*Environmental Proteomics Conference, Keystone, CO*
*January 2010*

# Approaches to Microbial Community Proteomics Data Analysis

**Angela D. Norbeck**
Senior Scientist
Pacific Northwest National Laboratory

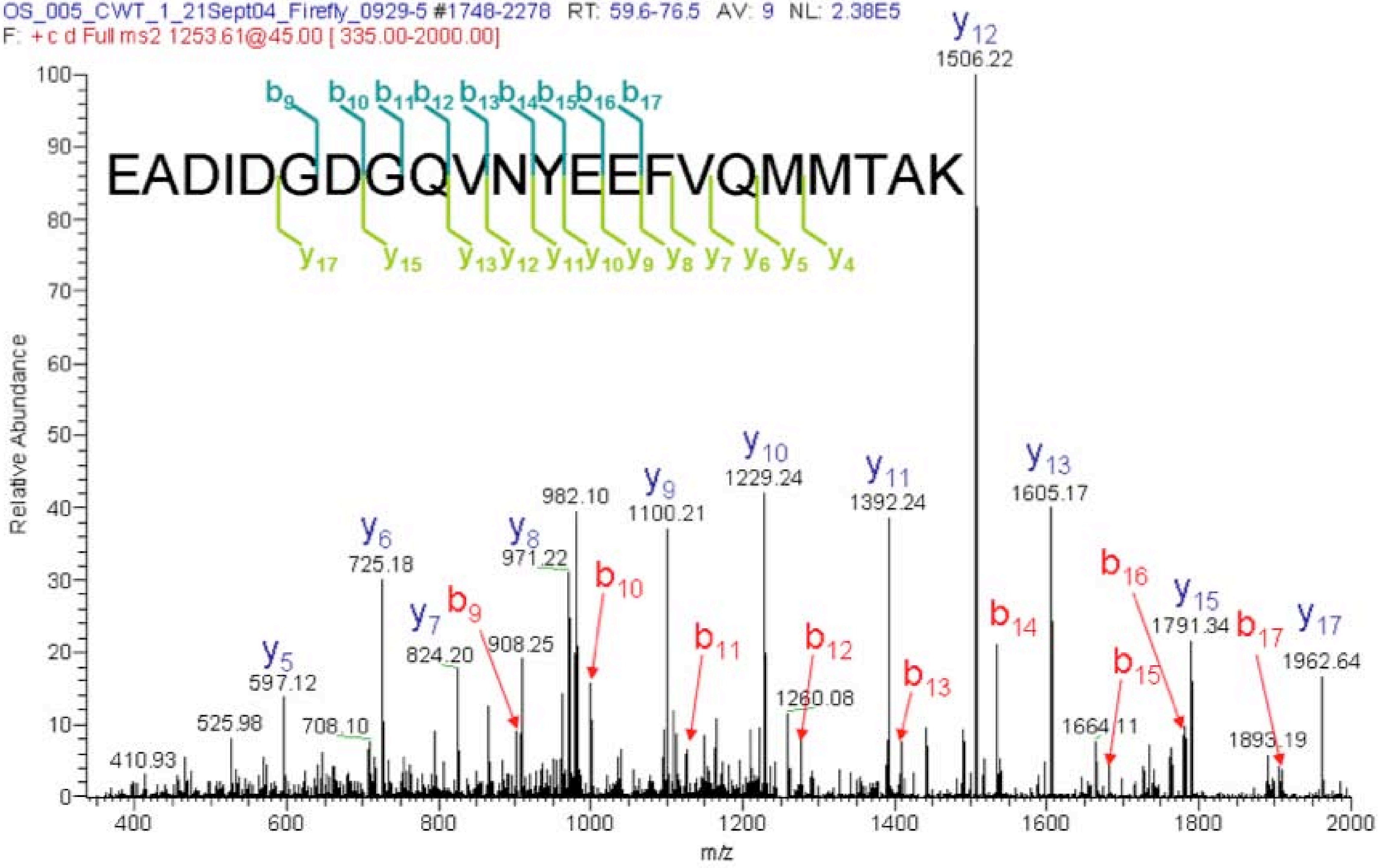
Pacific Northwest
NATIONAL LABORATORY

*Proudly Operated by* Battelle *Since 1965*

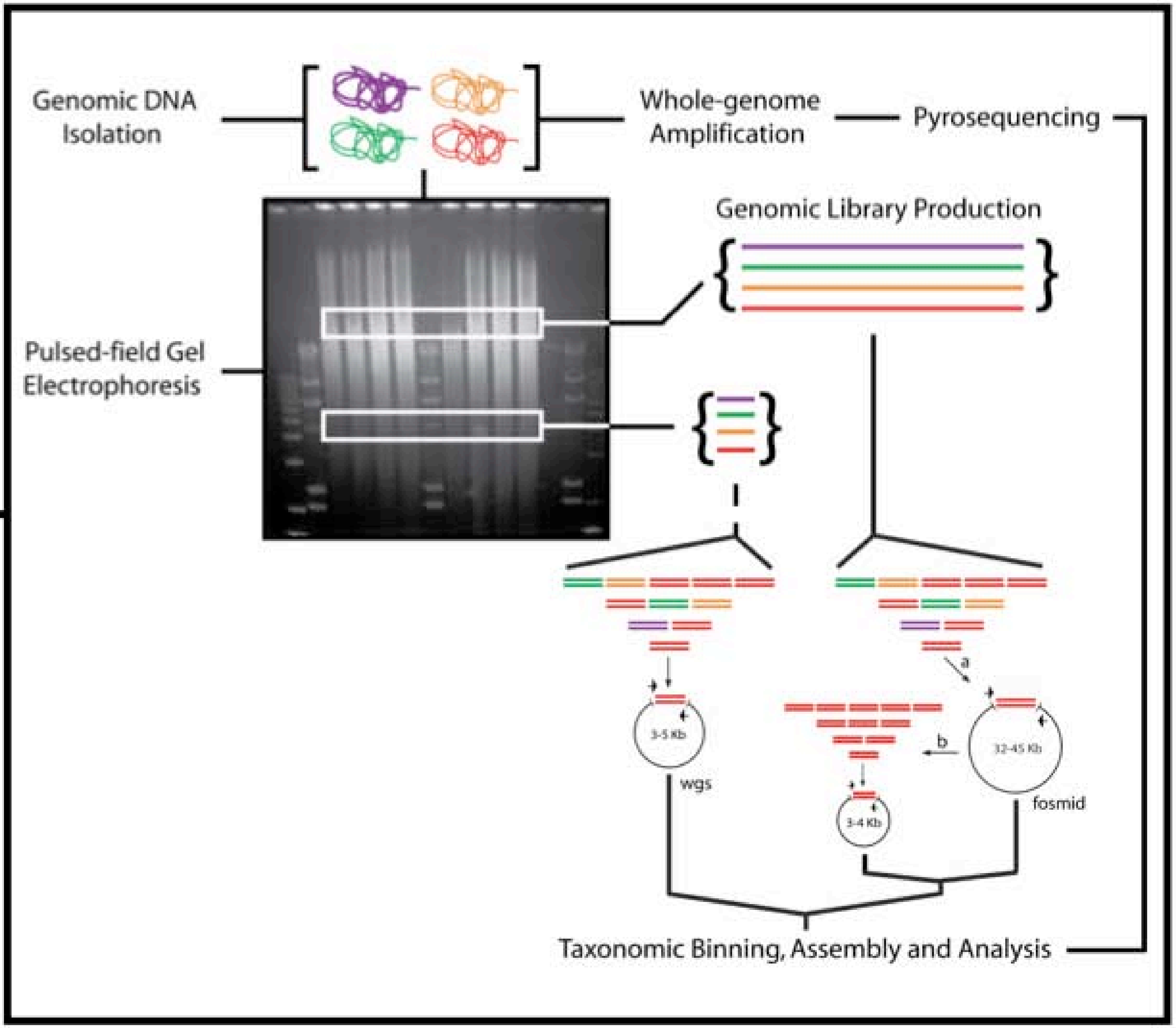
# Challenges

- **Extracting protein from samples**
  - Sample prep method development
- **Protein file needed for searching spectra**
  - Sequences from metagenome sequencing
  - Groups of annotated organism files
  - *In-silico* derived sequences
- **Finding proteins of interest within large results set**
  - Data analysis methods

# Protein files: Matching fragmentation (MS/MS) spectra to protein sequence files



# Metagenomic sequences: library production



Genomic DNA Isolation

Whole-genome Amplification

Pyrosequencing

Genomic Library Production

Pulsed-field Gel Electrophoresis

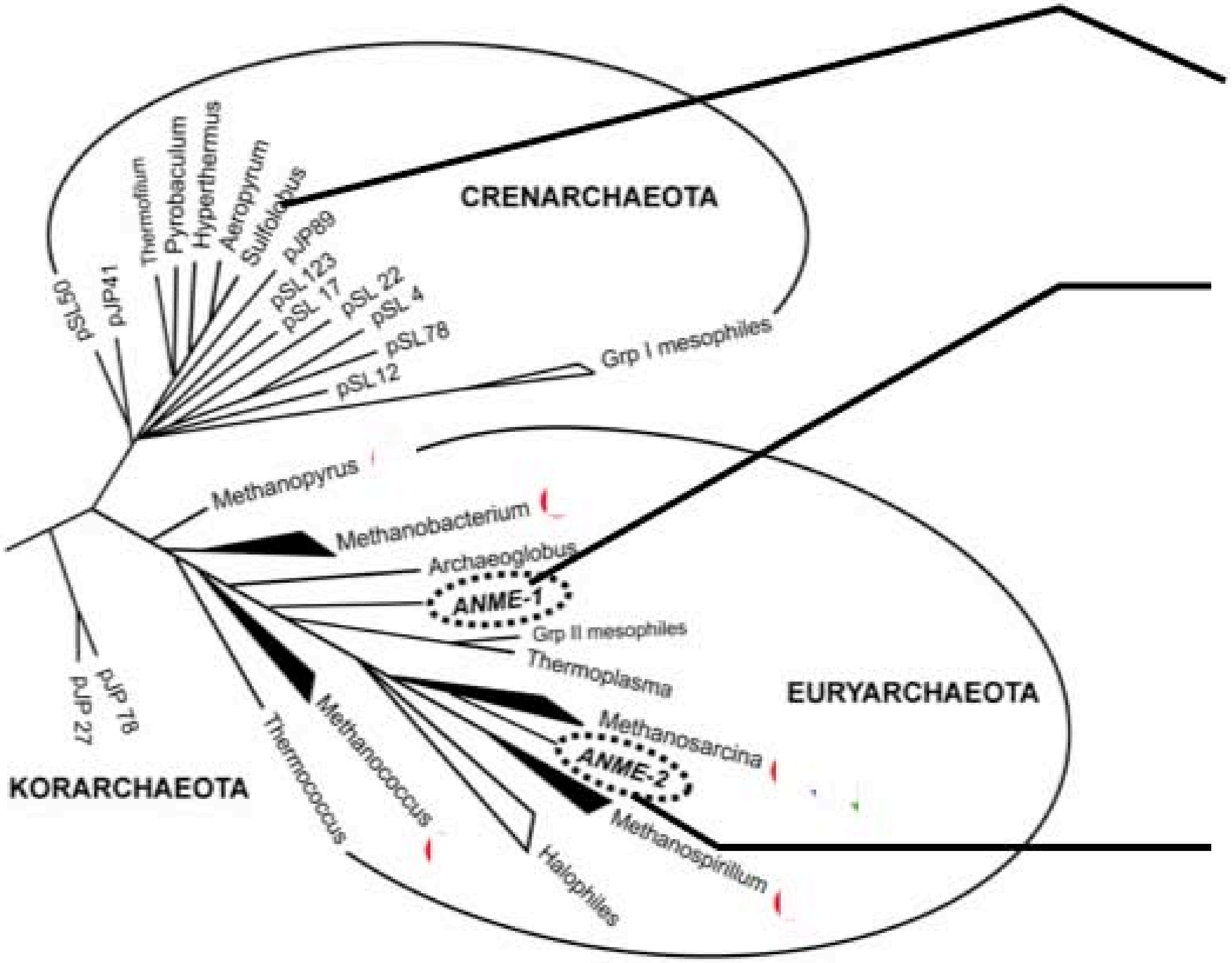
Taxonomic Binning, Assembly and Analysis

**wgs ~110 Mb** **fos ~12 Mb**

Slide from Steven Hallam, UBC

# Groups of annotated organism files

## Methane oxidizing Archaea



>gi_15605614_ref_NP_212987.1_ elongation factor Tu [Aquifex aeolicus VF5]

MAKEKFERTKEHVNVGTIGHVDHGKSTLTSAITCVL AAGLVEGGKAKCFKYEEIDKAPEEKERGITINIT

>gi_15605615_ref_NP_212988.1_ ribosomal protein S10 [Aquifex aeolicus VF5]

MEQEKIRIKLRAYDHRLLDQSVKQIIETVKRTGGVVK GPIPLPTRKRKWCVLRSPHKFDQSREHFEIREF

SRILDIIRFTPQTIEALMEISLPAGVDVEVKMRG

>gi_15605616_ref_NP_212989.1_ ribosomal protein L03 [Aquifex aeolicus VF5]

MPLGLIGEKVGMTRVLLKDGTAIPVTVIKFPVNYVVQ VKSQNTKDGYNALQIGAYEAKEKHLTKPLIGHF

K>gi_11498795_ref_NP_070024.1_ replication factor C large subunit [Archaeoglobus fulgidus DSM 4304]

DQRSWRVIERIVGEGAFNETISDEGEFLSSRIGKLK LIILDEVDNIHKKEDVGGEAALIRLIKRKPAQPL

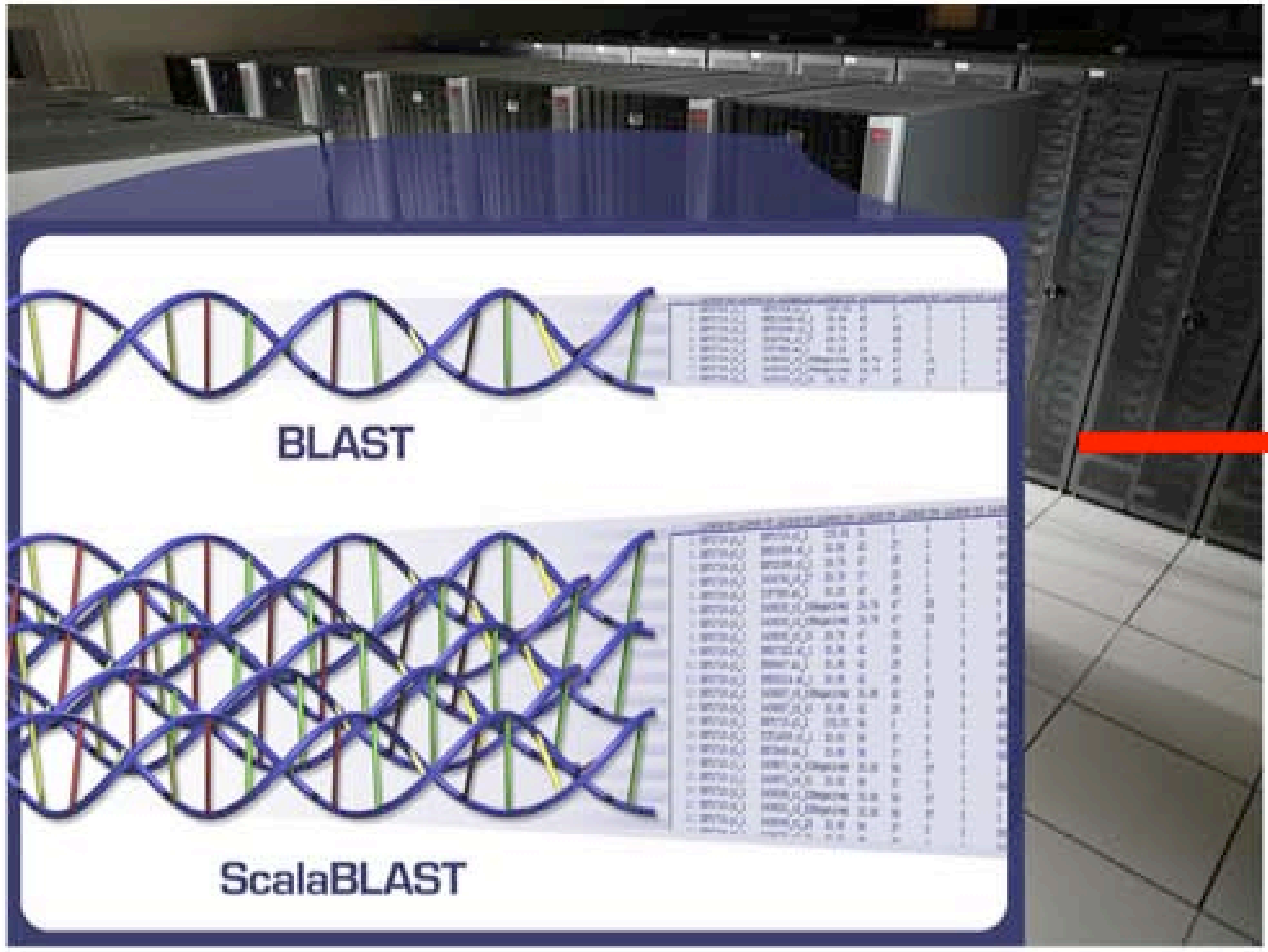
I>gi_220904928_ref_YP_002480240.1_ NADH (or F420H2) dehydrogenase, subunit C [Desulfovibrio desulfuricans subsp. desulfuricans str. ATCC 27774]

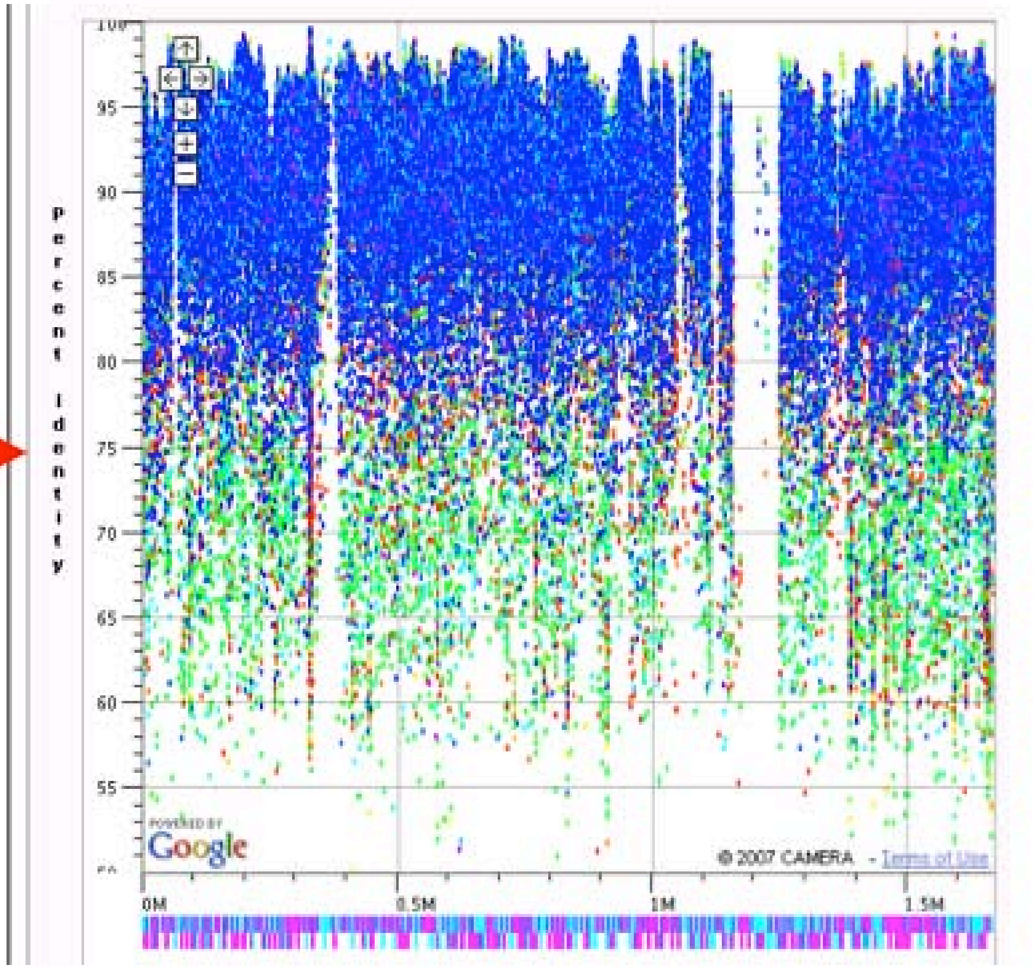
MESLEIADRLRGFFPEEVLDVREFRGQLAVLVRSGR ILELLAYLRDVLDMRHLQALCGVDNSRRNEPGLS

QGHPLRKEYPVKIPARGHEEWEGLTALRKRAAELD ALSWQGGARHE

Figure from Hinrichs, K. et al. (1999) *Nature* 398:802-805

# *In-silico* derived sequences: generating protein lookup tables from multiple organisms having similar function

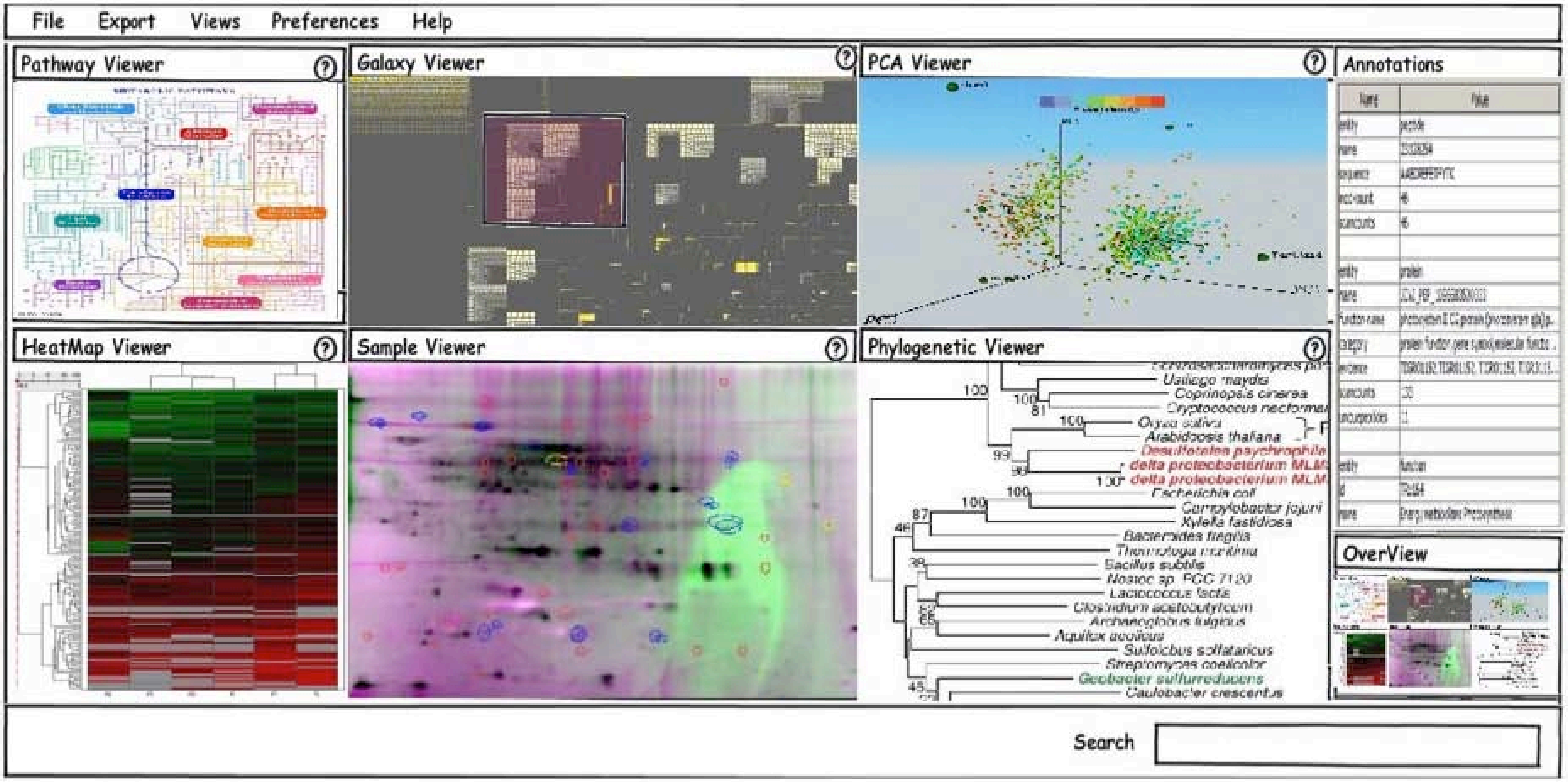




# Data analysis methods: software requirements are tough to meet

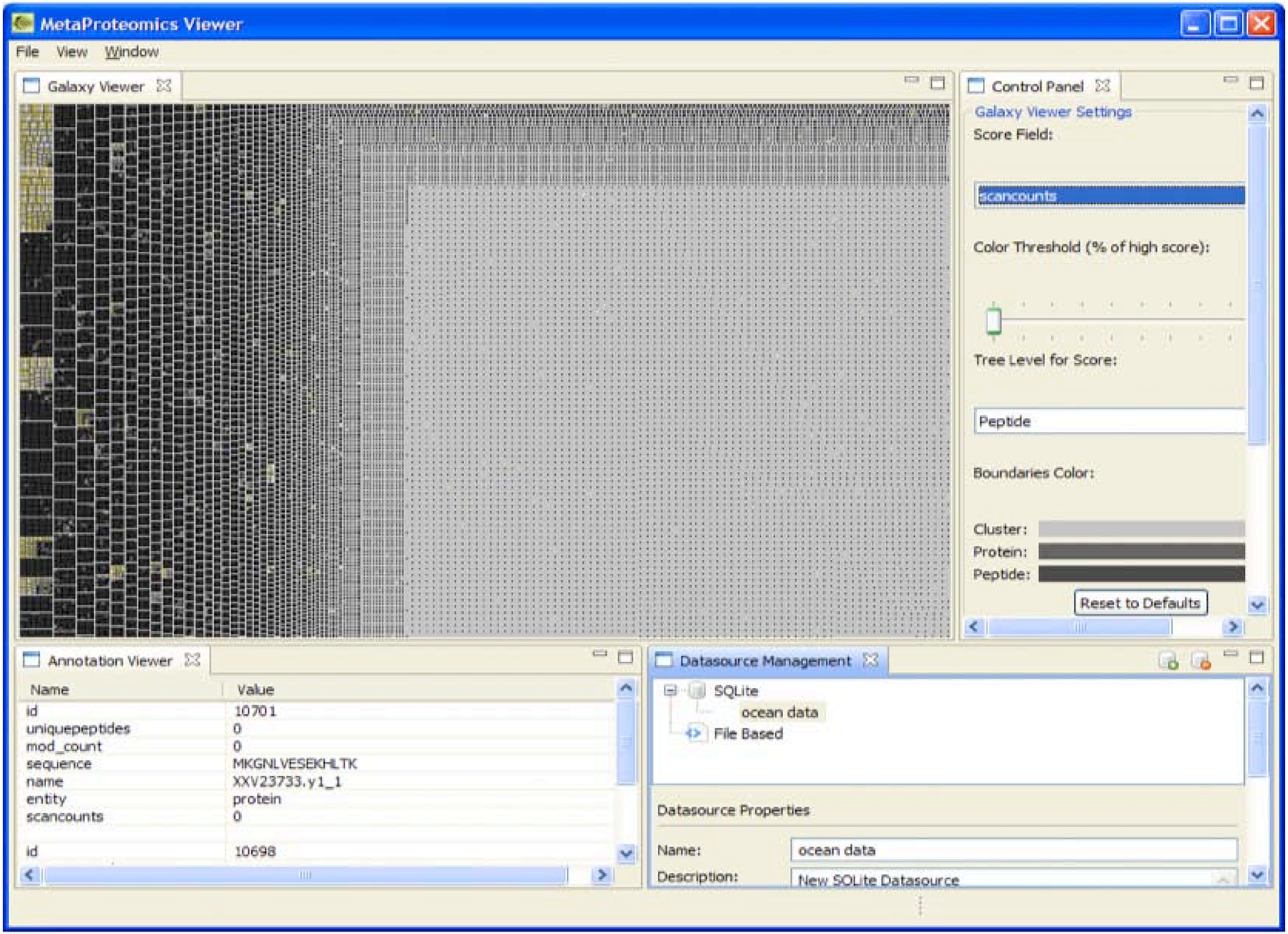
- **Globally view a community proteome using mass spectrometry and highlight identification regions**
  - Challenge: data files in GB with millions of rows
  - Data complexity: files are interrelated
  - No existing capability to detect patterns
  - Previous methods involved sorting lists and looking at top 1000 rows
- **Zoom in on regions and extract information about proteins**
  - Challenge: Redrawing of interface takes time and memory and information extraction about subset requires a new query
- **Change between different peptide and protein information**

# IMPROV – Integrated MetaPROteomics Viewer



| Name | Value |
|---|---|
| entity | peptide |
| name | [illegible] |
| sequence | [illegible] |
| spec count | 46 |
| spcounts | 46 |
| | |
| entity | protein |
| name | [illegible] |
| function name | photosystem II [illegible] protein ([illegible]) |
| category | protein function: gene type(s),molecular functio... |
| evidence | [illegible] |
| spcounts | [illegible] |
| uniquepeptides | [illegible] |
| | |
| entity | function |
| id | [illegible] |
| name | Energy metabolism: Photosynthesis |

# Galaxy View Navigation with Control Panel



Galaxy Viewer Settings

Score Field: scancounts

Color Threshold (% of high score):

Tree Level for Score: Peptide

Boundaries Color:

Cluster:

Protein:

Peptide:

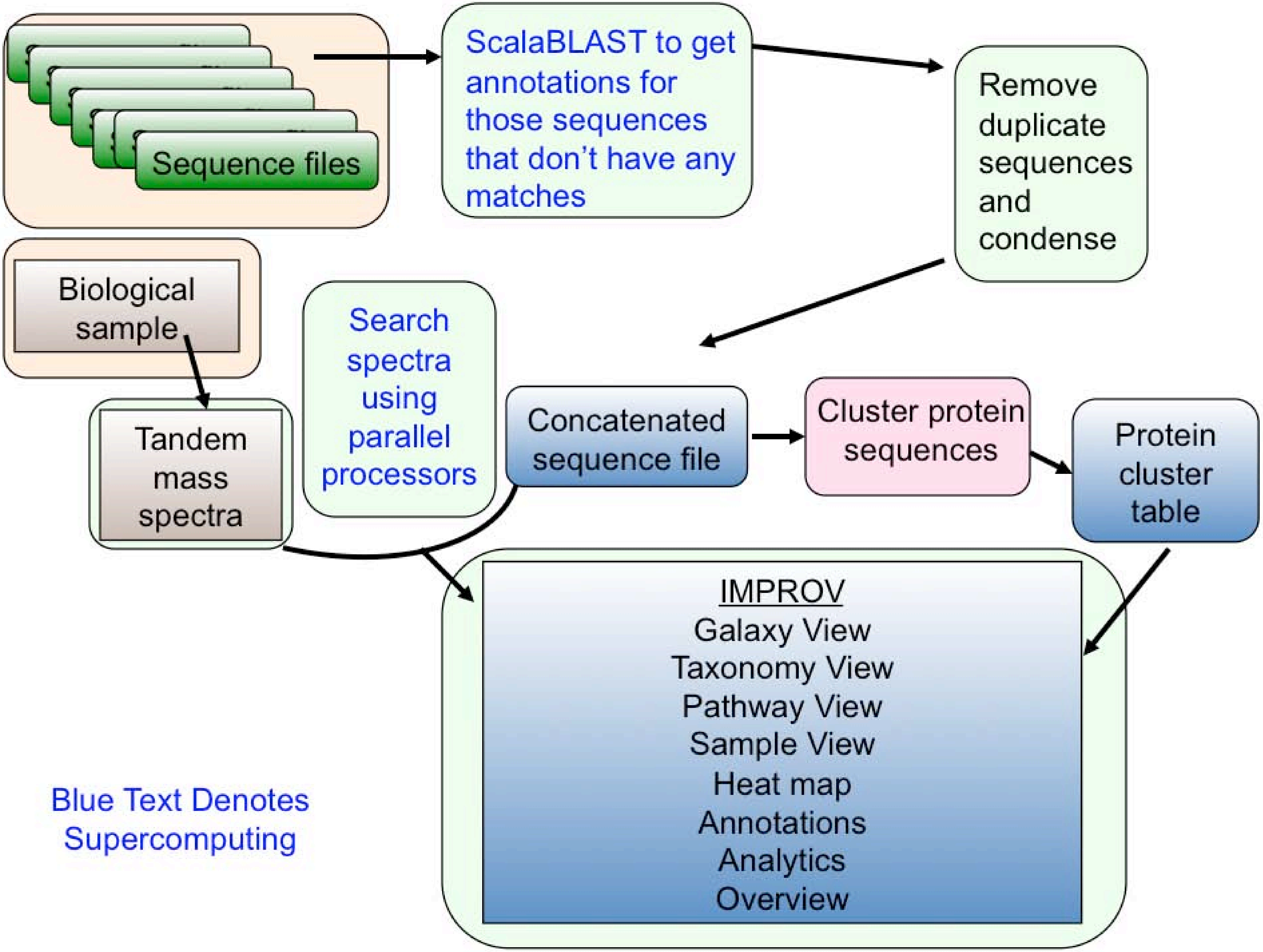
| Name | Value |
|---|---|
| id | 10701 |
| uniquepeptides | 0 |
| mod_count | 0 |
| sequence | MKGNLVESEKHLTK |
| name | XXV23733.y1_1 |
| entity | protein |
| scancounts | 0 |
| | |
| id | 10698 |

- SQLite
  - ocean data
- File Based

Datasource Properties

Name: ocean data

Description: New SQLite Datasource



Sequence files

ScalaBLAST to get annotations for those sequences that don't have any matches

Remove duplicate sequences and condense

Biological sample

Tandem mass spectra

Search spectra using parallel processors

Concatenated sequence file

Cluster protein sequences

Protein cluster table

IMPROV

Galaxy View

Taxonomy View

Pathway View

Sample View

Heat map

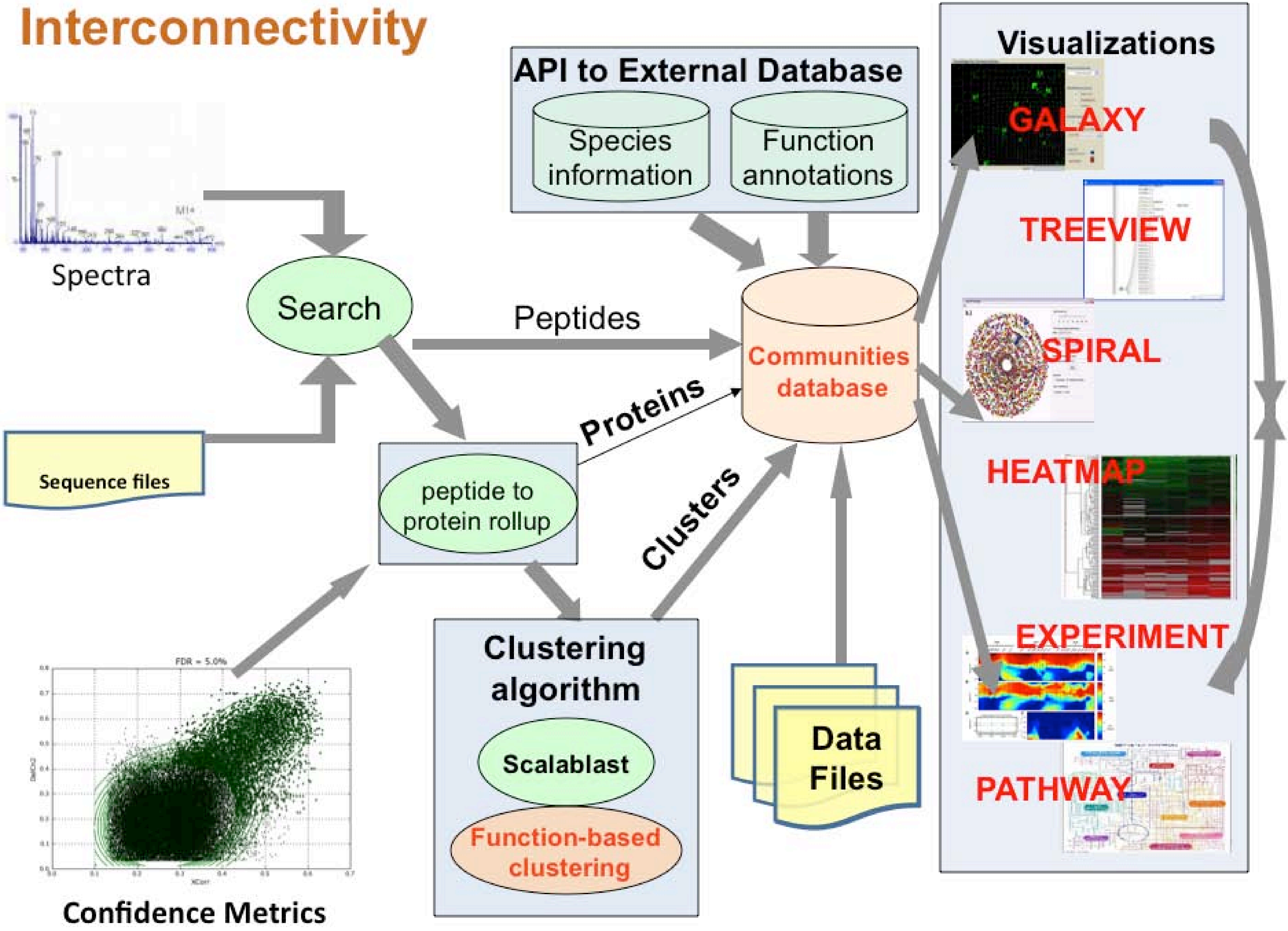
Annotations

Analytics

Overview

Blue Text Denotes Supercomputing

# Interconnectivity



Spectra

Search

Sequence files

Confidence Metrics

FDR = 5.0%

Peptides

Proteins

Clusters

peptide to protein rollup

API to External Database

Species information

Function annotations

Communities database

Clustering algorithm

Scalablast

Function-based clustering

Data Files

Visualizations

GALAXY

TREEVIEW

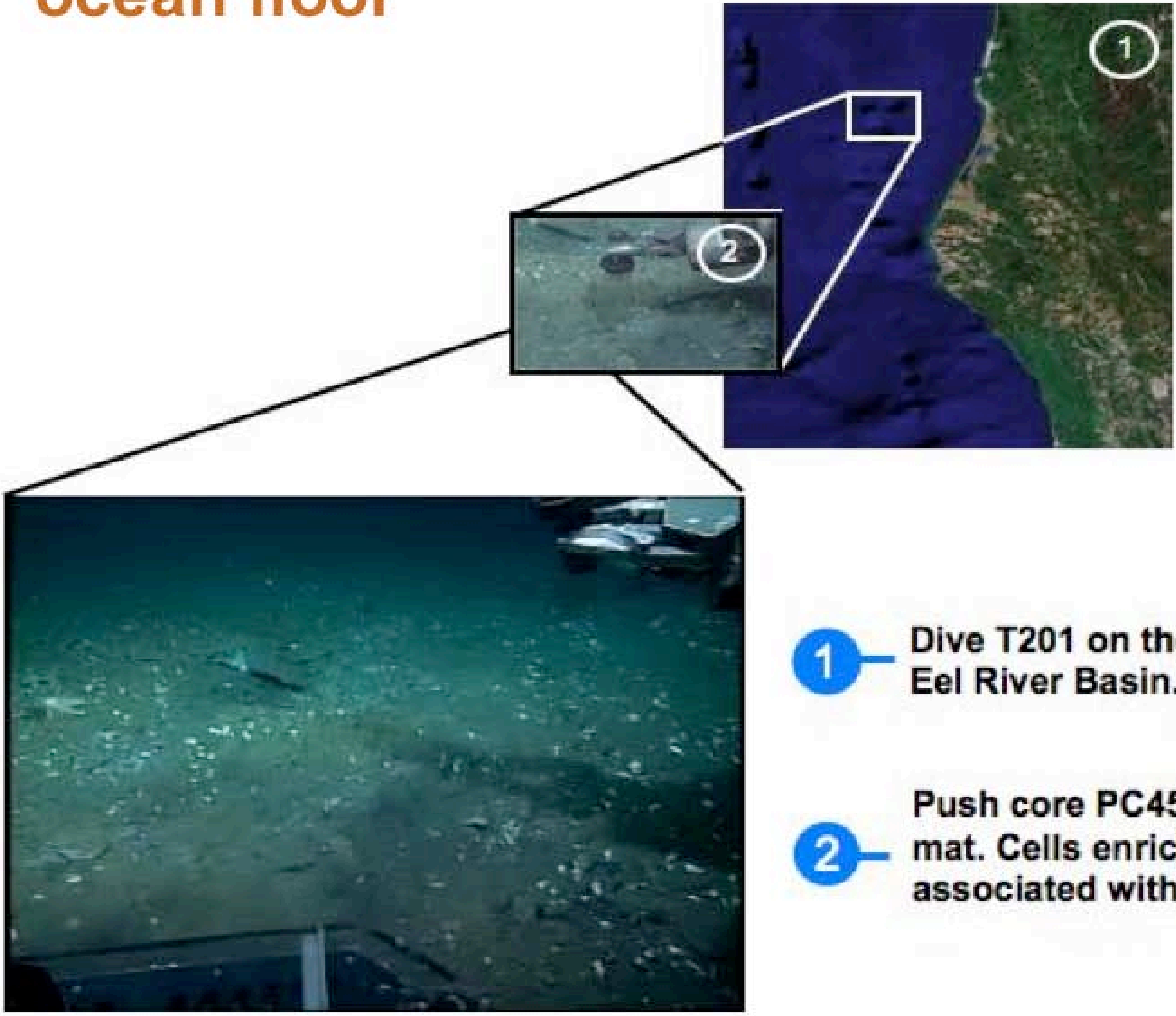
SPIRAL

HEATMAP

EXPERIMENT

PATHWAY

# Example: Proteomics exploration of the ocean floor

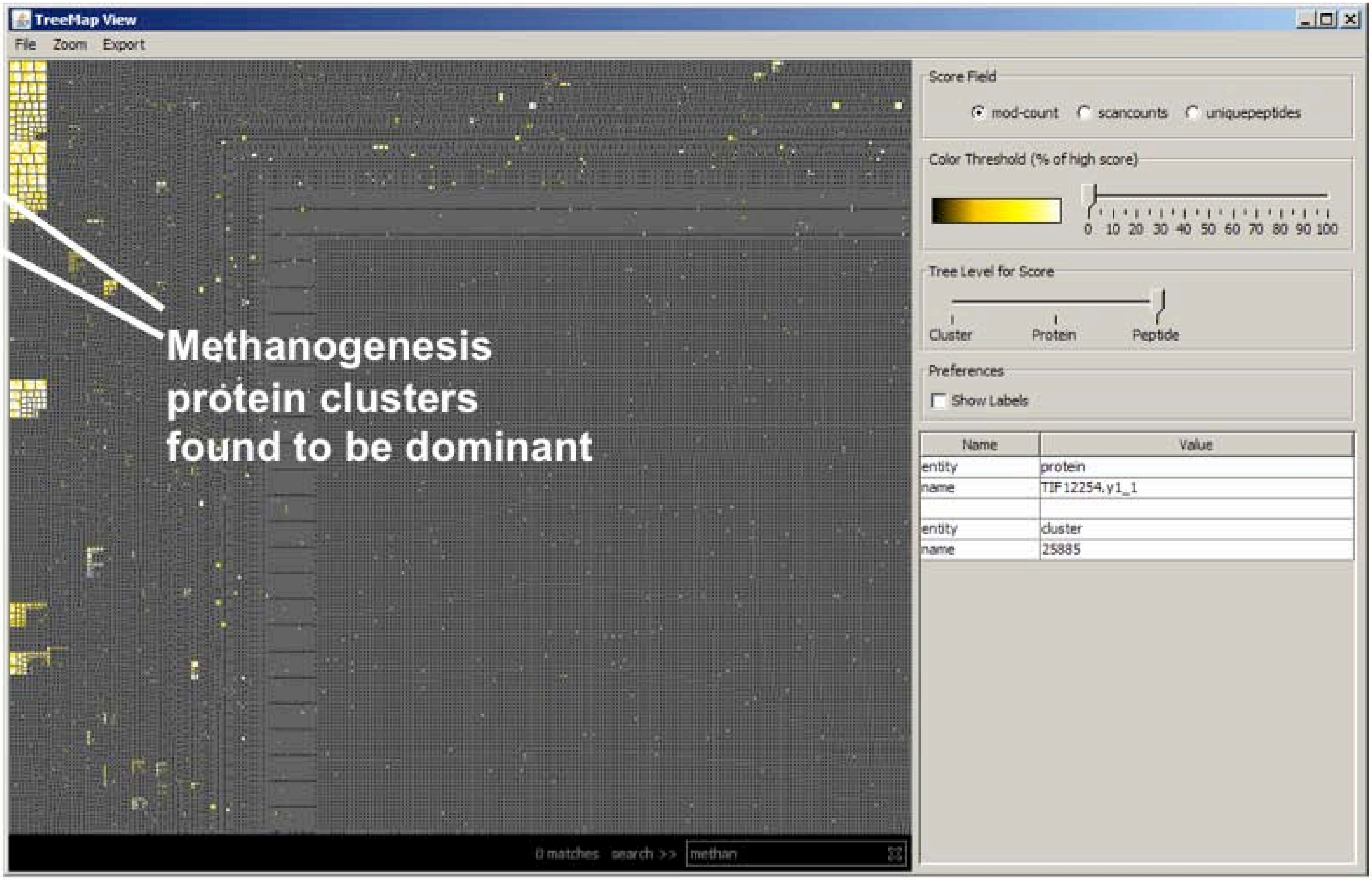


1 Dive T201 on the southern ridge of the Eel River Basin.

2 Push core PC45 sampled below bacterial mat. Cells enriched from 6-9 cm interval associated with methane oxidation.

Slide from Steven Hallam, UBC

# Application to Eel River Basin microbial community



Methanogenesis protein clusters found to be dominant

Score Field

● mod-count ○ scancounts ○ uniquepeptides

Color Threshold (% of high score)

Tree Level for Score: Cluster, Protein, Peptide (selected: Peptide)

Preferences

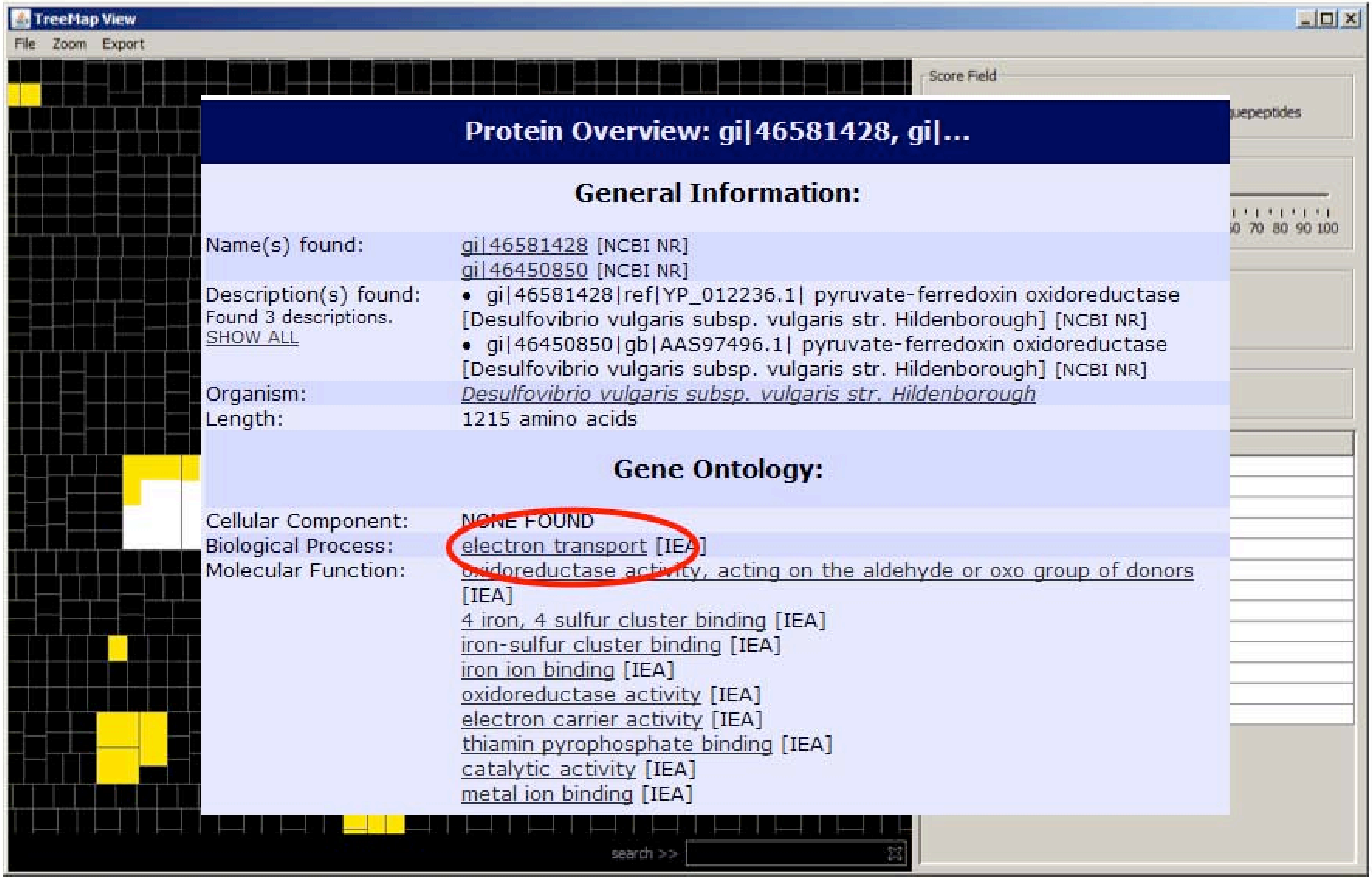
□ Show Labels

| Name | Value |
|---|---|
| entity | protein |
| name | TIF12254.y1_1 |
| | |
| entity | cluster |
| name | 25885 |

search >> methan

**172,000 proteins in 1700 clusters**

# New cellular functions discovered



## Protein Overview: gi|46581428, gi|...

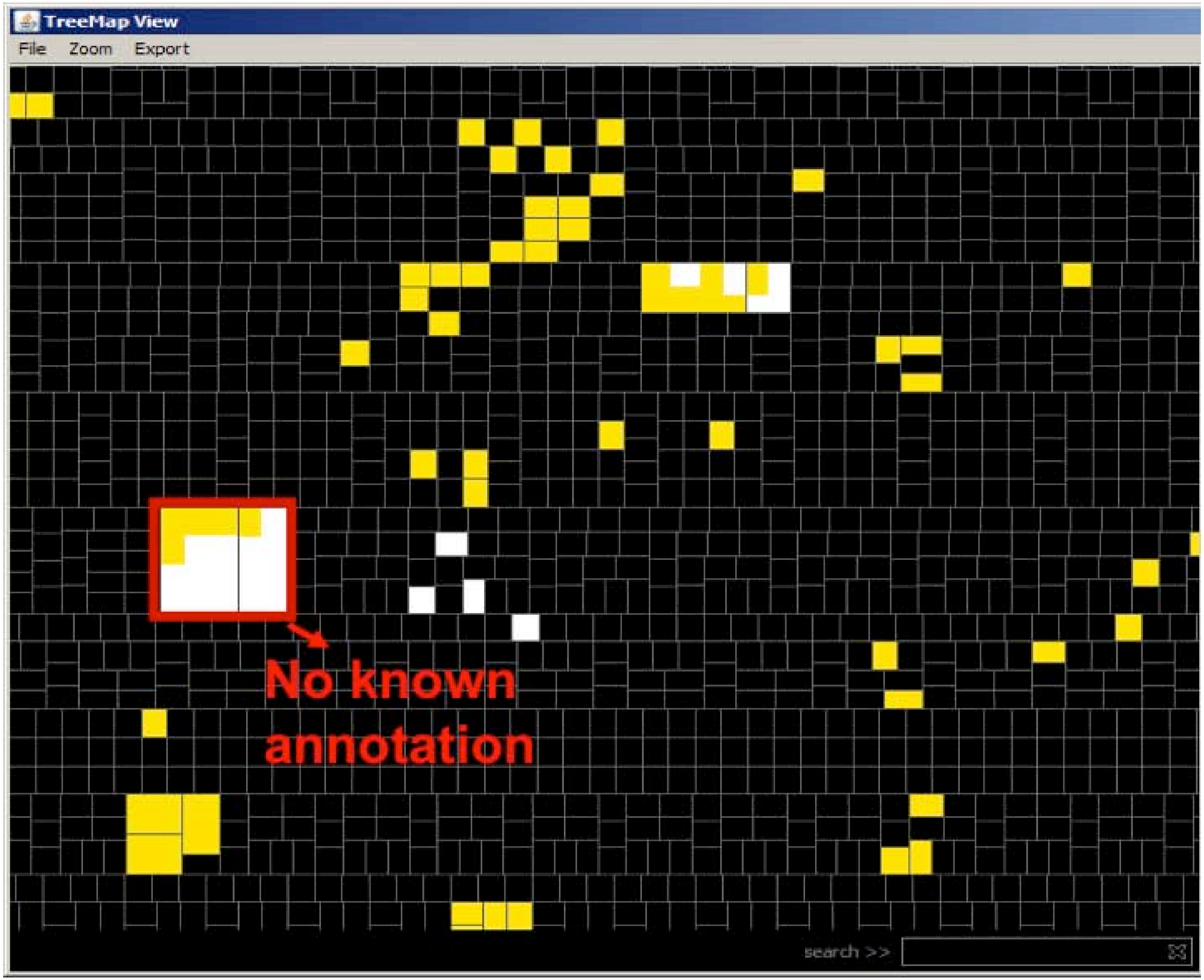
### General Information:

| | |
|---|---|
| Name(s) found: | gi\|46581428 [NCBI NR]<br>gi\|46450850 [NCBI NR] |
| Description(s) found:<br>Found 3 descriptions.<br>SHOW ALL | • gi\|46581428\|ref\|YP_012236.1\| pyruvate-ferredoxin oxidoreductase [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough] [NCBI NR]<br>• gi\|46450850\|gb\|AAS97496.1\| pyruvate-ferredoxin oxidoreductase [Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough] [NCBI NR] |
| Organism: | *Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough* |
| Length: | 1215 amino acids |

### Gene Ontology:

| | |
|---|---|
| Cellular Component: | NONE FOUND |
| Biological Process: | electron transport [IEA] |
| Molecular Function: | oxidoreductase activity, acting on the aldehyde or oxo group of donors [IEA]<br>4 iron, 4 sulfur cluster binding [IEA]<br>iron-sulfur cluster binding [IEA]<br>iron ion binding [IEA]<br>oxidoreductase activity [IEA]<br>electron carrier activity [IEA]<br>thiamin pyrophosphate binding [IEA]<br>catalytic activity [IEA]<br>metal ion binding [IEA] |

# Unknown functions explored

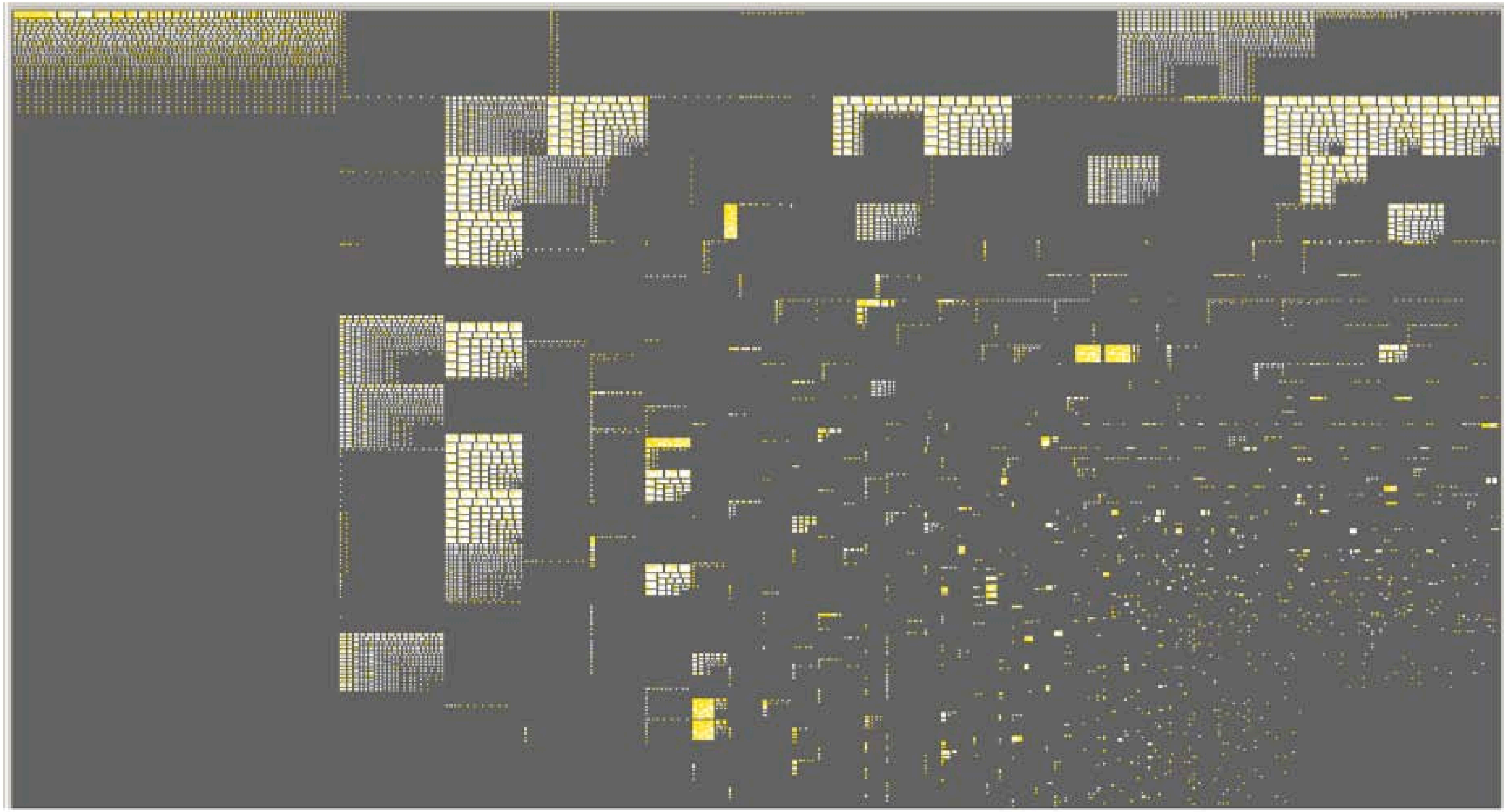


No known annotation

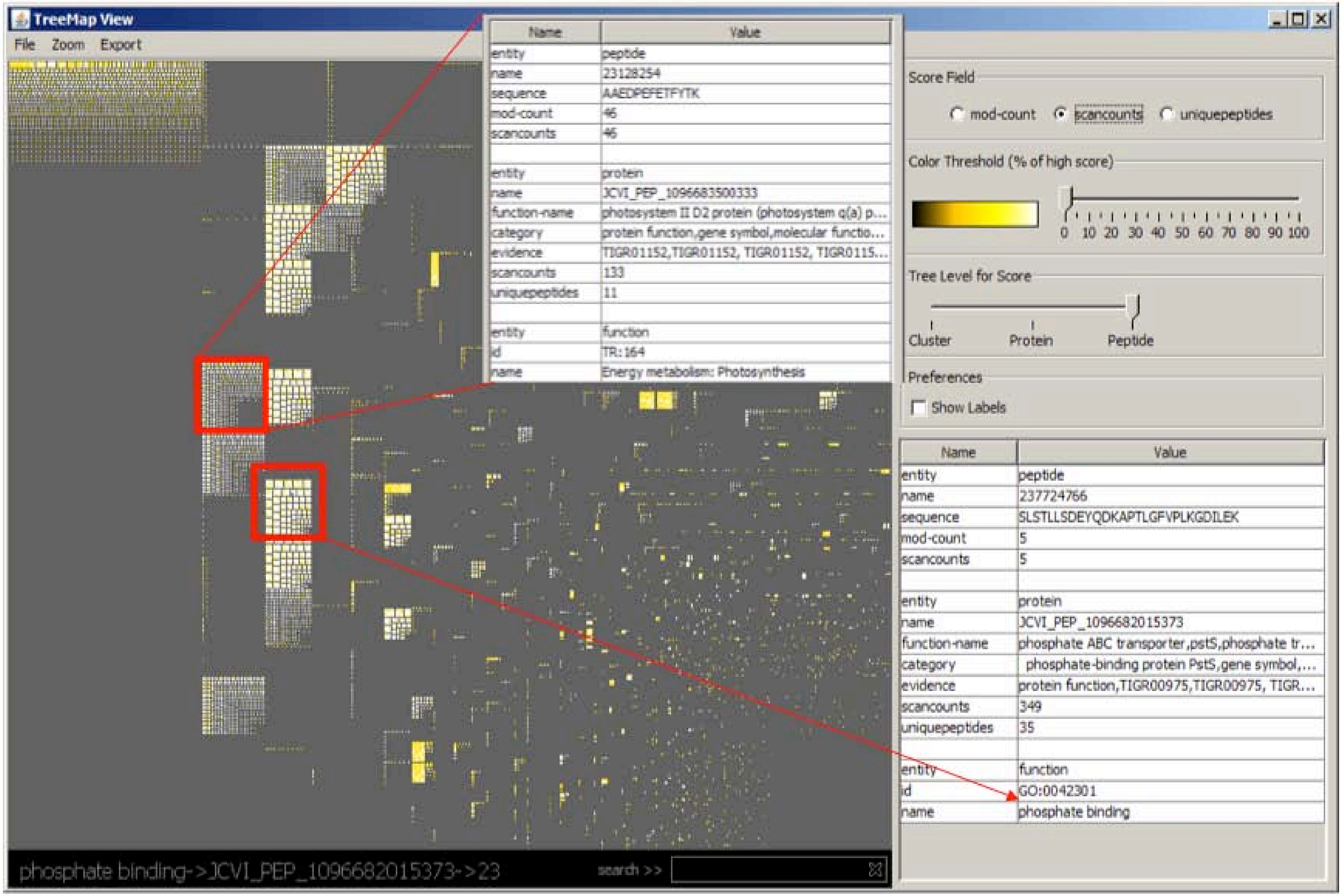
# Example: Proteomics exploration of a Sargasso Sea microbial community

- **2008 Global Ocean Survey protein file downloaded from CAMERA**
- **Over 6 x $10^6$ protein entries (2.2 GB file)**
- **Took longer than 1 month to search spectra**
- **Identified 16,000 clusters of proteins**
- **Peptide database is over 4 x “normal” (single organism) size**
  - Log file for database maintenance is an order of magnitude larger than usual

# Some proteins highly expressed, others not



# Shows annotations



| Name | Value |
|---|---|
| entity | peptide |
| name | 23128254 |
| sequence | AAEDPEFETFYTK |
| mod-count | 46 |
| scancounts | 46 |
| | |
| entity | protein |
| name | JCVI_PEP_1096683500333 |
| function-name | photosystem II D2 protein (photosystem q(a) p... |
| category | protein function,gene symbol,molecular functio... |
| evidence | TIGR01152,TIGR01152, TIGR01152, TIGR0115... |
| scancounts | 133 |
| uniquepeptides | 11 |
| | |
| entity | function |
| id | TR:164 |
| name | Energy metabolism: Photosynthesis |

Score Field

○ mod-count ● scancounts ○ uniquepeptides

Color Threshold (% of high score)

Tree Level for Score

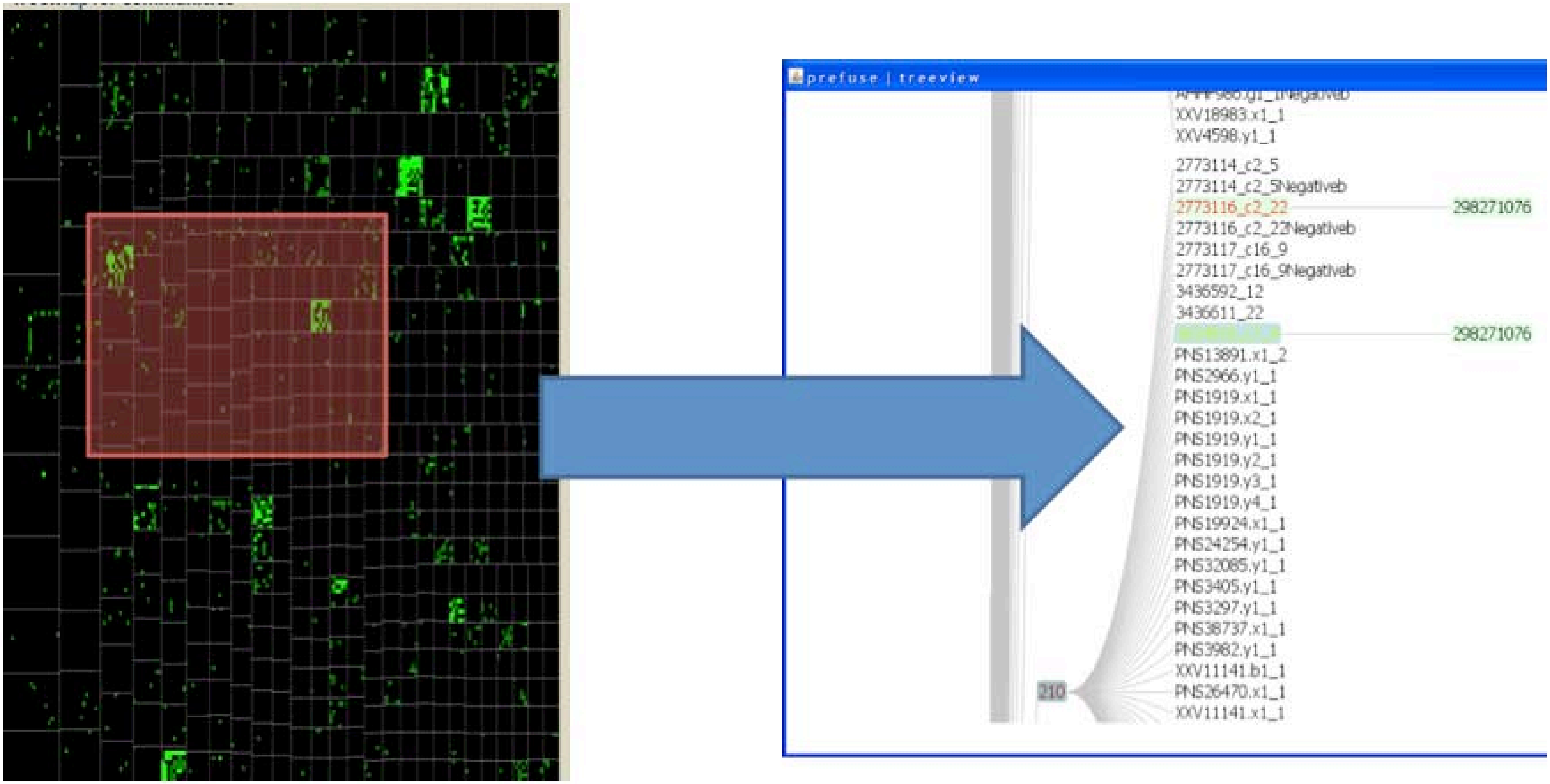
Preferences

□ Show Labels

| Name | Value |
|---|---|
| entity | peptide |
| name | 237724766 |
| sequence | SLSTLLSDEYQDKAPTLGFVPLKGDILEK |
| mod-count | 5 |
| scancounts | 5 |
| | |
| entity | protein |
| name | JCVI_PEP_1096682015373 |
| function-name | phosphate ABC transporter,pstS,phosphate tr... |
| category | phosphate-binding protein PstS,gene symbol,... |
| evidence | protein function,TIGR00975,TIGR00975, TIGR... |
| scancounts | 349 |
| uniquepeptides | 35 |
| | |
| entity | function |
| id | GO:0042301 |
| name | phosphate binding |

phosphate binding->JCVI_PEP_1096682015373->23

# Data mining of smaller, less abundantly expressed proteins is critical to understanding dynamics



XXV18983.x1_1
XXV4598.y1_1
2773114_c2_5
2773114_c2_5Negativeb
2773116_c2_22 298271076
2773116_c2_22Negativeb
2773117_c16_9
2773117_c16_9Negativeb
3436592_12
3436611_22
[illegible] 298271076
PNS13891.x1_2
PNS2966.y1_1
PNS1919.x1_1
PNS1919.x2_1
PNS1919.y1_1
PNS1919.y2_1
PNS1919.y3_1
PNS1919.y4_1
PNS19924.x1_1
PNS24254.y1_1
PNS32085.y1_1
PNS3405.y1_1
PNS3297.y1_1
PNS38737.x1_1
PNS3982.y1_1
XXV11141.b1_1
PNS26470.x1_1
XXV11141.x1_1

# Additional IMPROV features

- **Multiple ways to cluster proteins (i.e. by function, cell location, pathway..)**
- **Ability to toggle between peptide/protein spectra counts and peptide/protein abundance values or modification states**
- **Text string searching**
- **Zooming functions**
- **Ability to view and analyze multiple conditions and find clusters changing dynamically**
  - Spiral graphs for time course
- **Normalization of clusters by percent coverage**
  - Reorganization of map by highest coverage
- **Link out to other databases (Kegg, pfam..)**

# Acknowledgments

### Software Development

- Mudita Singhal
- Kelly Domico
- Getiria Onsongo

### Collaborators

- Dr. Steven Hallam
  University of British Columbia
- Dr. Stephen Giovannoni
  Oregon State University

### Informatics Team at PNNL

### High-Throughput Proteomics Production Team at PNNL

### Funding and Facilities

# Discussion Points

- What are the main kinds of biological applications for environmental proteomics?
- Should data repositories adhere to standard formats?
- How much data can or should be shared?
- How should orthologs amongst different organisms be used to relate functions in a community?