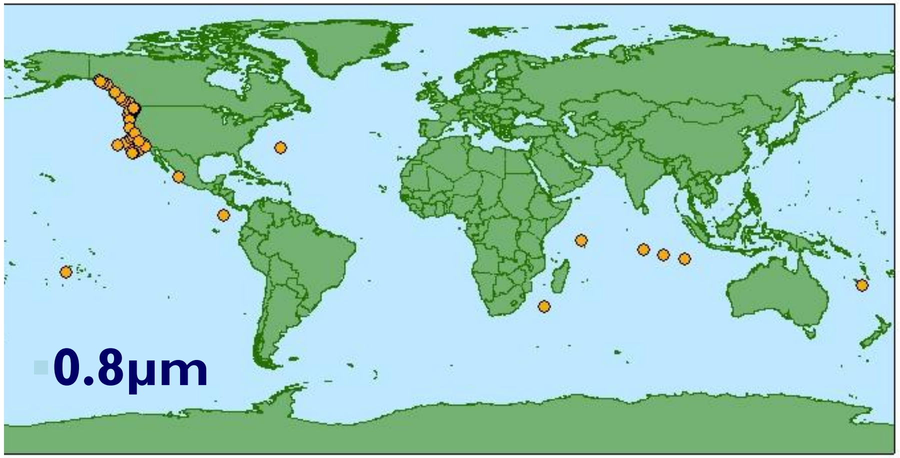
SA-80

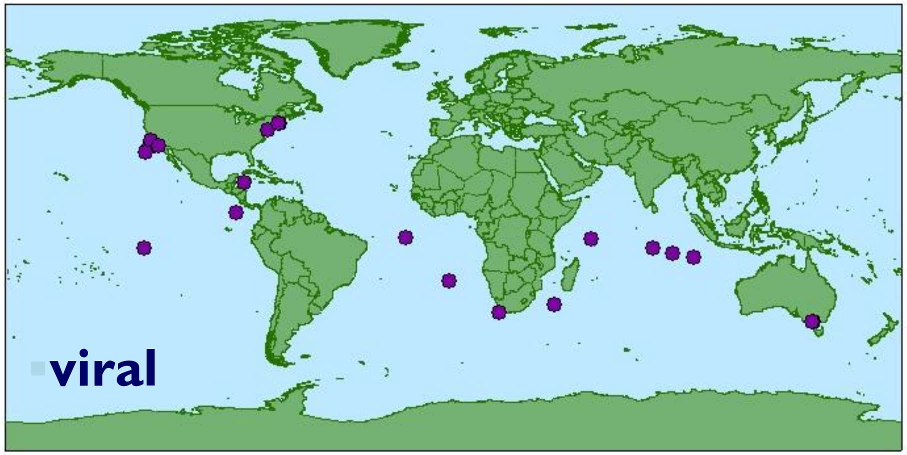
Sample Filtration

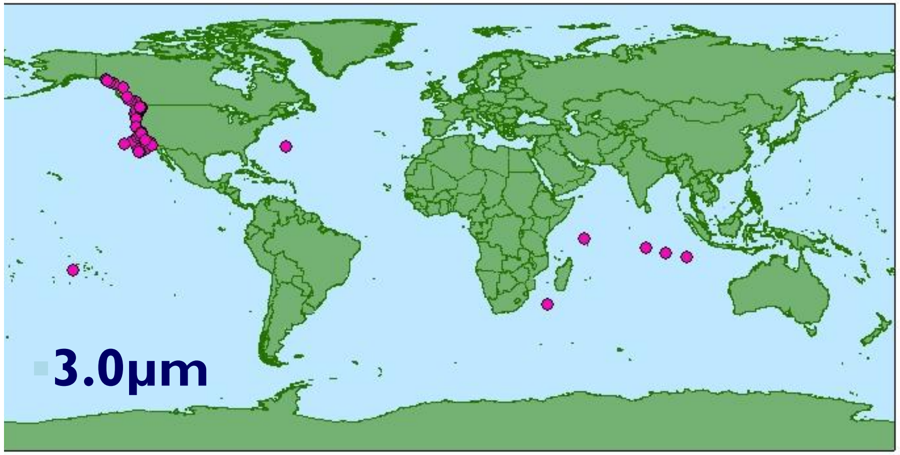


GOS circumnavigation data
229 stations and 291 samples





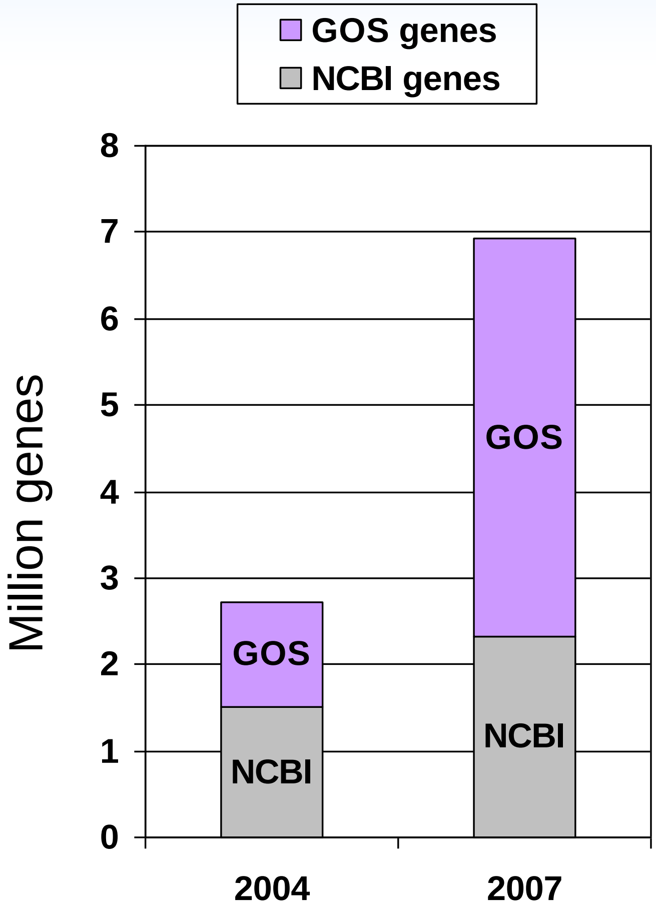


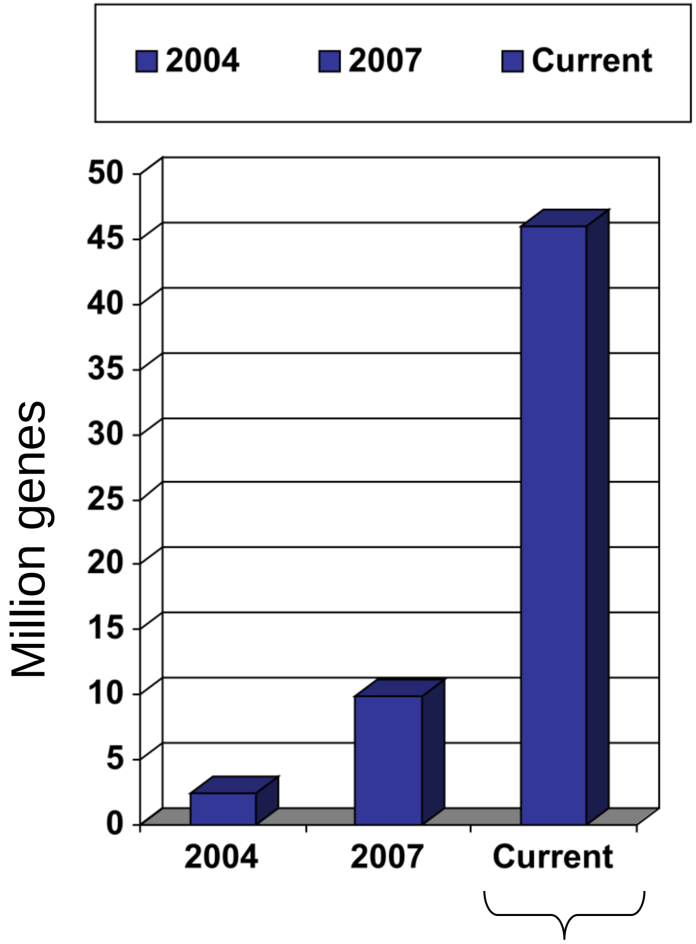


GOS data

	Reads	Proteins	Sequencing Technology
Phase I	7.6 Million	9.8 Million	Sanger
Circumnavigation	48 Million	~53Million	Sanger + 454

GOS dataset is expanding the protein universe



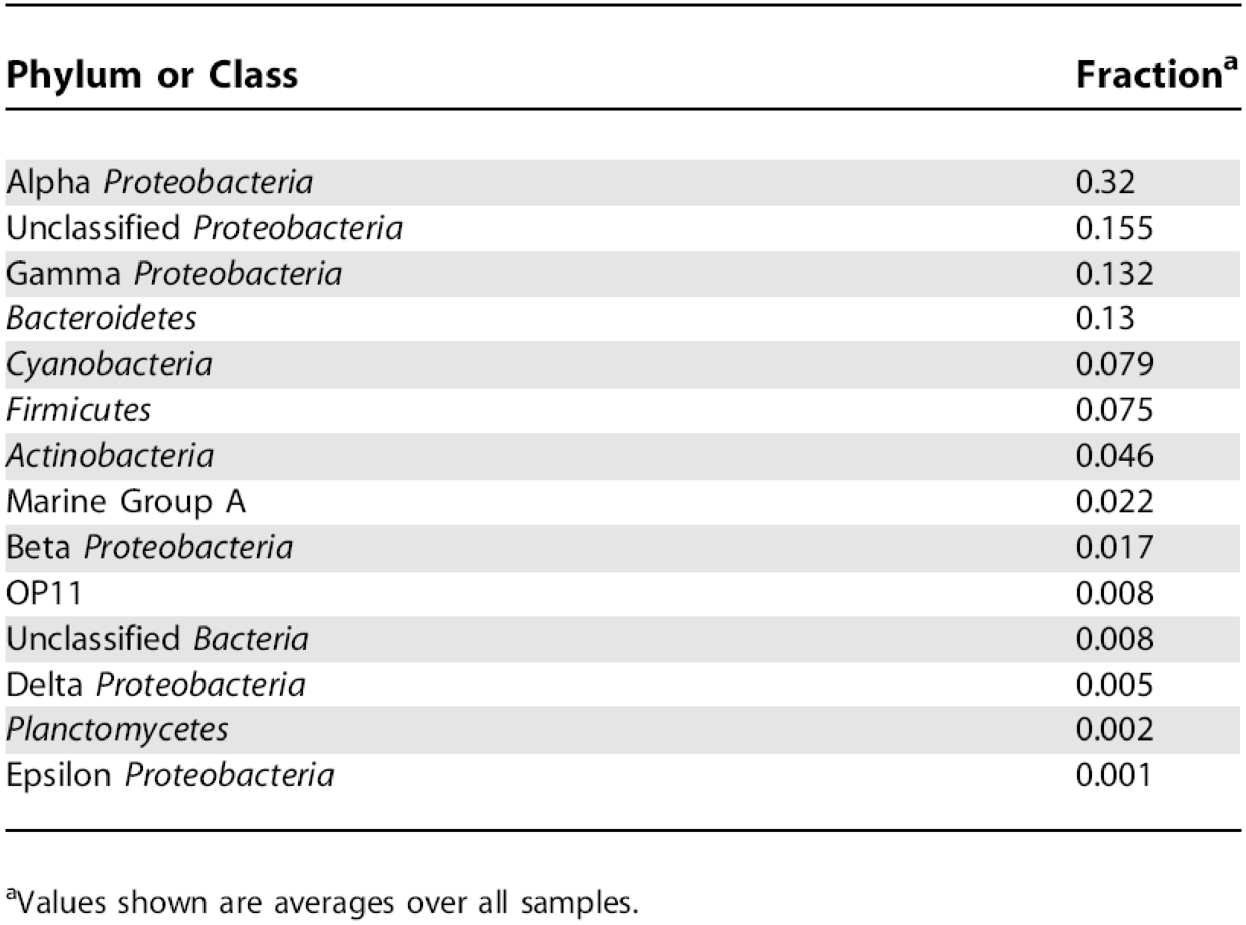


Extrapolation based on amount of GOS sequence data currently available but not yet released to public domain

J. Craig Venter INSTITUTE

Community makeup

Taxonomic makeup of GOS samples based on 16S data from shotgun sequencing

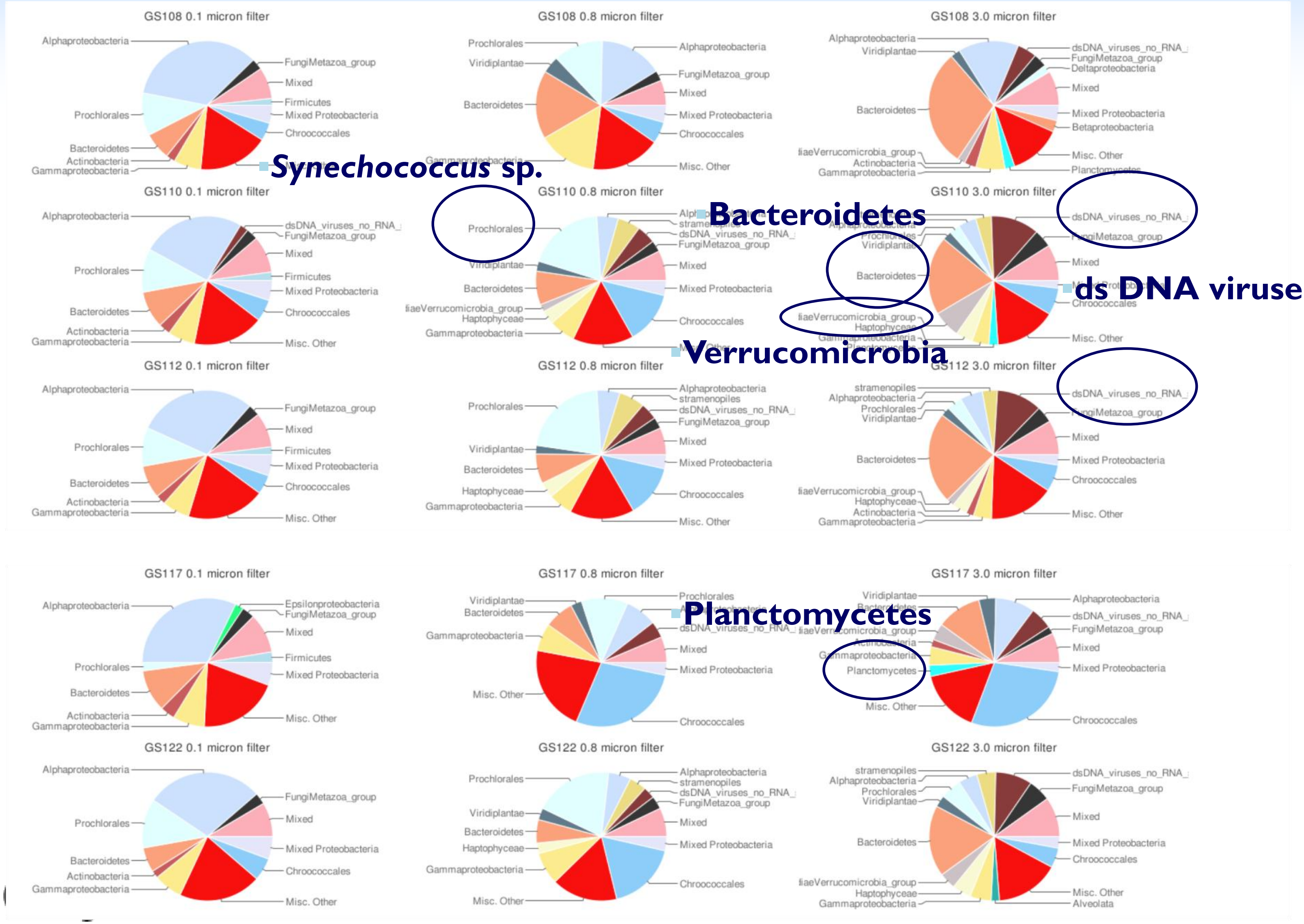


Phylum or Class	Fraction[a]
Alpha *Proteobacteria*	0.32
Unclassified *Proteobacteria*	0.155
Gamma *Proteobacteria*	0.132
Bacteroidetes	0.13
Cyanobacteria	0.079
Firmicutes	0.075
Actinobacteria	0.046
Marine Group A	0.022
Beta *Proteobacteria*	0.017
OP11	0.008
Unclassified *Bacteria*	0.008
Delta *Proteobacteria*	0.005
Planctomycetes	0.002
Epsilon *Proteobacteria*	0.001

[a] Values shown are averages over all samples.

Phylogenetic Distribution in the Indian Ocean across size-classes

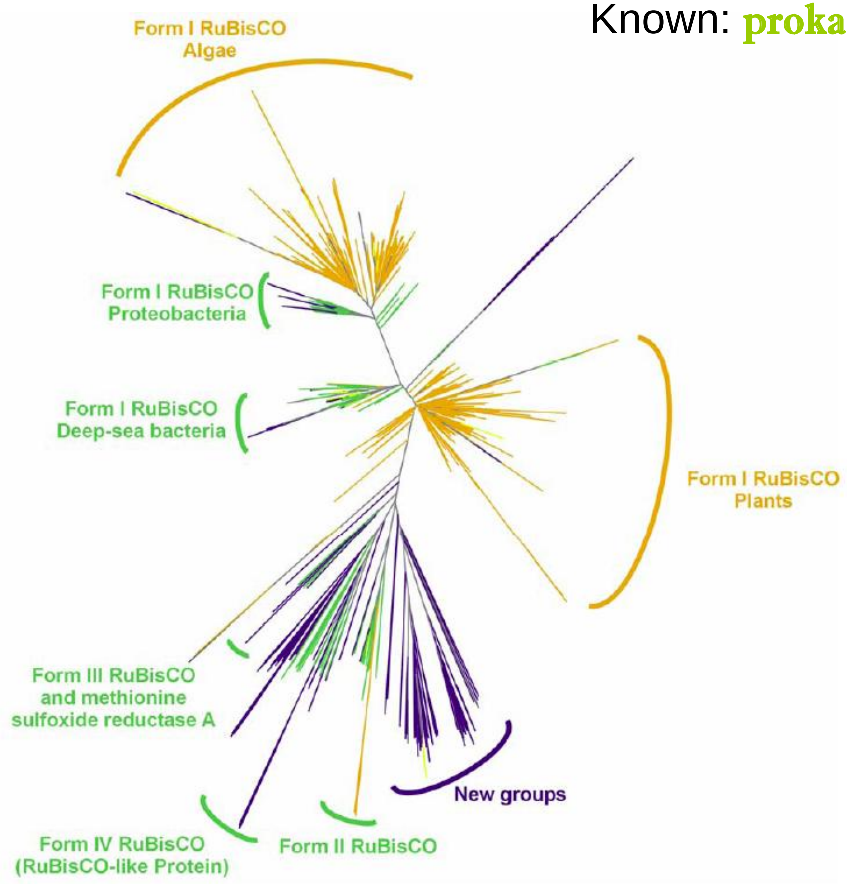
- 0.1 µm
- 0.8 µm
- 3.0 µm



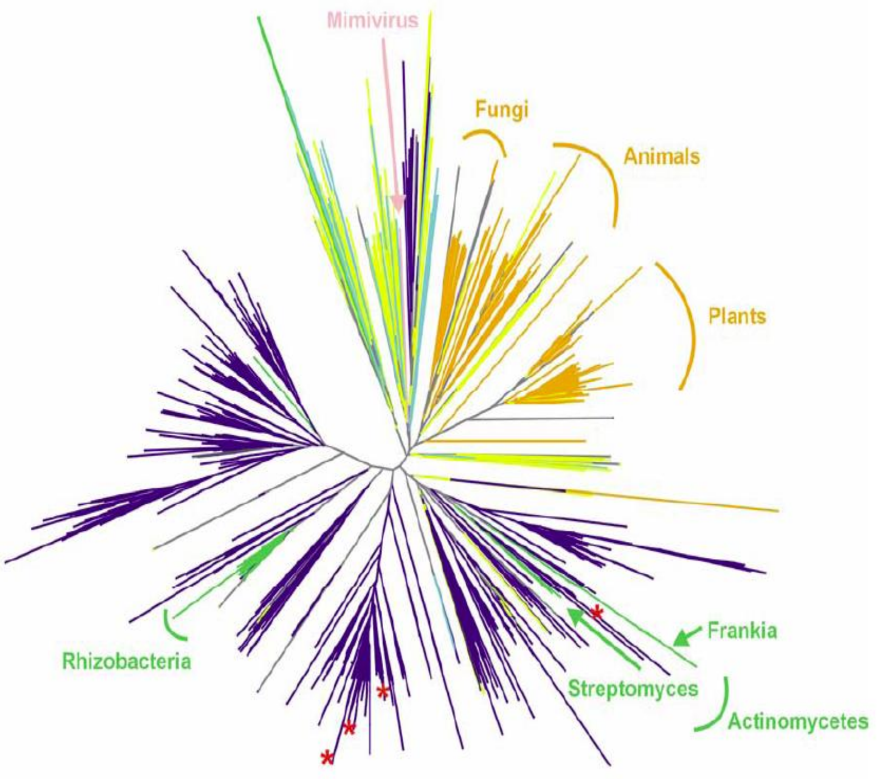
- *Synechococcus* sp.
- Bacteroidetes
- ds DNA viruses
- Verrucomicrobia
- Planctomycetes

GOS increases size and diversity of known protein families

GOS: prokaryotes, eukaryotes
Known: prokaryotes, eukaryotes



RuBisCO



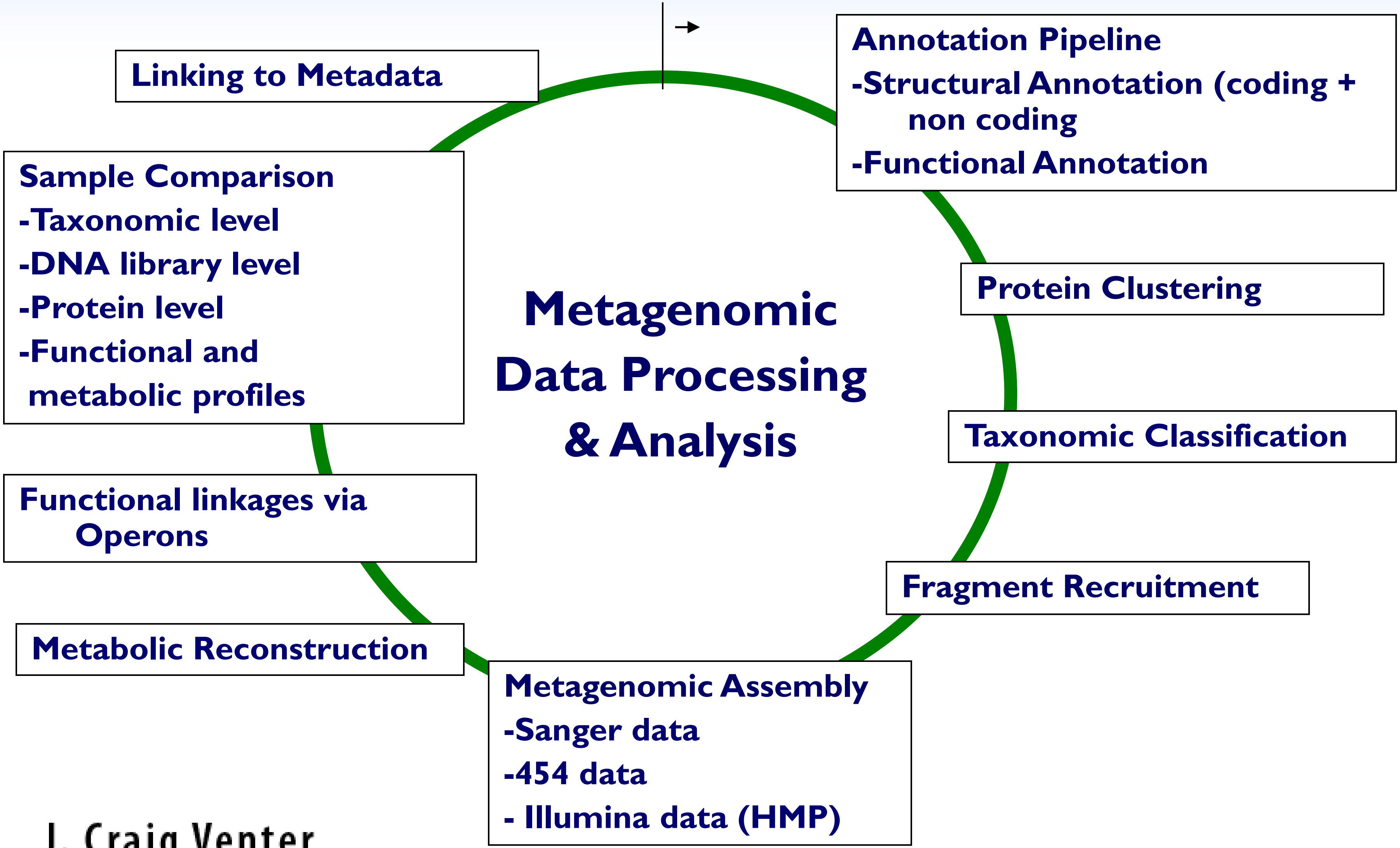
Glutamine synthetase (type II)

J. Craig Venter
INSTITUTE

Viruses in the Marine Environment

- Abundant: ~10^7 /ml^{-1} of surface seawater
- Diverse: VBR $\cong$ 10 ; ~ 10-fold greater diversity than microbial hosts
- Influence microbial diversity through infection and host cell lysis
- Mediators of horizontal gene transfer
- Influence biogeochemical cycling, particularly carbon

High-throughput Metagenomic Data Analysis



Metagenomic Data Processing & Analysis

Annotation Pipeline
-Structural Annotation (coding + non coding
-Functional Annotation

Protein Clustering

Taxonomic Classification

Fragment Recruitment

Metagenomic Assembly
-Sanger data
-454 data
- Illumina data (HMP)

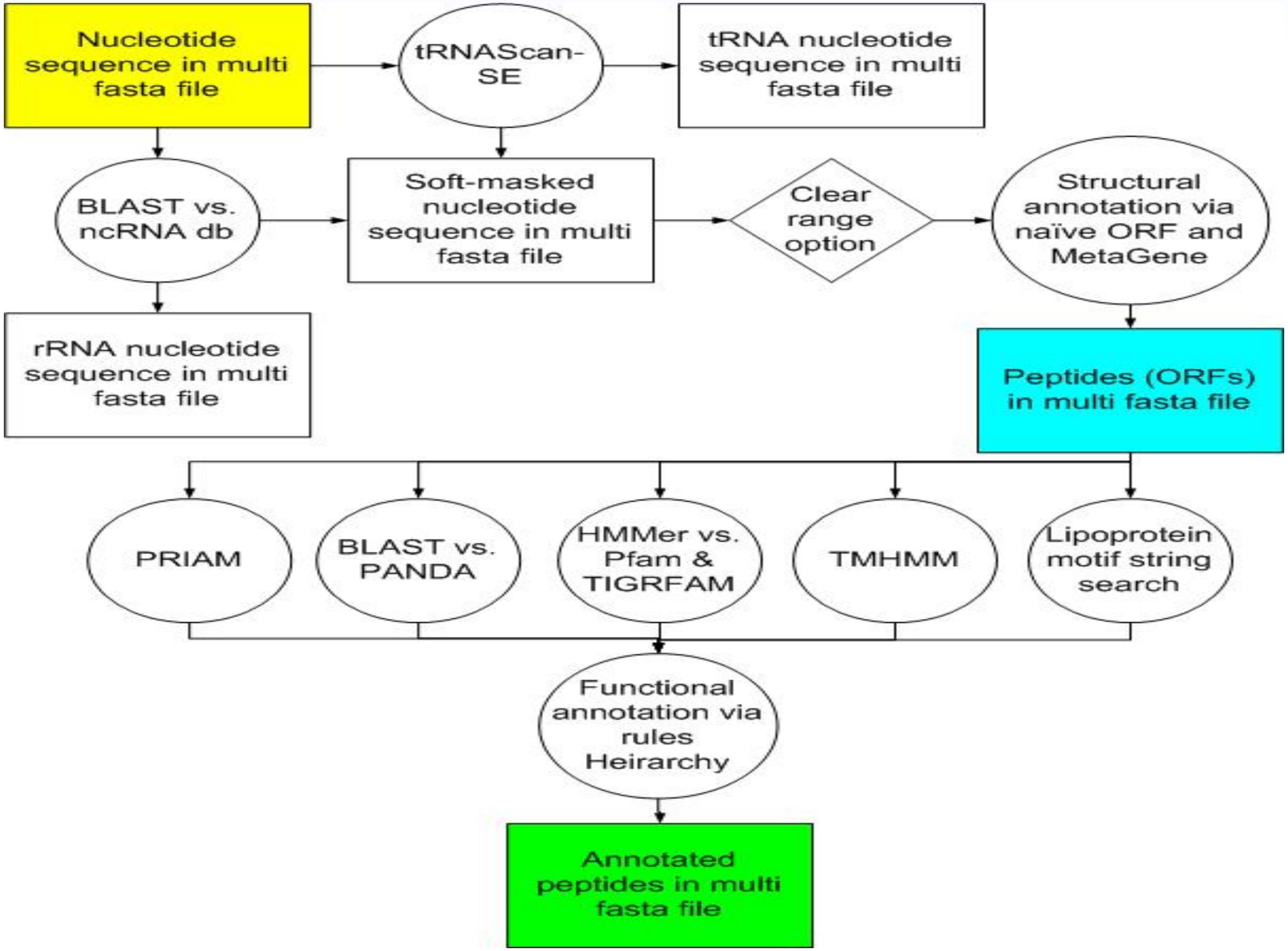
Metabolic Reconstruction

Functional linkages via Operons

Sample Comparison
-Taxonomic level
-DNA library level
-Protein level
-Functional and metabolic profiles

Linking to Metadata

Metagenomic Data Processing - Annotation pipeline


Nucleotide sequence in multi fasta file
tRNAScan-SE
tRNA nucleotide sequence in multi fasta file
BLAST vs. ncRNA db
Soft-masked nucleotide sequence in multi fasta file
Clear range option
Structural annotation via naïve ORF and MetaGene
rRNA nucleotide sequence in multi fasta file
Peptides (ORFs) in multi fasta file
PRIAM
BLAST vs. PANDA
HMMer vs. Pfam & TIGRFAM
TMHMM
Lipoprotein motif string search
Functional annotation via rules Heirarchy
Annotated peptides in multi fasta file

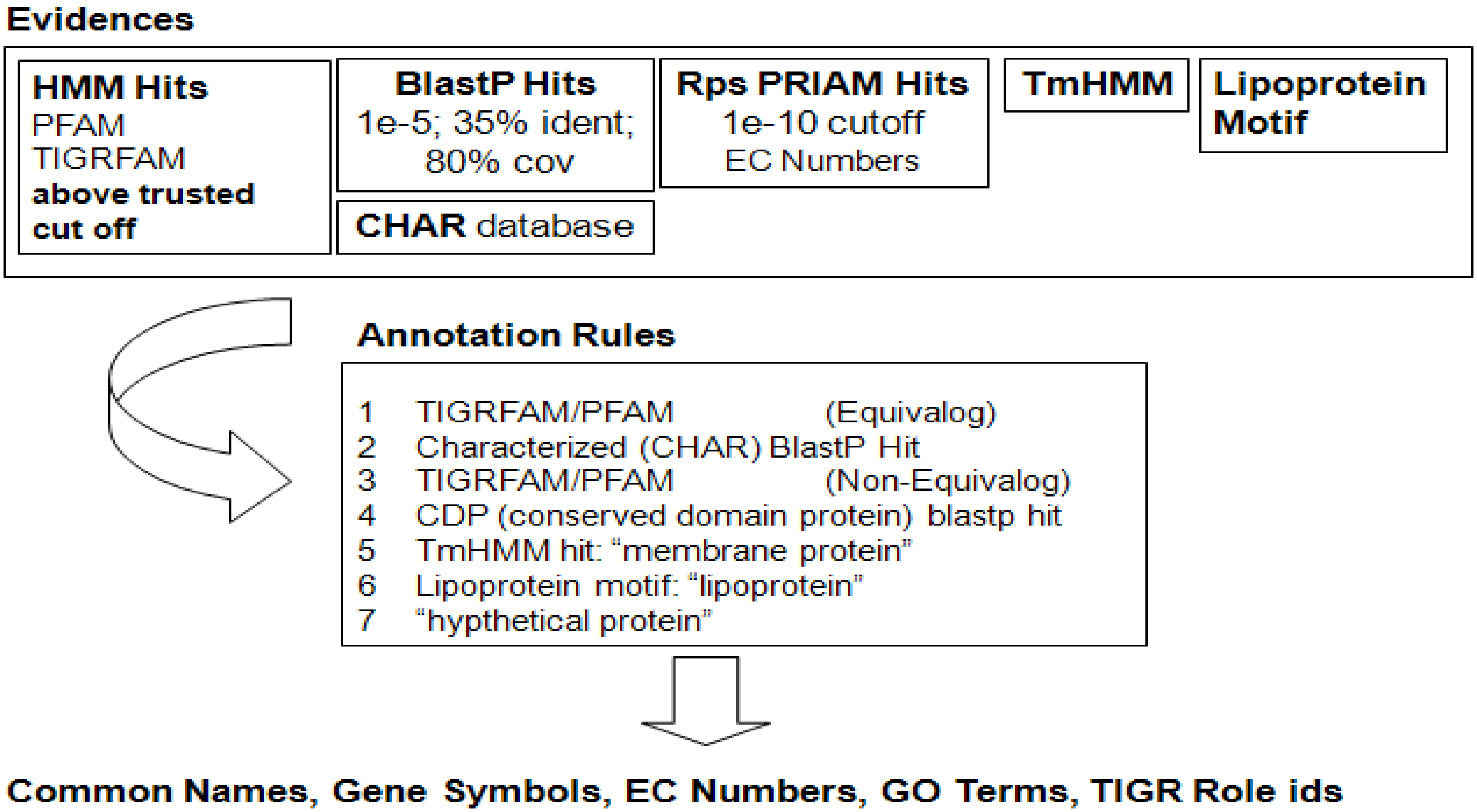
Structural Annotation

Functional Annotation

Published in SIGS

J. Craig Venter INSTITUTE

Annotation Rules Hierarchy



Evidences

HMM Hits
PFAM
TIGRFAM
above trusted cut off

BlastP Hits
1e-5; 35% ident; 80% cov

CHAR database

Rps PRIAM Hits
1e-10 cutoff
EC Numbers

TmHMM

Lipoprotein Motif

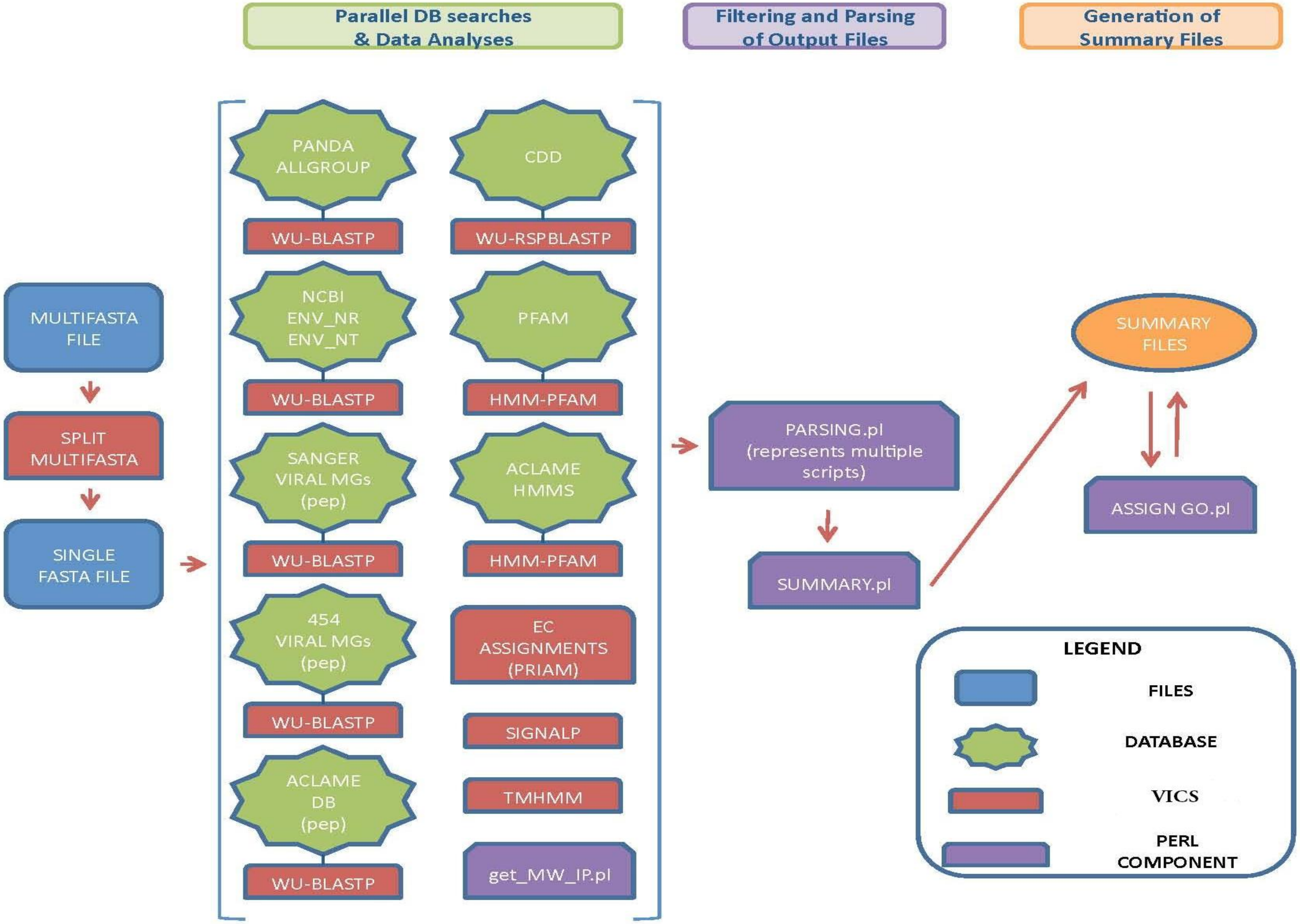
Annotation Rules

1 TIGRFAM/PFAM (Equivalog)
2 Characterized (CHAR) BlastP Hit
3 TIGRFAM/PFAM (Non-Equivalog)
4 CDP (conserved domain protein) blastp hit
5 TmHMM hit: "membrane protein"
6 Lipoprotein motif: "lipoprotein"
7 "hypthetical protein"

Common Names, Gene Symbols, EC Numbers, GO Terms, TIGR Role ids

J. Craig Venter
INSTITUTE

Viral Metagenomic (functional)Pipeline



J. Craig Venter INSTITUTE

Annotation Rules Hierarchy (Viral)

- PFAM/TIGRFAM_HMM, equivalog above trusted cutoff
- ACLAME_PEP, %id>= 50, coverage >= 80, e-value <= 10^{-10}
- ALLGROUP_PEP, %id>= 50, coverage >= 80, e-value <= 10^{-10}
- ACCLAME_HMM matches, > 90% coverage, e-value < 10^{-5}
- PFAM/TIGRFAM_HMM, non-equivalog above trusted cutoff
- CDD_RPS, %id>= 35%, coverage >= 90% of CDD-domain, e-value <= $1e^{-10}$
- FRAG_HMM, e-value < $1e^{-5}$
- ACLAME_PEP, %id >= 30%, coverage >= 70%, e-value <= $1e^{-5}$
- ALLGROUP_PEP, %id >= 30%, coverage >= 70%, e-value <= $1e^{-5}$
- No evidence -> hypothetical protein

Metagenomic Assembly

Advantages

- Provides genomic context
- Reduces redundancy and complexity
- Improves annotation
- Mechanism to isolate environment specific gene regions

Challenges

- Coverage dependent
- Variation can limit the length of assemblies
- Can mask diversity

•Celera Hybrid Assembler has been updated to work with 454 Titanium reads

•Will further optimize assembly process to capture environmental diversity

Metagenomic Data Processing - Continued

- **Protein Clustering** : JCVI's Protein clustering (S. Yooseph)
- **Taxonomic Classification** : APIS (J. Badger)
- **Fragment Recruitment** :Advanced Reference Viewer (D. Rusch)
- **Metagenomic Assembly** : Celera Assembler (G. Sutton & J. Miller)
- **Sample Comparison**

Making sense of everything in the context of **METADATA**

General Questions

- Who are they?

  Species , Taxonomic distribution…

- How many?

  Distribution across sites and filters

- What are they doing?

  Functional profiles

  Metabolic profiles

MR Specific Questions

- Metabolic profiles across sites and filters
- Pathways coverage and abundance
- What known characterized pathways and how many?
- What novel pathways are there?
- Metabolic network

Metabolic Reconstruction

- From the Annotation Pipeline (orf based)

Proteins → EC assignment→ Pathways prediction

(EC to MetaCyc/Kegg mapping)

Sources for EC : TIGRFAM
PFAM
High confidence blast hit to Uniref100/Panda
RPSblast to EC profiles from PRIAM

- From BlastX to a Functional database (read based)

Reads → Blastx Metacyc/Kegg → Pathways prediction

Browse/analyze/compare pathways across datasets in the context of annotation and Metadata

METAREP
JCVI Metagenomics Reports

website www.jcvi.org/metarep
source code http://github.com/jcvi/METAREP
blog http://blogs.jcvi.org/tag/metarep
contact metarep-support@jcvi.org

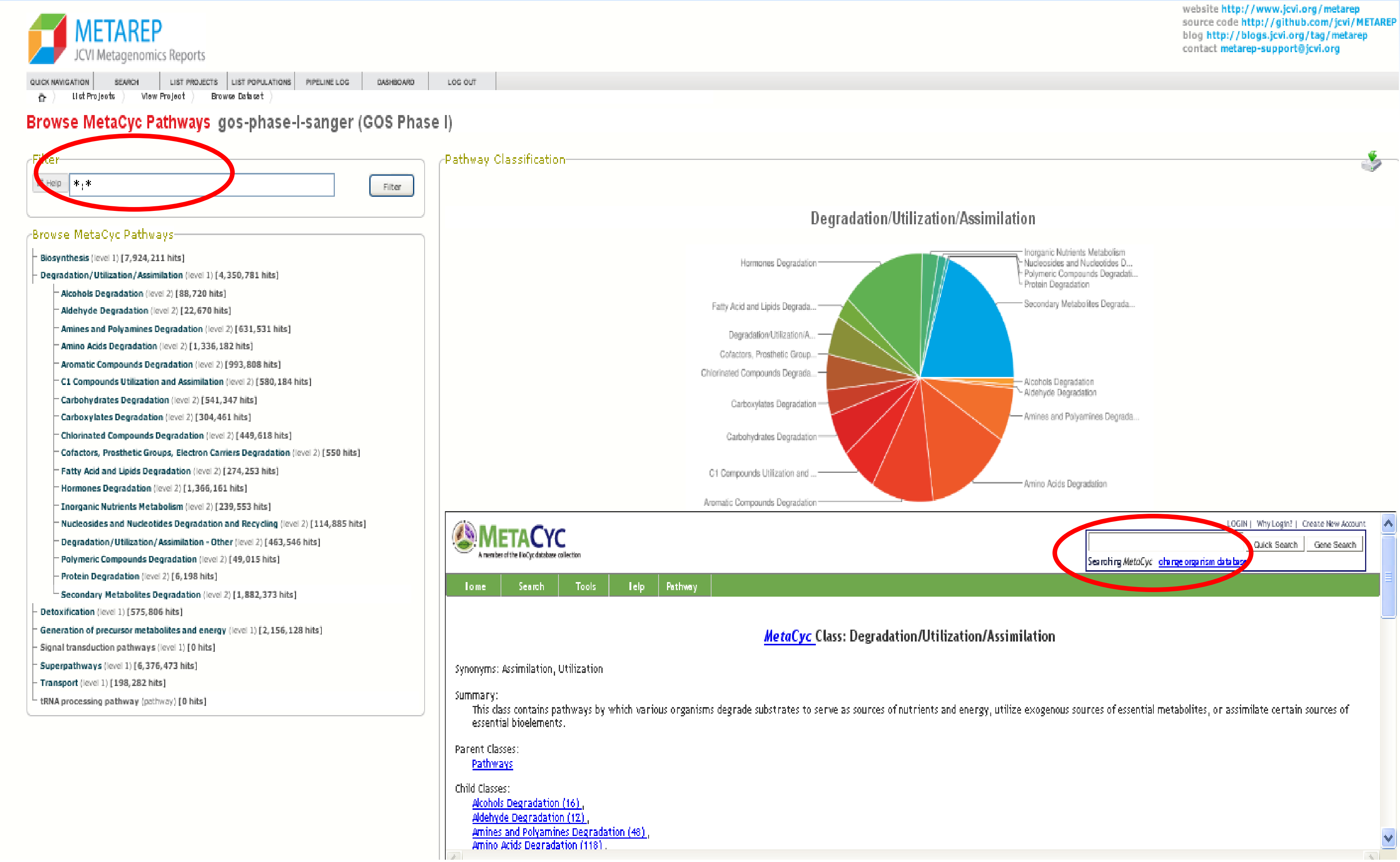
METAREP is a web interface designed to help scientists to view, query and compare annotation data derived from proteins called on metagenomics reads

Developer : Johannes Goll
Published in Bioinformatics

www.jcvi.org/metarep

J. Craig Venter
INSTITUTE

Browse pathways


METAREP
JCVI Metagenomics Reports
QUICK NAVIGATION
SEARCH
LIST PROJECTS
LIST POPULATIONS
PIPELINE LOG
DASHBOARD
LOG OUT
List Projects
View Project
Browse Dataset
Filter
Help
Filter
Pathway Classification
Degradation/Utilization/Assimilation
Inorganic Nutrients Metabolism
Nucleosides and Nucleotides D...
Polymeric Compounds Degradati...
Protein Degradation
Secondary Metabolites Degrada...
Hormones Degradation
Fatty Acid and Lipids Degrada...
Degradation/Utilization/A...
Cofactors, Prosthetic Group...
Chlorinated Compounds Degrada...
Carboxylates Degradation
Carbohydrates Degradation
C1 Compounds Utilization and ...
Aromatic Compounds Degradation
Alcohols Degradation
Aldehyde Degradation
Amines and Polyamines Degrada...
Amino Acids Degradation
METACYC
A member of the BioCyc database collection
LOGIN | Why Login? | Create New Account
Quick Search
Gene Search
Searching MetaCyc change organism database
Home
Search
Tools
Help
Pathway

website http://www.jcvi.org/metarep
source code http://github.com/jcvi/METAREP
blog http://blogs.jcvi.org/tag/metarep
contact metarep-support@jcvi.org

Browse MetaCyc Pathways gos-phase-I-sanger (GOS Phase I)

Filter: *;*

Browse MetaCyc Pathways

- Biosynthesis (level 1) [7,924,211 hits]
- Degradation/Utilization/Assimilation (level 1) [4,350,781 hits]
  - Alcohols Degradation (level 2) [88,720 hits]
  - Aldehyde Degradation (level 2) [22,670 hits]
  - Amines and Polyamines Degradation (level 2) [631,531 hits]
  - Amino Acids Degradation (level 2) [1,336,182 hits]
  - Aromatic Compounds Degradation (level 2) [993,808 hits]
  - C1 Compounds Utilization and Assimilation (level 2) [580,184 hits]
  - Carbohydrates Degradation (level 2) [541,347 hits]
  - Carboxylates Degradation (level 2) [304,461 hits]
  - Chlorinated Compounds Degradation (level 2) [449,618 hits]
  - Cofactors, Prosthetic Groups, Electron Carriers Degradation (level 2) [550 hits]
  - Fatty Acid and Lipids Degradation (level 2) [274,253 hits]
  - Hormones Degradation (level 2) [1,366,161 hits]
  - Inorganic Nutrients Metabolism (level 2) [239,553 hits]
  - Nucleosides and Nucleotides Degradation and Recycling (level 2) [114,885 hits]
  - Degradation/Utilization/Assimilation - Other (level 2) [463,546 hits]
  - Polymeric Compounds Degradation (level 2) [49,015 hits]
  - Protein Degradation (level 2) [6,198 hits]
  - Secondary Metabolites Degradation (level 2) [1,882,373 hits]
- Detoxification (level 1) [575,806 hits]
- Generation of precursor metabolites and energy (level 1) [2,156,128 hits]
- Signal transduction pathways (level 1) [0 hits]
- Superpathways (level 1) [6,376,473 hits]
- Transport (level 1) [198,282 hits]
- tRNA processing pathway (pathway) [0 hits]

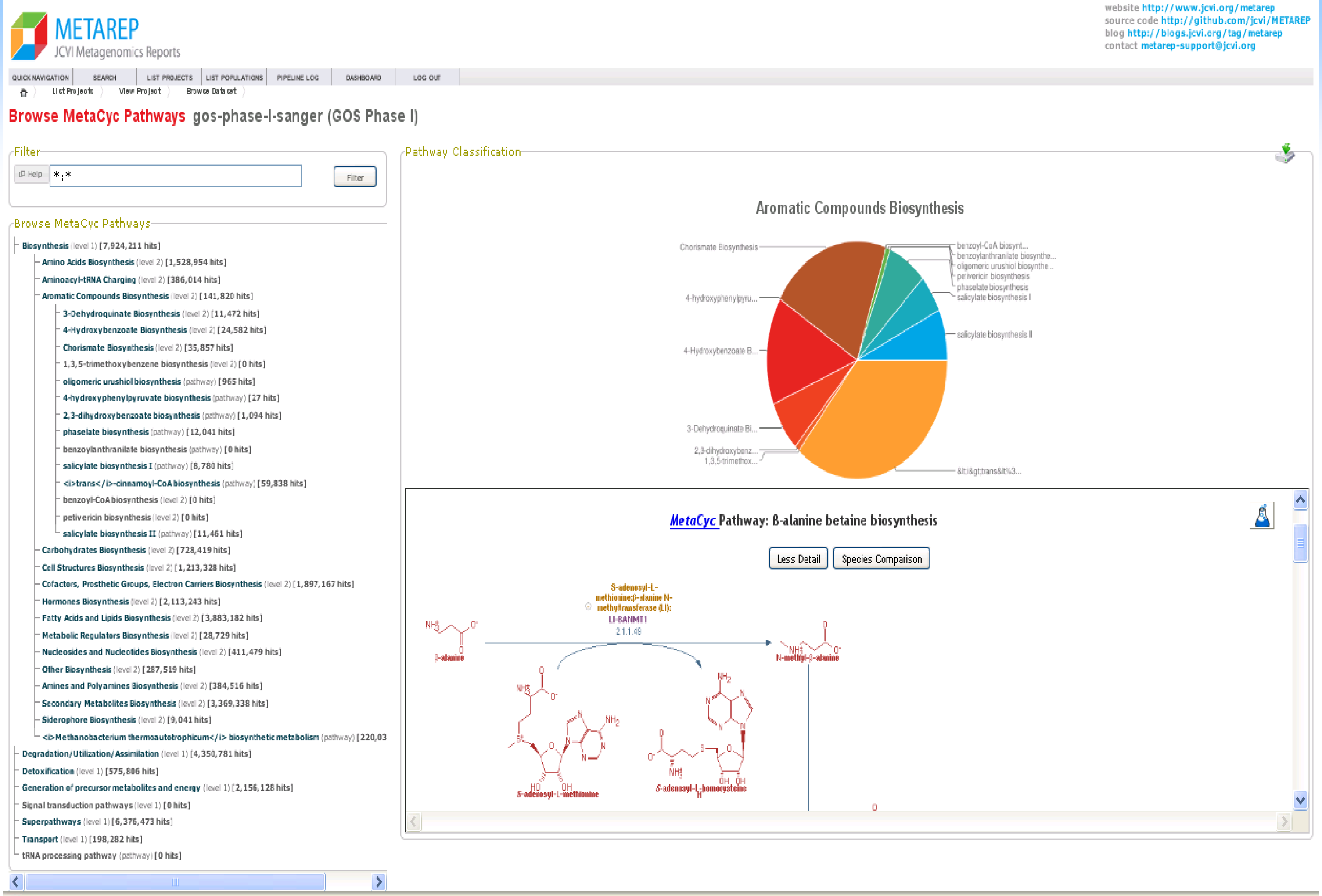
MetaCyc Class: Degradation/Utilization/Assimilation

Synonyms: Assimilation, Utilization

Summary:
This class contains pathways by which various organisms degrade substrates to serve as sources of nutrients and energy, utilize exogenous sources of essential metabolites, or assimilate certain sources of essential bioelements.

Parent Classes:
Pathways

Child Classes:
Alcohols Degradation (16) ,
Aldehyde Degradation (12) ,
Amines and Polyamines Degradation (48) ,
Amino Acids Degradation (118) ,



website http://www.jcvi.org/metarep
source code http://github.com/jcvi/METAREP
blog http://blogs.jcvi.org/tag/metarep
contact metarep-support@jcvi.org

Browse MetaCyc Pathways gos-phase-I-sanger (GOS Phase I)

Filter

:

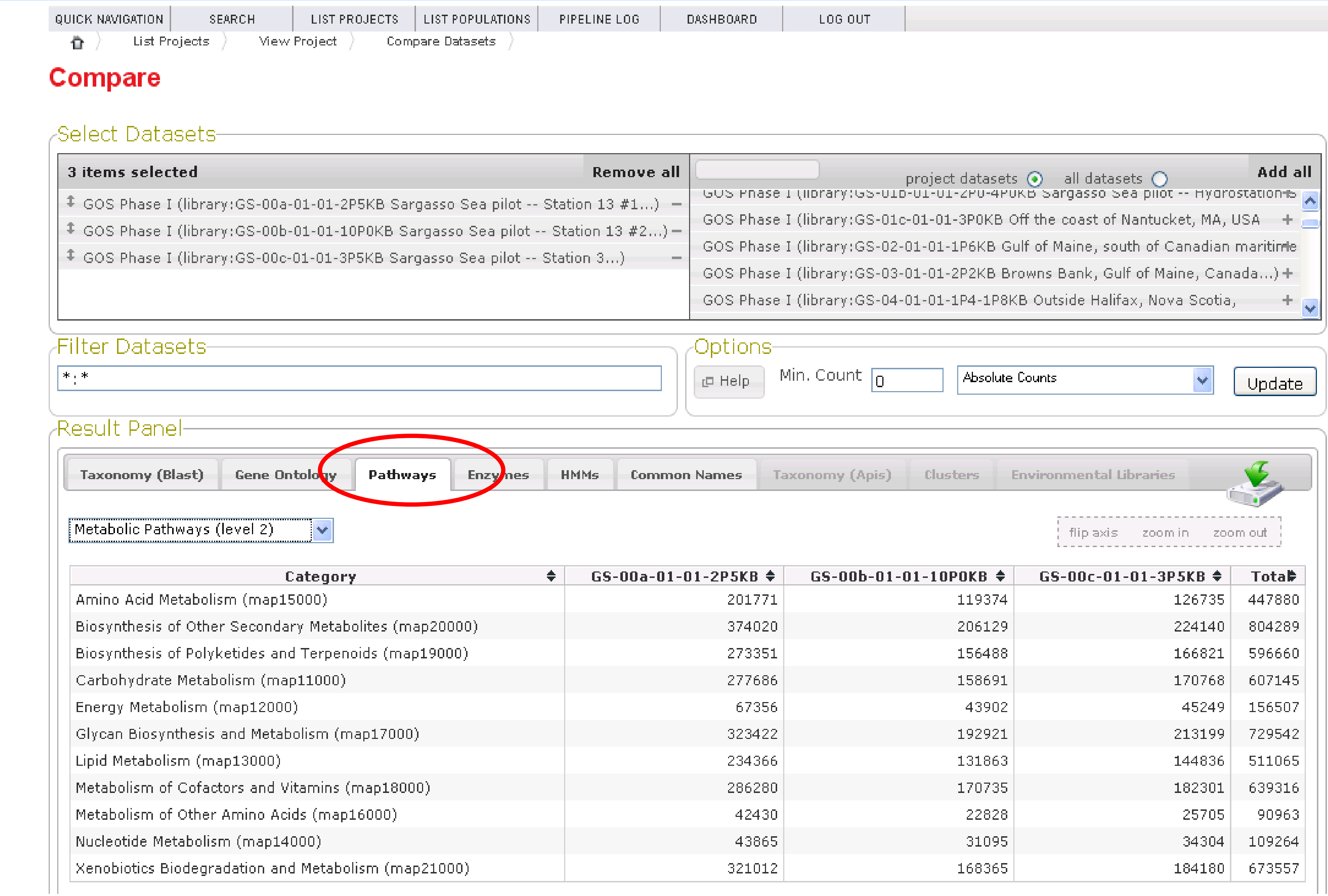
Browse MetaCyc Pathways

- Biosynthesis (level 1) [7,924,211 hits]
  - Amino Acids Biosynthesis (level 2) [1,528,954 hits]
  - Aminoacyl-tRNA Charging (level 2) [386,014 hits]
  - Aromatic Compounds Biosynthesis (level 2) [141,820 hits]
    - 3-Dehydroquinate Biosynthesis (level 2) [11,472 hits]
    - 4-Hydroxybenzoate Biosynthesis (level 2) [24,582 hits]
    - Chorismate Biosynthesis (level 2) [35,857 hits]
    - 1,3,5-trimethoxybenzene biosynthesis (level 2) [0 hits]
    - oligomeric urushiol biosynthesis (pathway) [965 hits]
    - 4-hydroxyphenylpyruvate biosynthesis (pathway) [27 hits]
    - 2,3-dihydroxybenzoate biosynthesis (pathway) [1,094 hits]
    - phaselate biosynthesis (pathway) [12,041 hits]
    - benzoylanthranilate biosynthesis (pathway) [0 hits]
    - salicylate biosynthesis I (pathway) [8,780 hits]
    - <i>trans</i>-cinnamoyl-CoA biosynthesis (pathway) [59,838 hits]
    - benzoyl-CoA biosynthesis (level 2) [0 hits]
    - petivericin biosynthesis (level 2) [0 hits]
    - salicylate biosynthesis II (pathway) [11,461 hits]
  - Carbohydrates Biosynthesis (level 2) [728,419 hits]
  - Cell Structures Biosynthesis (level 2) [1,213,328 hits]
  - Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis (level 2) [1,897,167 hits]
  - Hormones Biosynthesis (level 2) [2,113,243 hits]
  - Fatty Acids and Lipids Biosynthesis (level 2) [3,883,182 hits]
  - Metabolic Regulators Biosynthesis (level 2) [28,729 hits]
  - Nucleosides and Nucleotides Biosynthesis (level 2) [411,479 hits]
  - Other Biosynthesis (level 2) [287,519 hits]
  - Amines and Polyamines Biosynthesis (level 2) [384,516 hits]
  - Secondary Metabolites Biosynthesis (level 2) [3,369,338 hits]
  - Siderophore Biosynthesis (level 2) [9,041 hits]
  - <i>Methanobacterium thermoautotrophicum</i> biosynthetic metabolism (pathway) [220,03
- Degradation/Utilization/Assimilation (level 1) [4,350,781 hits]
- Detoxification (level 1) [575,806 hits]
- Generation of precursor metabolites and energy (level 1) [2,156,128 hits]
- Signal transduction pathways (level 1) [0 hits]
- Superpathways (level 1) [6,376,473 hits]
- Transport (level 1) [198,282 hits]
- tRNA processing pathway (pathway) [0 hits]

Pathway Classification

MetaCyc Pathway: β-alanine betaine biosynthesis

Compare pathways across datasets



Compare

Select Datasets

3 items selected — Remove all

- GOS Phase I (library:GS-00a-01-01-2P5KB Sargasso Sea pilot -- Station 13 #1...)
- GOS Phase I (library:GS-00b-01-01-10P0KB Sargasso Sea pilot -- Station 13 #2...)
- GOS Phase I (library:GS-00c-01-01-3P5KB Sargasso Sea pilot -- Station 3...)

project datasets ● all datasets ○ — Add all

- GOS Phase I (library:GS-01b-01-01-2P0-4P0KB Sargasso Sea pilot -- Hydrostation S
- GOS Phase I (library:GS-01c-01-01-3P0KB Off the coast of Nantucket, MA, USA
- GOS Phase I (library:GS-02-01-01-1P6KB Gulf of Maine, south of Canadian maritime
- GOS Phase I (library:GS-03-01-01-2P2KB Browns Bank, Gulf of Maine, Canada...)
- GOS Phase I (library:GS-04-01-01-1P4-1P8KB Outside Halifax, Nova Scotia,

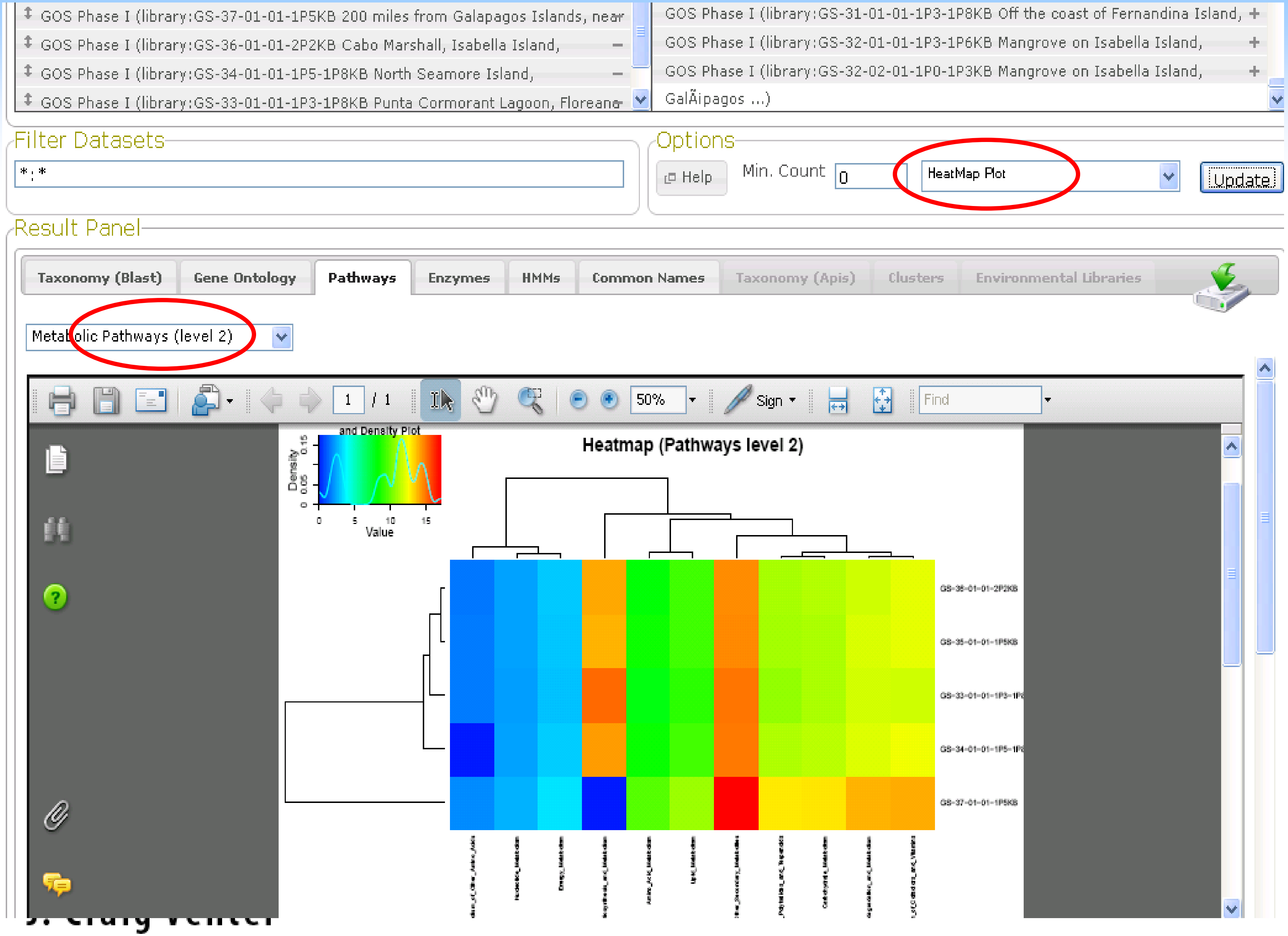
Filter Datasets: *:*

Options: Help — Min. Count: 0 — Absolute Counts — Update

Result Panel

Metabolic Pathways (level 2)

Category	GS-00a-01-01-2P5KB	GS-00b-01-01-10P0KB	GS-00c-01-01-3P5KB	Total
Amino Acid Metabolism (map15000)	201771	119374	126735	447880
Biosynthesis of Other Secondary Metabolites (map20000)	374020	206129	224140	804289
Biosynthesis of Polyketides and Terpenoids (map19000)	273351	156488	166821	596660
Carbohydrate Metabolism (map11000)	277686	158691	170768	607145
Energy Metabolism (map12000)	67356	43902	45249	156507
Glycan Biosynthesis and Metabolism (map17000)	323422	192921	213199	729542
Lipid Metabolism (map13000)	234366	131863	144836	511065
Metabolism of Cofactors and Vitamins (map18000)	286280	170735	182301	639316
Metabolism of Other Amino Acids (map16000)	42430	22828	25705	90963
Nucleotide Metabolism (map14000)	43865	31095	34304	109264
Xenobiotics Biodegradation and Metabolism (map21000)	321012	168365	184180	673557



GOS Phase I (library:GS-37-01-01-1P5KB 200 miles from Galapagos Islands, near

GOS Phase I (library:GS-36-01-01-2P2KB Cabo Marshall, Isabella Island,

GOS Phase I (library:GS-34-01-01-1P5-1P8KB North Seamore Island,

GOS Phase I (library:GS-33-01-01-1P3-1P8KB Punta Cormorant Lagoon, Floreana

GOS Phase I (library:GS-31-01-01-1P3-1P8KB Off the coast of Fernandina Island,

GOS Phase I (library:GS-32-01-01-1P3-1P6KB Mangrove on Isabella Island,

GOS Phase I (library:GS-32-02-01-1P0-1P3KB Mangrove on Isabella Island,

GalÃipagos ...)

Filter Datasets: *;*

Options

Min. Count: 0

HeatMap Plot

Result Panel

Metabolic Pathways (level 2)

Heatmap (Pathways level 2)

Pathways Tools for GOS

- Metagenomic specific predictions - Incorporate taxonomic resolution when predicting pathways
- Confidence Scores for the pathways
- Incorporate more annotation evidence types in predictions other than EC
- Ability to overlay and visualize expression data
- Full integration of pathways tools into Metarep
- Performance enhancements to handle metagenomic data volume

Conclusion

- Who are they?

  Species , Taxonomic distribution…

- How many?

  Distribution across sites and filters

- What are they doing?

  Functional profiles

  Metabolic profiles

Acknowledgements

Metagenomic PI's & Coordinators

Shibu Yooseph
Barbara Methe

Metagenomic Bioinformatics & Software Engineers

Johannes Goll
Jeff Hoover
Alex Richter
Aaron Tenney
Daniel Brami
Monika Bihan
Kelvin Li

GOS Funded by

DOE Genomics: GTL Program
Gordon and Betty Moore Foundation
J. Craig Venter Science Foundation

Metagenomic PI's

Doug Rusch
Andy Allen
Shannon Williamson
Andrey Tovtchigretchko
Jonathan Badger

Postdocs

Seung-Jin Sul
Youngik Yang

Leadership

Robert Friedman, Karen Nelson
& J. Craig Venter

J. Craig Venter
INSTITUTE

Questions

Thank You