Viewing and Interpreting Data within a Biological Context

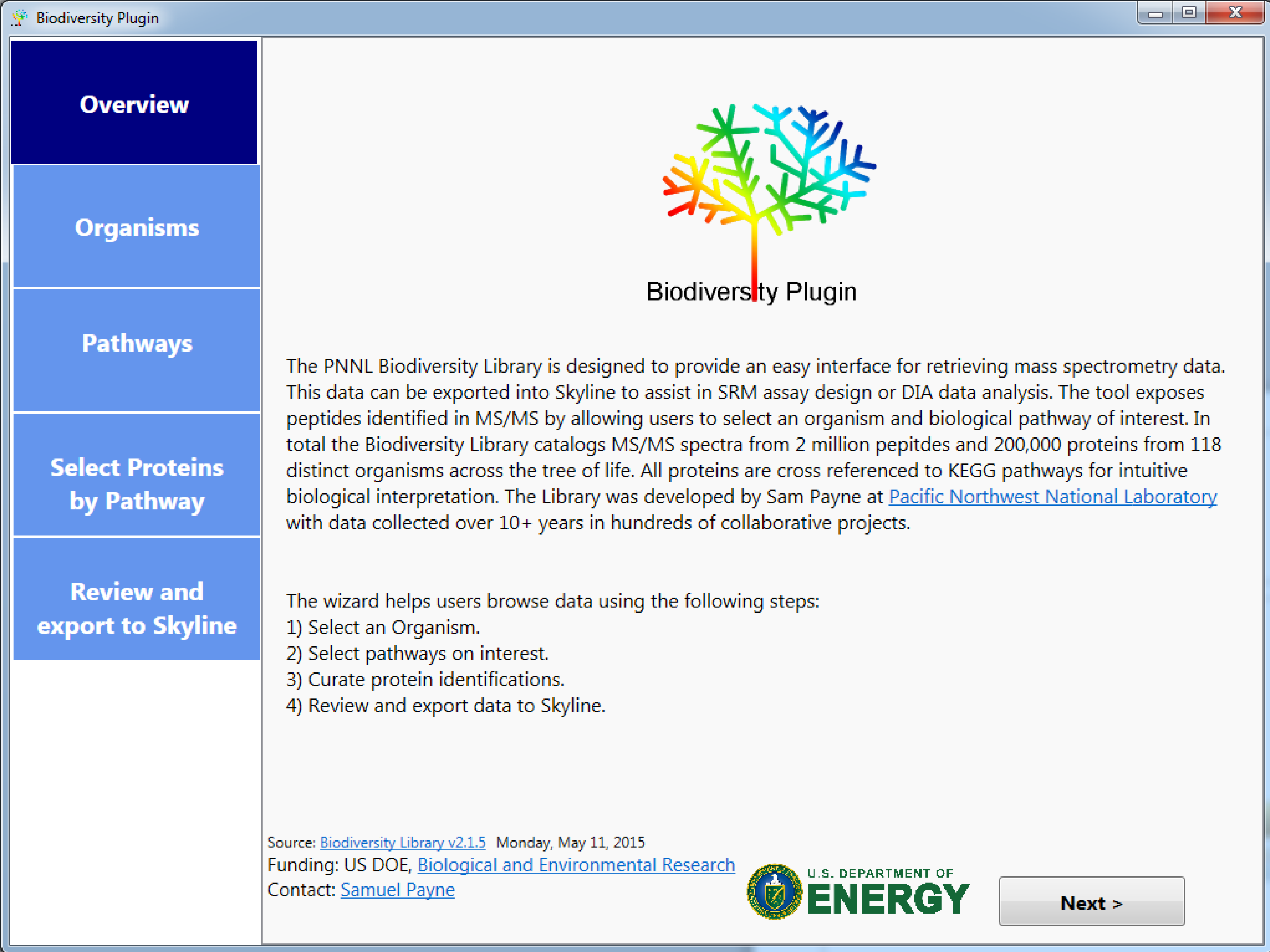
Samuel Payne

Pacific Northwest National Laboratory

Goals

- Biodiversity Library and Plugin – Expose public data
  - Easy download/consumption of data on proteomeXchange
  - Planning experiments from a biological perspective

- Active Data Canvas – Provide visual data analysis
  - Getting away from spreadsheets
  - Collaborating on data analysis

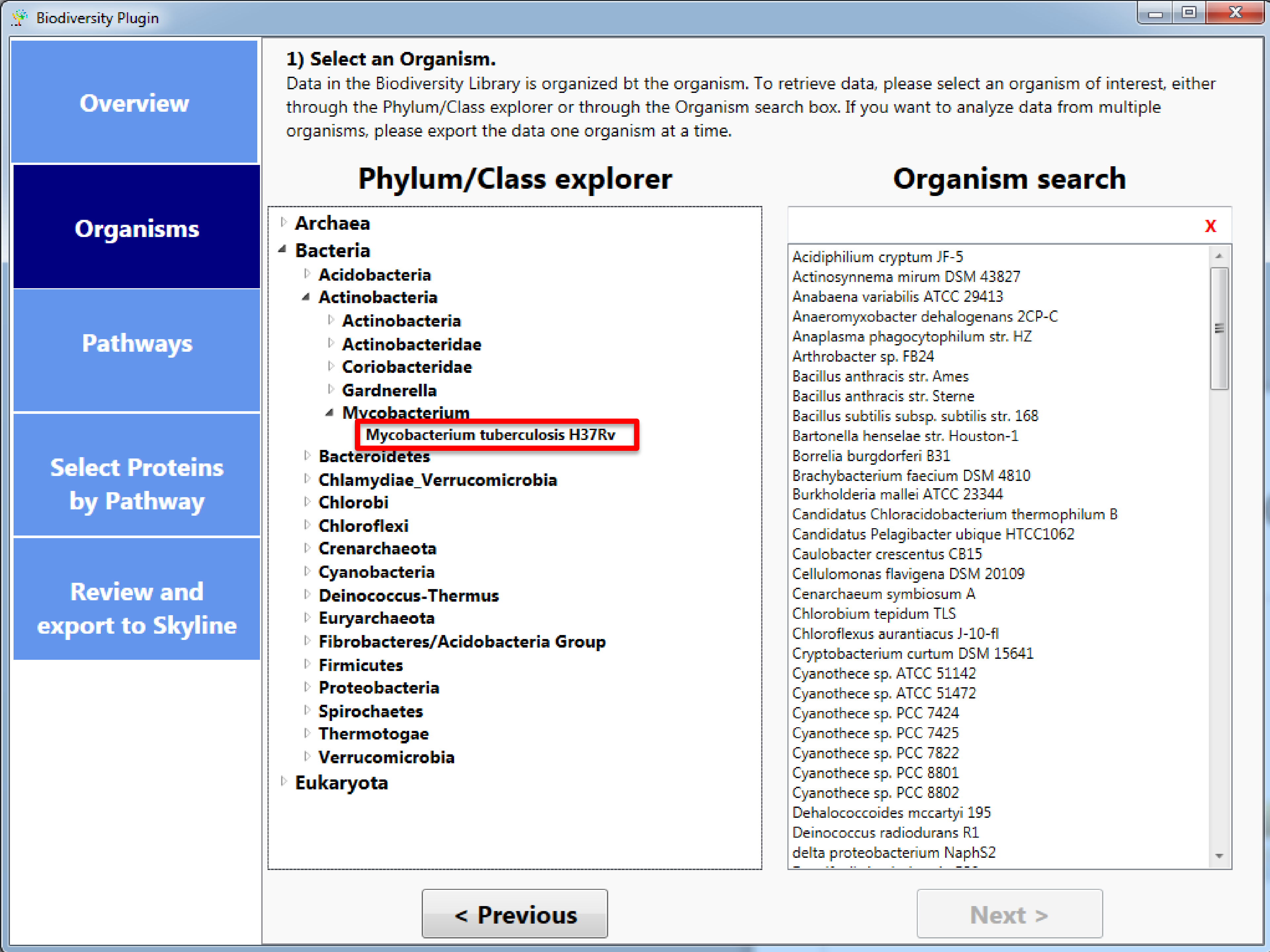


The PNNL Biodiversity Library is designed to provide an easy interface for retrieving mass spectrometry data. This data can be exported into Skyline to assist in SRM assay design or DIA data analysis. The tool exposes peptides identified in MS/MS by allowing users to select an organism and biological pathway of interest. In total the Biodiversity Library catalogs MS/MS spectra from 2 million pepitdes and 200,000 proteins from 118 distinct organisms across the tree of life. All proteins are cross referenced to KEGG pathways for intuitive biological interpretation. The Library was developed by Sam Payne at Pacific Northwest National Laboratory with data collected over 10+ years in hundreds of collaborative projects.

The wizard helps users browse data using the following steps:

1) Select an Organism.
2) Select pathways on interest.
3) Curate protein identifications.
4) Review and export data to Skyline.

Source: Biodiversity Library v2.1.5 Monday, May 11, 2015
Funding: US DOE, Biological and Environmental Research
Contact: Samuel Payne



1) Select an Organism.

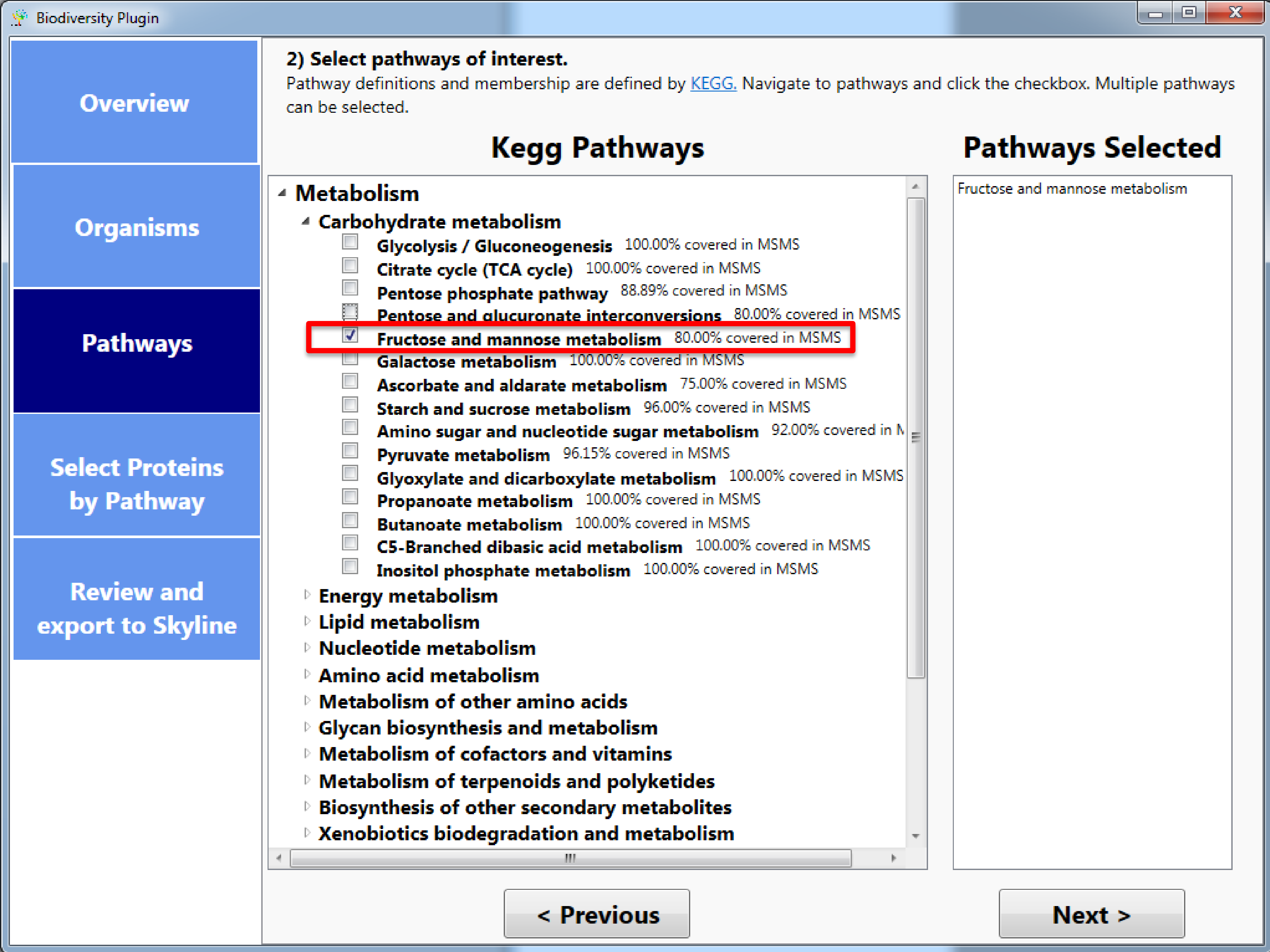
Data in the Biodiversity Library is organized bt the organism. To retrieve data, please select an organism of interest, either through the Phylum/Class explorer or through the Organism search box. If you want to analyze data from multiple organisms, please export the data one organism at a time.

Phylum/Class explorer

- Archaea
- Bacteria
  - Acidobacteria
  - Actinobacteria
    - Actinobacteria
    - Actinobacteridae
    - Coriobacteridae
    - Gardnerella
    - Mycobacterium
      - Mycobacterium tuberculosis H37Rv
  - Bacteroidetes
  - Chlamydiae_Verrucomicrobia
  - Chlorobi
  - Chloroflexi
  - Crenarchaeota
  - Cyanobacteria
  - Deinococcus-Thermus
  - Euryarchaeota
  - Fibrobacteres/Acidobacteria Group
  - Firmicutes
  - Proteobacteria
  - Spirochaetes
  - Thermotogae
  - Verrucomicrobia
- Eukaryota

Organism search

Acidiphilium cryptum JF-5
Actinosynnema mirum DSM 43827
Anabaena variabilis ATCC 29413
Anaeromyxobacter dehalogenans 2CP-C
Anaplasma phagocytophilum str. HZ
Arthrobacter sp. FB24
Bacillus anthracis str. Ames
Bacillus anthracis str. Sterne
Bacillus subtilis subsp. subtilis str. 168
Bartonella henselae str. Houston-1
Borrelia burgdorferi B31
Brachybacterium faecium DSM 4810
Burkholderia mallei ATCC 23344
Candidatus Chloracidobacterium thermophilum B
Candidatus Pelagibacter ubique HTCC1062
Caulobacter crescentus CB15
Cellulomonas flavigena DSM 20109
Cenarchaeum symbiosum A
Chlorobium tepidum TLS
Chloroflexus aurantiacus J-10-fl
Cryptobacterium curtum DSM 15641
Cyanothece sp. ATCC 51142
Cyanothece sp. ATCC 51472
Cyanothece sp. PCC 7424
Cyanothece sp. PCC 7425
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 8801
Cyanothece sp. PCC 8802
Dehalococcoides mccartyi 195
Deinococcus radiodurans R1
delta proteobacterium NaphS2



2) Select pathways of interest.

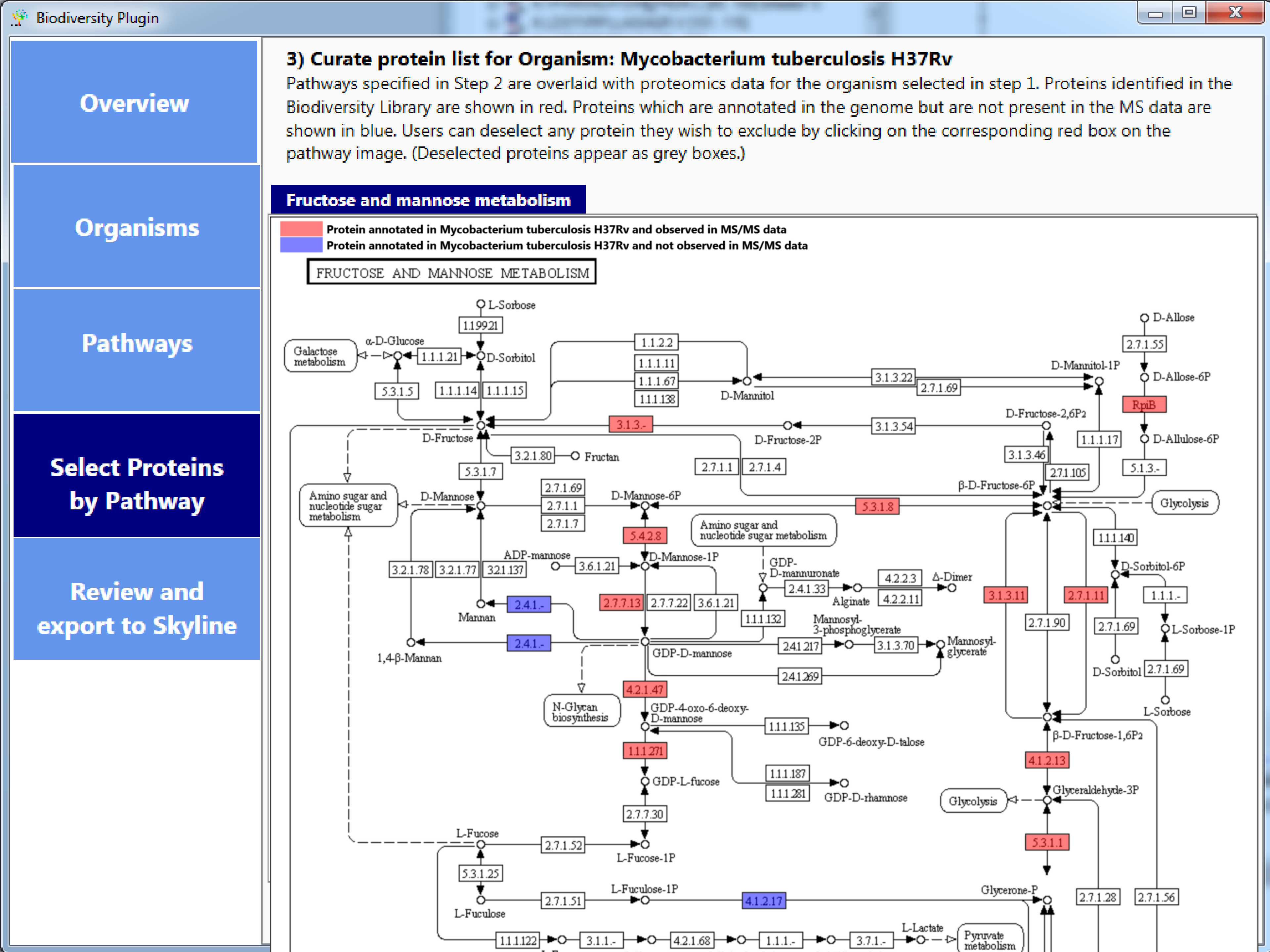
Pathway definitions and membership are defined by KEGG. Navigate to pathways and click the checkbox. Multiple pathways can be selected.

Kegg Pathways

- **Metabolism**
  - **Carbohydrate metabolism**
    - ☐ **Glycolysis / Gluconeogenesis** 100.00% covered in MSMS
    - ☐ **Citrate cycle (TCA cycle)** 100.00% covered in MSMS
    - ☐ **Pentose phosphate pathway** 88.89% covered in MSMS
    - ☐ **Pentose and glucuronate interconversions** 80.00% covered in MSMS
    - ☑ **Fructose and mannose metabolism** 80.00% covered in MSMS
    - ☐ **Galactose metabolism** 100.00% covered in MSMS
    - ☐ **Ascorbate and aldarate metabolism** 75.00% covered in MSMS
    - ☐ **Starch and sucrose metabolism** 96.00% covered in MSMS
    - ☐ **Amino sugar and nucleotide sugar metabolism** 92.00% covered in
    - ☐ **Pyruvate metabolism** 96.15% covered in MSMS
    - ☐ **Glyoxylate and dicarboxylate metabolism** 100.00% covered in MSMS
    - ☐ **Propanoate metabolism** 100.00% covered in MSMS
    - ☐ **Butanoate metabolism** 100.00% covered in MSMS
    - ☐ **C5-Branched dibasic acid metabolism** 100.00% covered in MSMS
    - ☐ **Inositol phosphate metabolism** 100.00% covered in MSMS
  - **Energy metabolism**
  - **Lipid metabolism**
  - **Nucleotide metabolism**
  - **Amino acid metabolism**
  - **Metabolism of other amino acids**
  - **Glycan biosynthesis and metabolism**
  - **Metabolism of cofactors and vitamins**
  - **Metabolism of terpenoids and polyketides**
  - **Biosynthesis of other secondary metabolites**
  - **Xenobiotics biodegradation and metabolism**

Pathways Selected

Fructose and mannose metabolism



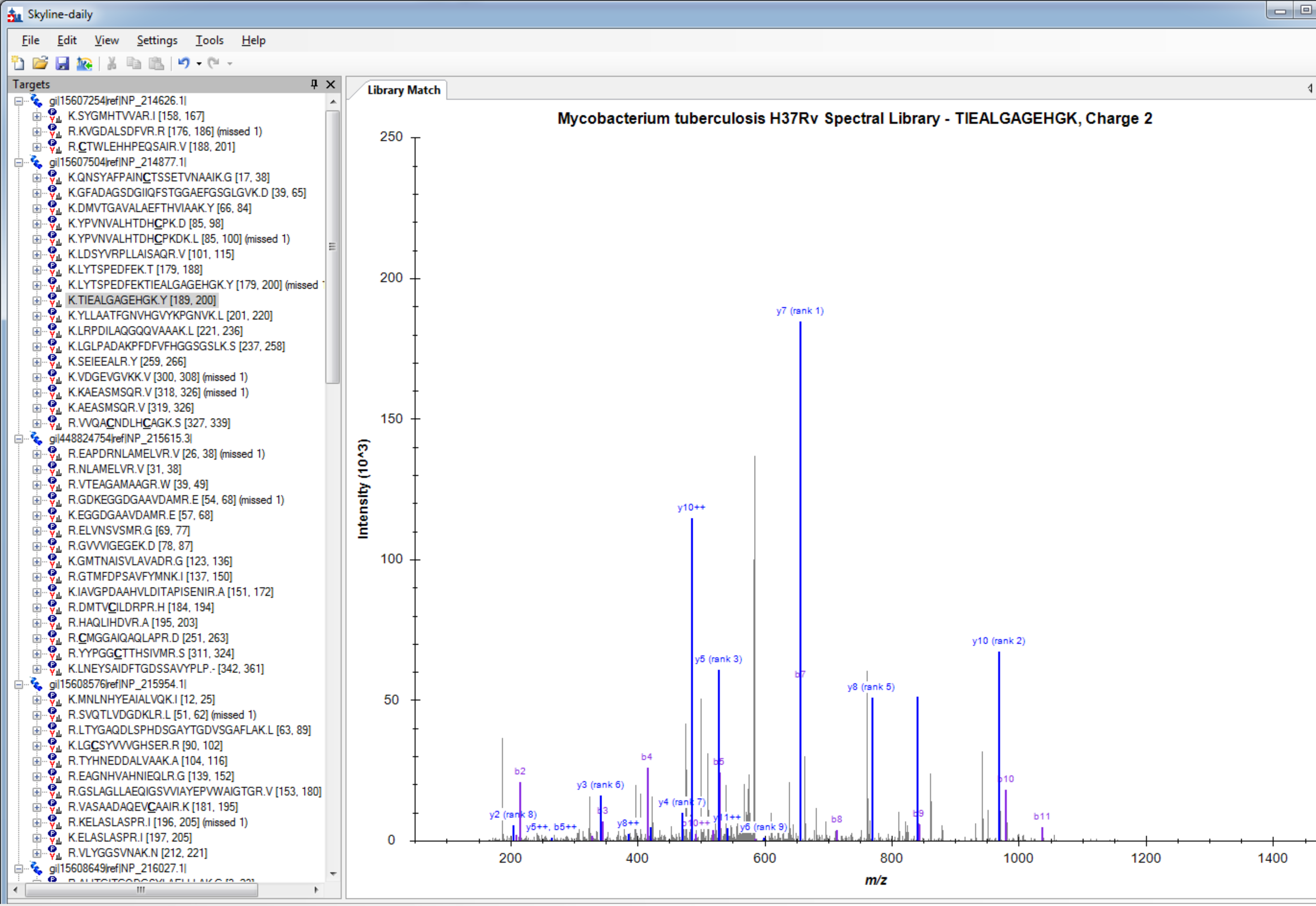
3) Curate protein list for Organism: Mycobacterium tuberculosis H37Rv

Pathways specified in Step 2 are overlaid with proteomics data for the organism selected in step 1. Proteins identified in the Biodiversity Library are shown in red. Proteins which are annotated in the genome but are not present in the MS data are shown in blue. Users can deselect any protein they wish to exclude by clicking on the corresponding red box on the pathway image. (Deselected proteins appear as grey boxes.)

Fructose and mannose metabolism

🟥 Protein annotated in Mycobacterium tuberculosis H37Rv and observed in MS/MS data

🟦 Protein annotated in Mycobacterium tuberculosis H37Rv and not observed in MS/MS data



Targets

- gi|15607254|ref|NP_214626.1|
  - K.SYGMHTVVAR.I [158, 167]
  - R.KVGDALSDFVR.R [176, 186] (missed 1)
  - R.CTWLEHHPEQSAIR.V [188, 201]
- gi|15607504|ref|NP_214877.1|
  - K.QNSYAFPAINCTSSETVNAAIK.G [17, 38]
  - K.GFADAGSDGIIQFSTGGAEFGSGLGVK.D [39, 65]
  - K.DMVTGAVALAEFTHVIAAK.Y [66, 84]
  - K.YPVNVALHTDHCPK.D [85, 98]
  - K.YPVNVALHTDHCPKDK.L [85, 100] (missed 1)
  - K.LDSYVRPLLAISAQR.V [101, 115]
  - K.LYTSPEDFEK.T [179, 188]
  - K.LYTSPEDFEKTIEALGAGEHGK.Y [179, 200] (missed
  - K.TIEALGAGEHGK.Y [189, 200]
  - K.YLLAATFGNVHGVYKPGNVK.L [201, 220]
  - K.LRPDILAQGQQVAAAK.L [221, 236]
  - K.LGLPADAKPFDFVFHGGSGSLK.S [237, 258]
  - K.SEIEEALR.Y [259, 266]
  - K.VDGEVGVKK.V [300, 308] (missed 1)
  - K.KAEASMSQR.V [318, 326] (missed 1)
  - K.AEASMSQR.V [319, 326]
  - R.VVQACNDLHCAGK.S [327, 339]
- gi|448824754|ref|NP_215615.3|
  - R.EAPDRNLAMELVR.V [26, 38] (missed 1)
  - R.NLAMELVR.V [31, 38]
  - R.VTEAGAMAAGR.W [39, 49]
  - R.GDKEGGDGAAVDAMR.E [54, 68] (missed 1)
  - K.EGGDGAAVDAMR.E [57, 68]
  - R.ELVNSVSMR.G [69, 77]
  - R.GVVVIGEGEK.D [78, 87]
  - K.GMTNAISVLAVADR.G [123, 136]
  - R.GTMFDPSAVFYMNK.I [137, 150]
  - K.IAVGPDAAHVLDITAPISENIR.A [151, 172]
  - R.DMTVCILDRPR.H [184, 194]
  - R.HAQLIHDVR.A [195, 203]
  - R.CMGGAIQAQLAPR.D [251, 263]
  - R.YYPGGCTTHSIVMR.S [311, 324]
  - K.LNEYSAIDFTGDSSAVYPLP.- [342, 361]
- gi|15608576|ref|NP_215954.1|
  - K.MNLNHYEAIALVQK.I [12, 25]
  - R.SVQTLVDGDKLR.L [51, 62] (missed 1)
  - R.LTYGAQDLSPHDSGAYTGDVSGAFLAK.L [63, 89]
  - K.LGCSYVVVGHSER.R [90, 102]
  - R.TYHNEDDALVAAK.A [104, 116]
  - R.EAGNHVAHNIEQLR.G [139, 152]
  - R.GSLAGLLAEQIGSVVIAYEPVWAIGTGR.V [153, 180]
  - R.VASAADAQEVCAAIR.K [181, 195]
  - R.KELASLASPR.I [196, 205] (missed 1)
  - K.ELASLASPR.I [197, 205]
  - R.VLYGGSVNAK.N [212, 221]
- gi|15608649|ref|NP_216027.1|

Library/Plugin Uses

- Designing SRM assays
- Spectrum library searches
- DIA scoring

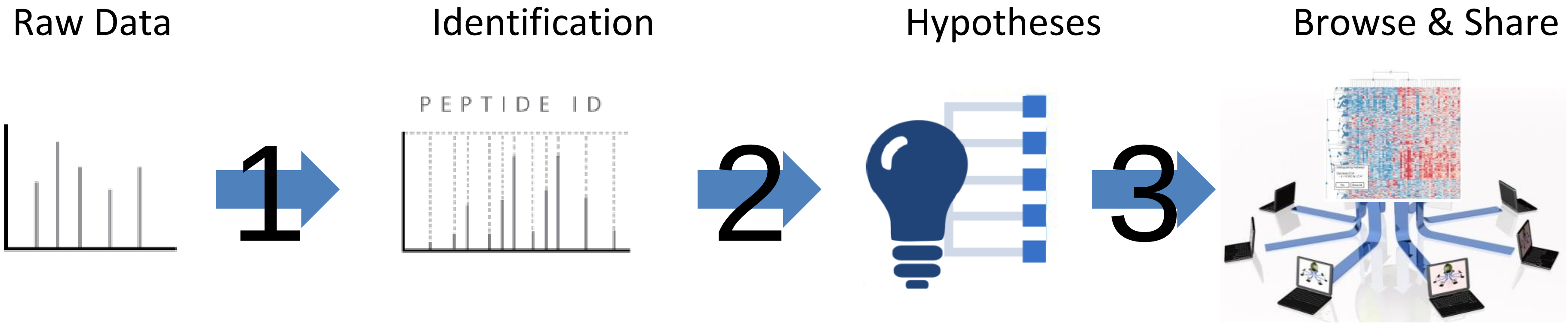
For ANY organism

- Bioinformatics data mining
  - Proteotypic machine learning
  - LC elution prediction
  - Scoring functions

- 112 organisms
- 3 million peptides @ FDR ~ 0.00009
- All data publicly available at MassIVE

A Meaningful Context

- I have data, what now?



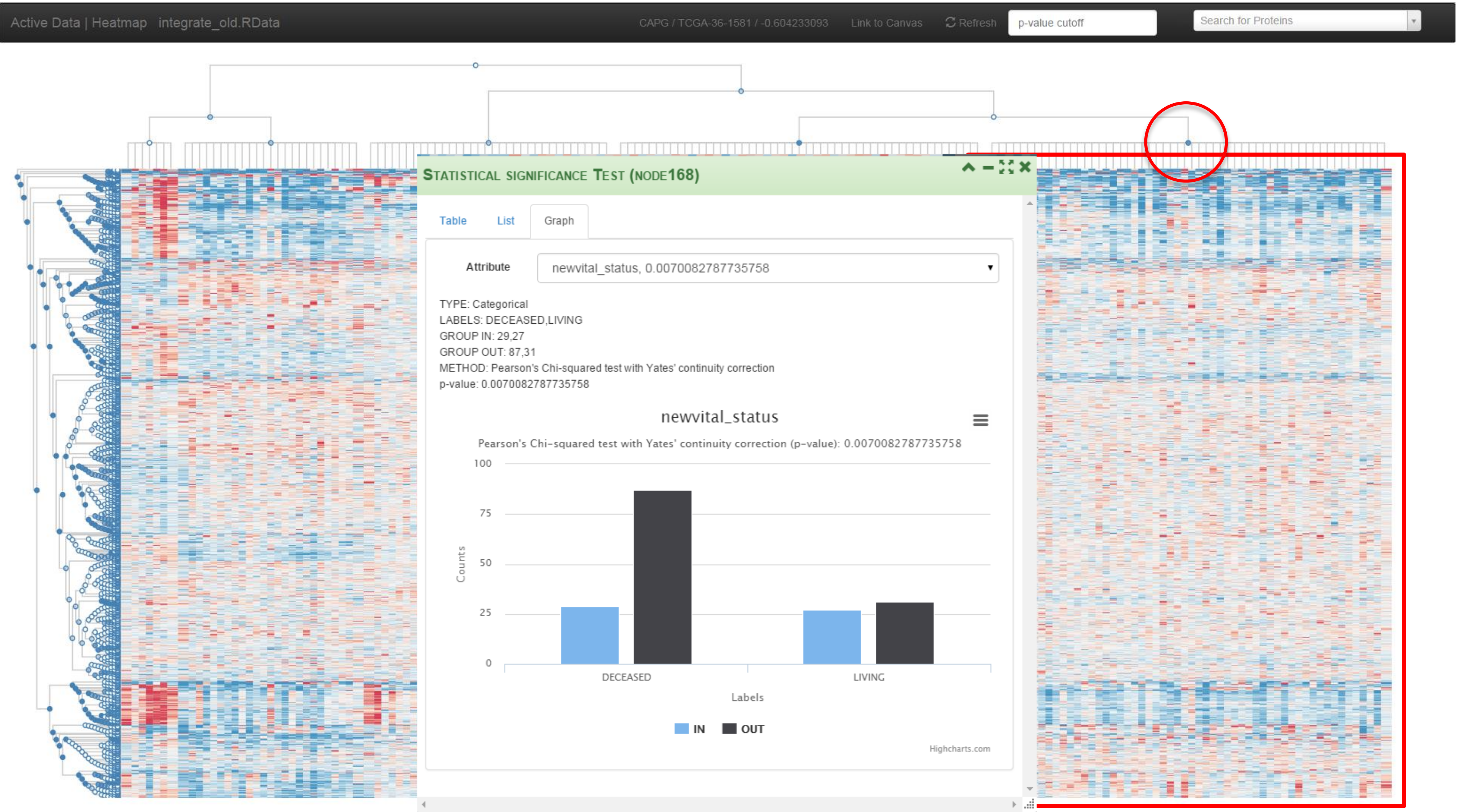
Heatmap View

Ovarian Cancer Data

174 tumor samples

1,608 proteins

Heatmap View



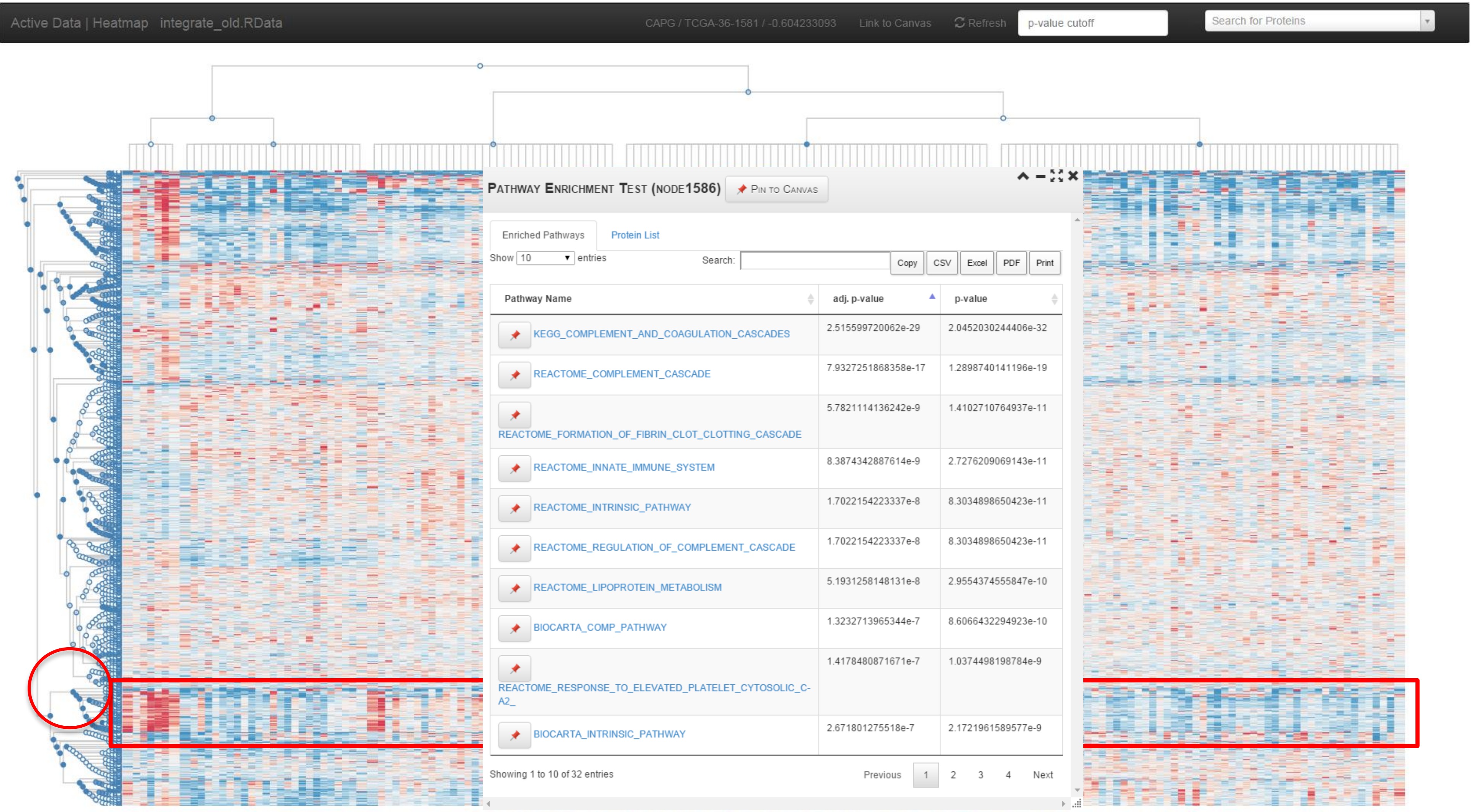
STATISTICAL SIGNIFICANCE TEST (NODE168)

Attribute: newvital_status, 0.0070082787735758

TYPE: Categorical
LABELS: DECEASED,LIVING
GROUP IN: 29,27
GROUP OUT: 87,31
METHOD: Pearson's Chi-squared test with Yates' continuity correction
p-value: 0.0070082787735758

Pearson's Chi-squared test with Yates' continuity correction (p-value): 0.0070082787735758

Heatmap View



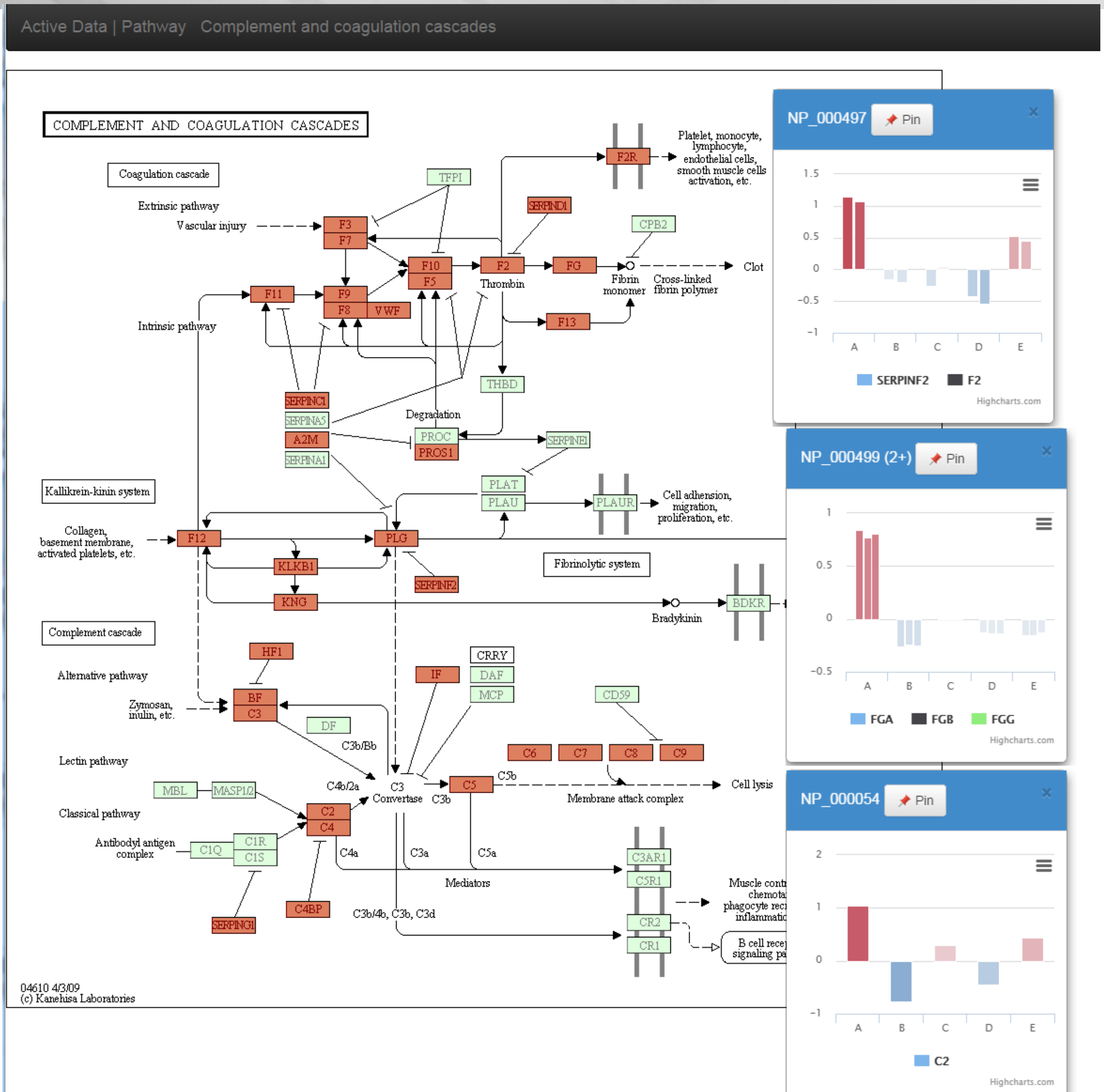
Pathway Enrichment Test (node1586)

Pathway Name	adj. p-value	p-value
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	2.515599720062e-29	2.0452030244406e-32
REACTOME_COMPLEMENT_CASCADE	7.9327251868358e-17	1.2898740141196e-19
REACTOME_FORMATION_OF_FIBRIN_CLOT_CLOTTING_CASCADE	5.7821114136242e-9	1.4102710764937e-11
REACTOME_INNATE_IMMUNE_SYSTEM	8.3874342887614e-9	2.7276209069143e-11
REACTOME_INTRINSIC_PATHWAY	1.7022154223337e-8	8.3034898650423e-11
REACTOME_REGULATION_OF_COMPLEMENT_CASCADE	1.7022154223337e-8	8.3034898650423e-11
REACTOME_LIPOPROTEIN_METABOLISM	5.1931258148131e-8	2.9554374555847e-10
BIOCARTA_COMP_PATHWAY	1.3232713965344e-7	8.6066432294923e-10
REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2_	1.4178480871671e-7	1.0374498198784e-9
BIOCARTA_INTRINSIC_PATHWAY	2.671801275518e-7	2.1721961589577e-9

Showing 1 to 10 of 32 entries

KEGG Overlay

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



Acknowledgements

- Biodiversity Library and Plugin
  - Grant Fujimoto
  - Michael Degan, Chris Overall, Matt Monroe, Lillian Ryadinskiy
- Active Data Canvas
  - Joon-Yong Lee, Nick Cramer
  - Ryan Wilson, Gary Kiebel, Grant Fujimoto
- Dick Smith
- Poster Wednesday 312 (Biodiversity Library)
- Poster Thursday 466 (Active Data Canvas)

National Institute of
General Medical Sciences

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA