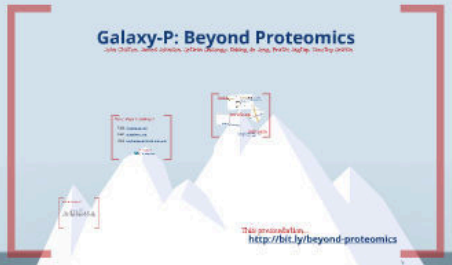


Toolbox Filters

Large Numbers of Files

Windows

LWR

Tool Macros

Cloud

Binary Data Types

Galaxy-P: Beyond Proteomics

Galaxy-P: Beyond Proteomics

John Chilton, James Johnson, Getiria Onsongo, Ebbing de Jong, Pratik Jagtap, Timothy Griffin

Three ways to Galaxy-P

Public usegalaxyp.org

Local getgalaxyp.org

Cloud biocloudcentral.msi.umn.edu

...or maybe 4

Twitter @usegalaxyp

Tools

Workflows

Outreach

What is Galaxy-P?

This presentation...

http://bit.ly/beyond-proteomics

What is Galaxy-P?

nearing end of year 1

Three year NSF funded grant to build mass spec & proteomics data analysis platform on top of Galaxy.

Three ways to Galaxy-P

Public **usegalaxyp.org**

Local **getgalaxyp.org**

Cloud **biocloudcentral.msi.umn.edu**

...or maybe 4

Twitter **@usegalaxyp**

Tools

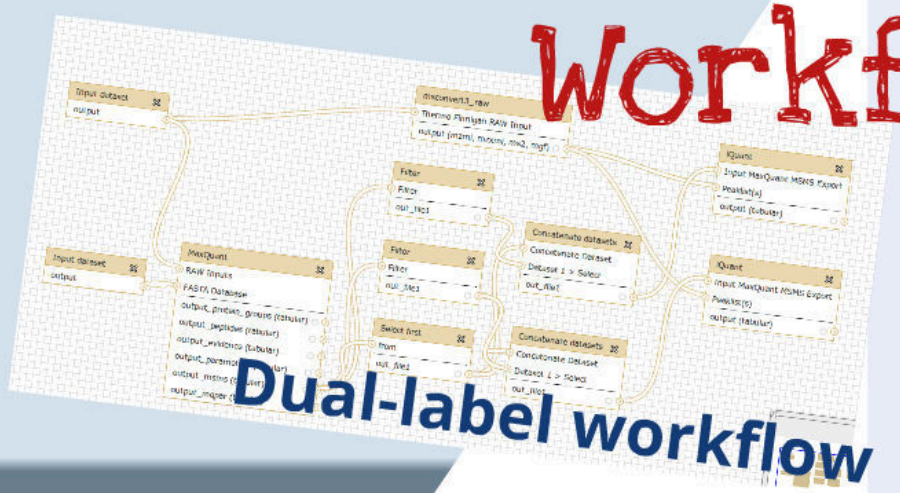
Category Proteomics

search repository name, description

Name ↓	Synopsis	Metadata Revisions	Tools Verified	Owner
appendfdr	Add false discovery rate to tabular data.	0:ef7cc296f063	no	galaxyp
bumbershoot	Tool wrappers for applications in the Tabb lab's Bumbershoot suite.	2:e07c19262374	no	galaxyp
dbbuilder	This tool allows users to download protein databases from common sources.	2:6c624803c5f9	no	galaxyp
	Galaxy tool wrapper for the			

Proteomics category has 3rd most repositories on tool shed.

Workflows



Dual-label workflow

Jagtap's Minnesota Two Step

Outreach

Huge presence at ASMS 2013, generated a lot of buzz.

Galaxy-P: Beyond Proteomics

John Chilton, James Johnson, Getiria Onsongo, Ebbing de Jong, Pratik Jagtap, Timothy Griffin

Three ways to Galaxy-P

Public usegalaxyp.org

Local getgalaxyp.org

Cloud biocloudcentral.msi.umn.edu

...or maybe 4

Twitter @usegalaxyp

Tools

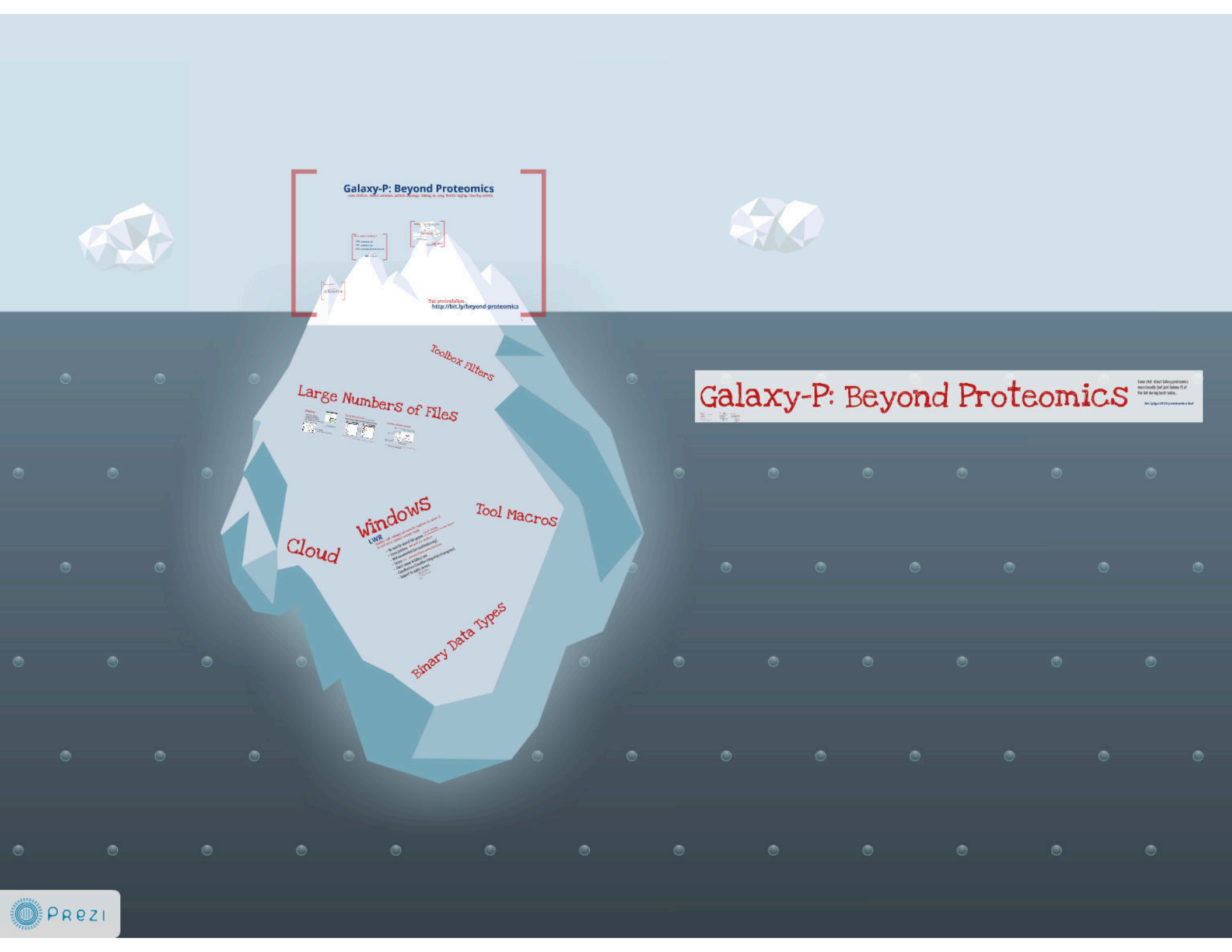
Workflows

Outreach

What is Galaxy-P?

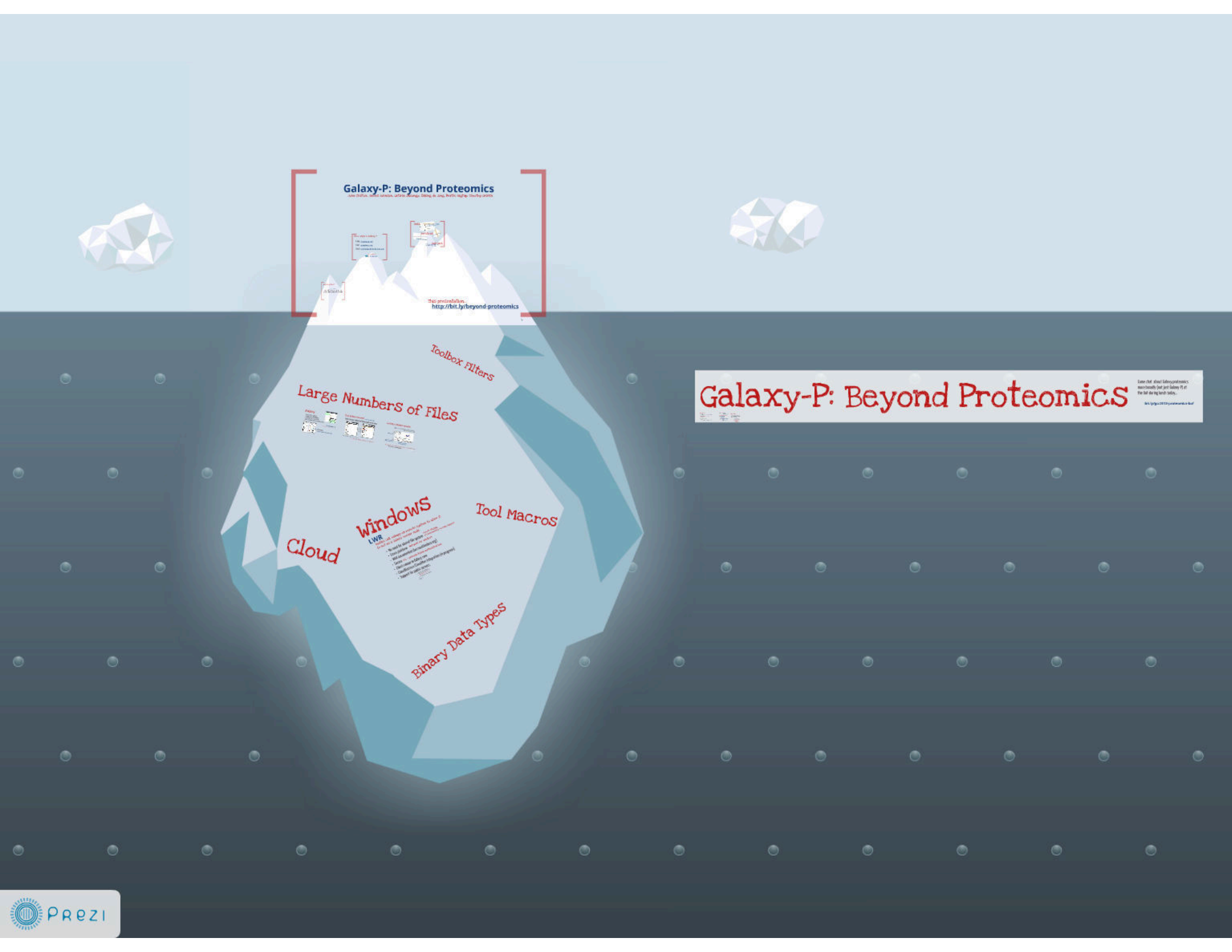
This presentation...

http://bit.ly/beyond-proteomics


Galaxy-P: Beyond Proteomics
http://bit.ly/beyond-proteomics
Toolbox Filters
Large Numbers of Files
Windows
LWR
Tool Macros
Cloud
Binary Data Types
Galaxy-P: Beyond Proteomics

Come chat about Galaxy proteomics more broadly (not just Galaxy-P) at the BoF during lunch today...

bit.ly/gcc2013-proteomics-bof


Galaxy-P: Beyond Proteomics
http://bit.ly/beyond-proteomics
Toolbox Filters
Large Numbers of Files
Windows
LWR
Tool Macros
Cloud
Binary Data Types
Galaxy-P: Beyond Proteomics

windows

LWR

Install LWR webapp on remote system to allow it to act as a Galaxy worker node.

- No need for shared file system — remote staging w/experimental caching support
- Cross-platform — Not just for windows
- Well documented (lwr.readthedocs.org)
- Secure — SSL + private token authentication
- Client runner in Galaxy now
- CloudBioLinux/CloudMan integration (in progress)
- Support for public servers.

public servers

- Point LWR at a toolbox to describe what it can run.
- Markup your standard Galaxy tool XML to lock down how it runs. then any Galaxy instance can run these tools on your resource
- Easily share access to specialized hardware, large datasets, etc...
- One click access...

http://bit.ly/galaxy-extras

```
<tool>
  ...
  <default_destination runner="lwr">
    <param id="url">https://example.com:8913/</param>
  </default_destination>
  ...
```

Publish tools with default public server instances to the tool shed

ne click access

http://bit.ly/galaxy-extras

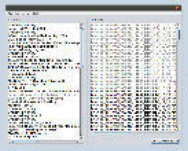
```
<tool>
  ...
  <default_destination runner="lwr">
    <param id="url">https://examle.com:8913/</param>
  </default_destination>
  ...
```

Publish tools with default public server instances to the tool shed and provide one-click access to your compute and data.

Large Numbers of Files

JGalaxy

- Java Web Start Application
- Batch Download and Upload Files
- Get around browser upload limits
- Many API enhancements contributed upstream



bit.ly/galaxy-extras

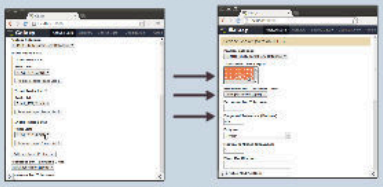
Optional Galaxy extensions for direct access

Built on blend4j

github.com/jmchilton/blend4j

Tool Enhancements (in your Galaxy now)

Tools can now take in multiple inputs at all once.

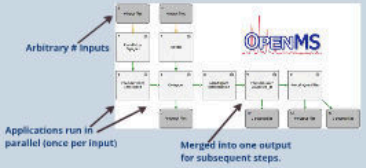


Convenient for a specifying a few inputs, necessary for dozens or hundreds.

Replace dataset <repeat> elements with <input type="data" multiple="true">

Workflow Enhancements

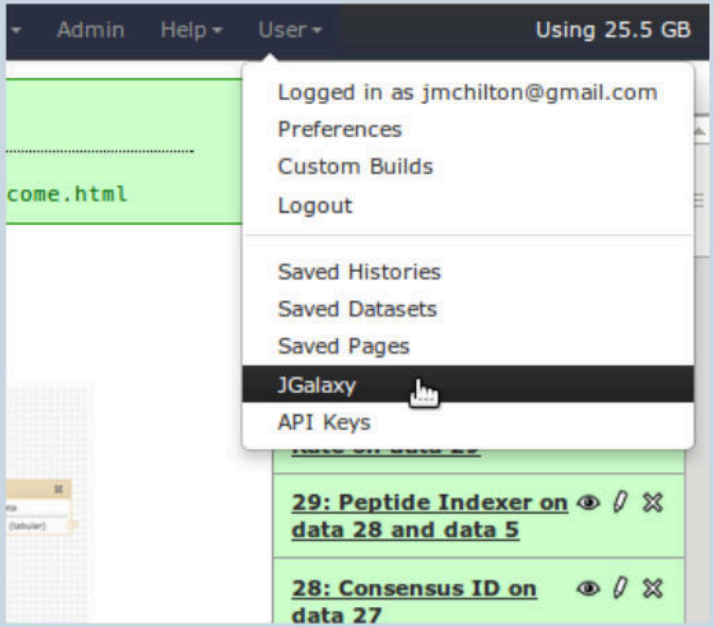
"An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics (J. Proteome Res., 2013, PMID: 23391308)."



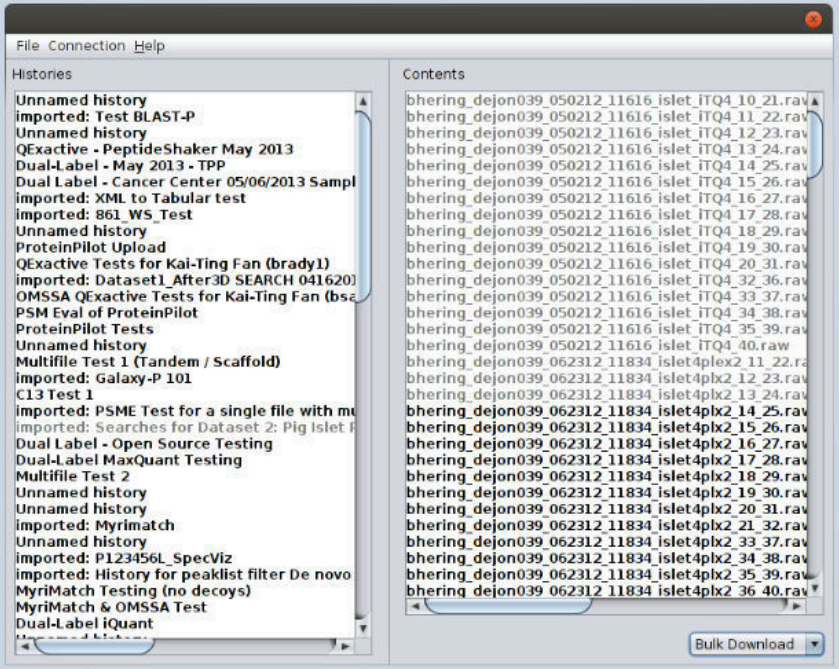
Galaxy can run these tools, but it cannot build this workflow. Galaxy-P can!

http://toolshed.g2.bx.psu.edu/view/jpaprotsys/openms

JGalaxy



- Java Web Start Application
- Batch Download and Upload Files
- Get around browser upload limits
- Many API enhancements contributed upstream



bit.ly/galaxy-extras

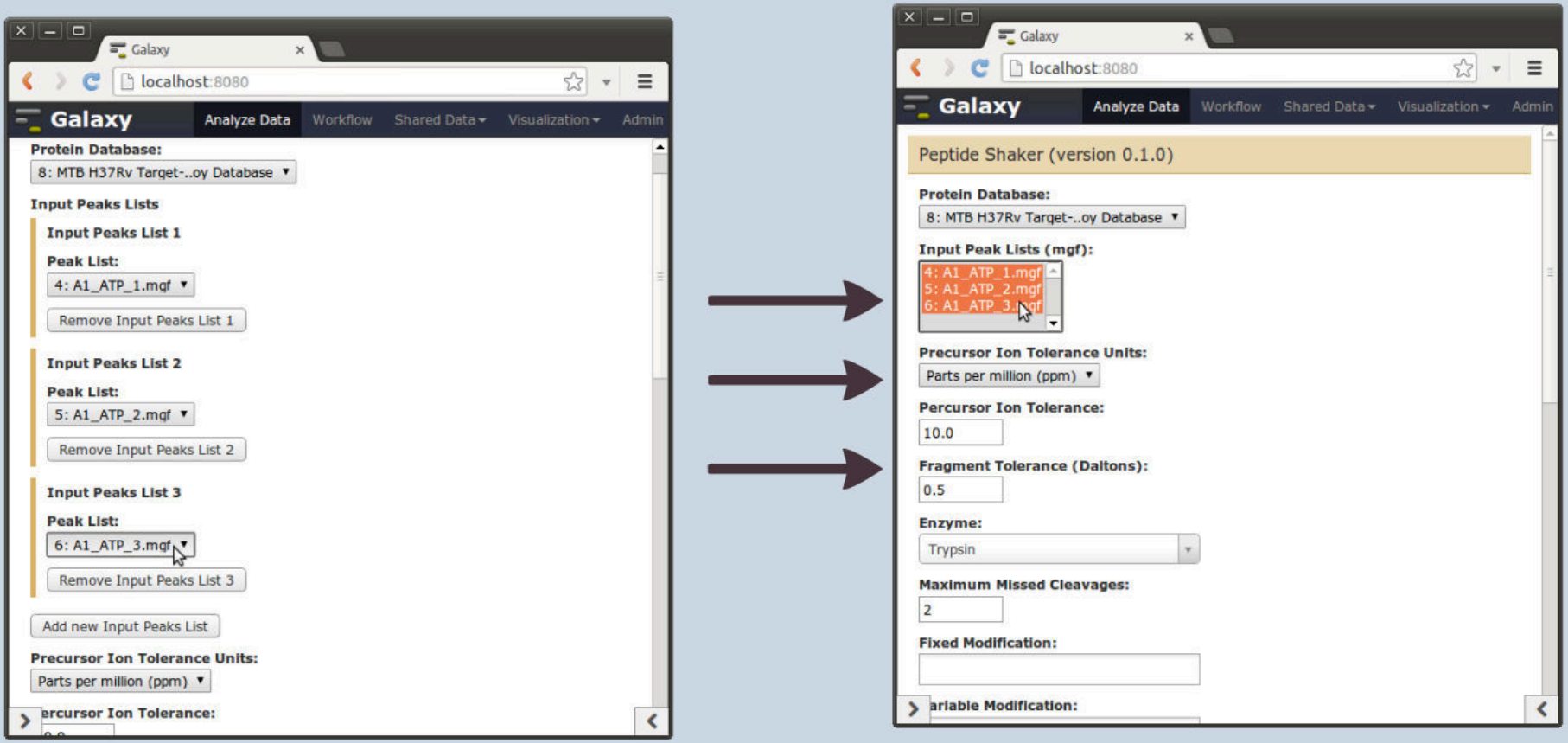
Optional Galaxy extensions for direct access

Built on blend4j

github.com/jmchilton/blend4j

Tool Enhancements (in your Galaxy now)

Tools can now take in multiple inputs at all once.

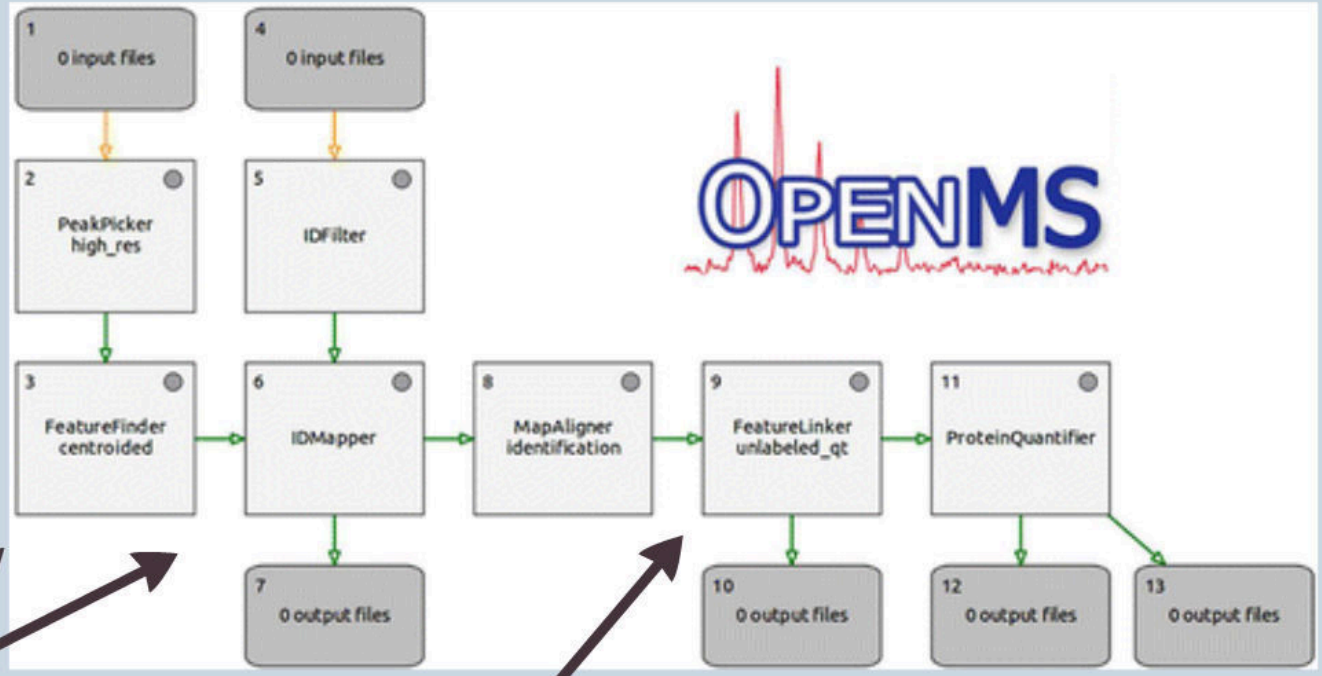


Convenient for a specifying a few inputs, necessary for dozens or hundreds.

Replace dataset ‹repeat› elements with ‹input type="data" multiple="true"›.

Workflow Enhancements

"An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics (J. Proteome Res., 2013, PMID: 23391308)."



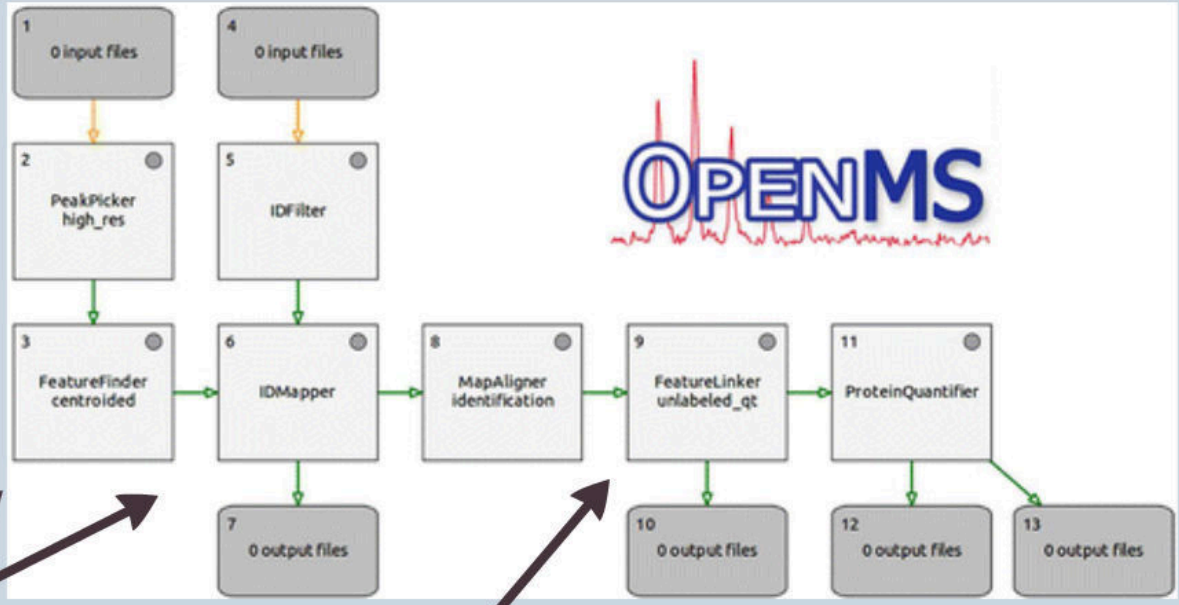
Arbitrary # Inputs

Applications run in parallel (once per input)

Merged into one output for subsequent steps.

Workflow Enhancements

"An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics (J. Proteome Res., 2013, PMID: 23391308)."



Arbitrary # Inputs

Applications run in parallel (once per input)

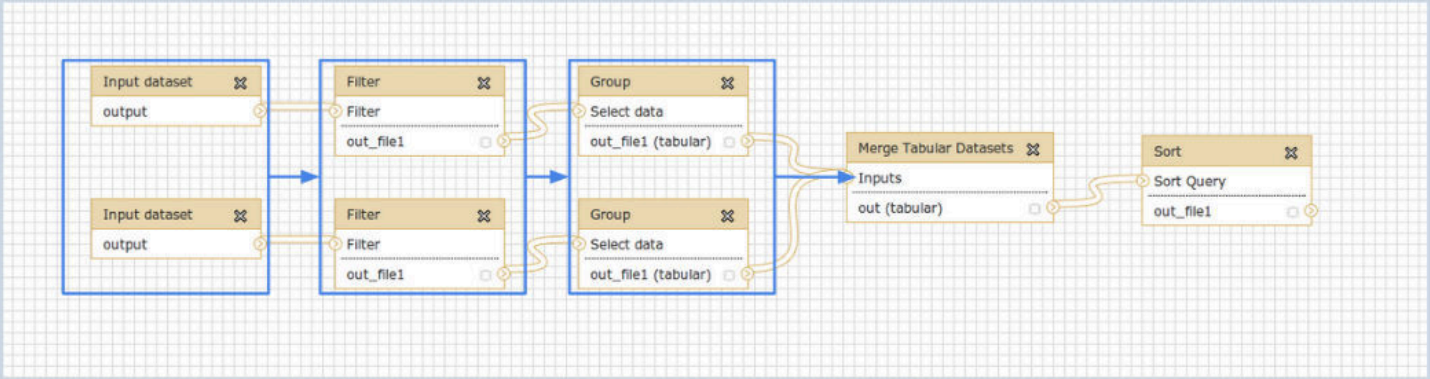
Merged into one output for subsequent steps.

Galaxy can run these tools, but it cannot build this workflow, Galaxy-P can!

http://toolshed.g2.bx.psu.edu/view/galaxyp/openms

Galaxy-P can!

Multiple File Datasets



...group multiple files into a single dataset to "flatten" workflow.

Not just for toy workflows...

Jagtap's "Shamelessly Seamless" Proteogenomics workflow

- 150+ steps - Multiple identifications, BLAST, custom tools for spectral validation and genome mapping
- Takes 3 days to run on a Fractionated 52 RAW file sample
- Multiple LWR steps.
- Tools and datatypes are NOT adapted for multiple file datasets

Not just for proteomics...

nothing proteomics related

bit.ly/galaxy-extras

Track or merge multiple file datasets into your Galaxy.

Other Applications:

Imaging:

"...it is needed for the community. I don't think we have other options for our requirements. [The] multiple file datasets implementation was a real saver for us."

Alex Khassapov, CSIRO

Genomics:

I am using it with success for chip-seq and rna-seq analysis.

Hagai Cohen, Hebrew University

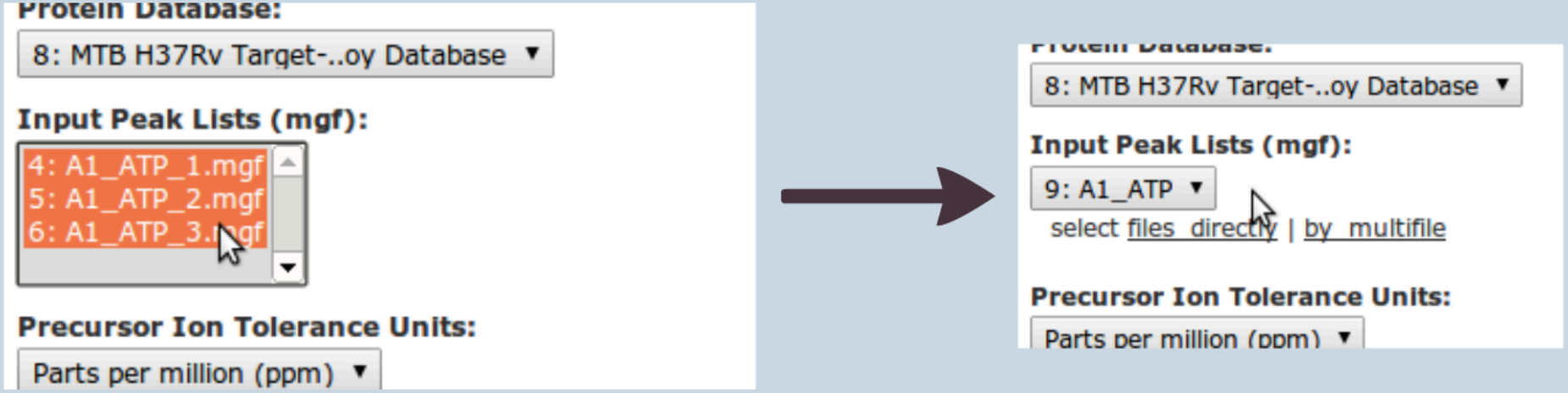
Not just improved grouping...

Addresses additional limitations of Galaxy.

Sample tracking throughout complex workflows...

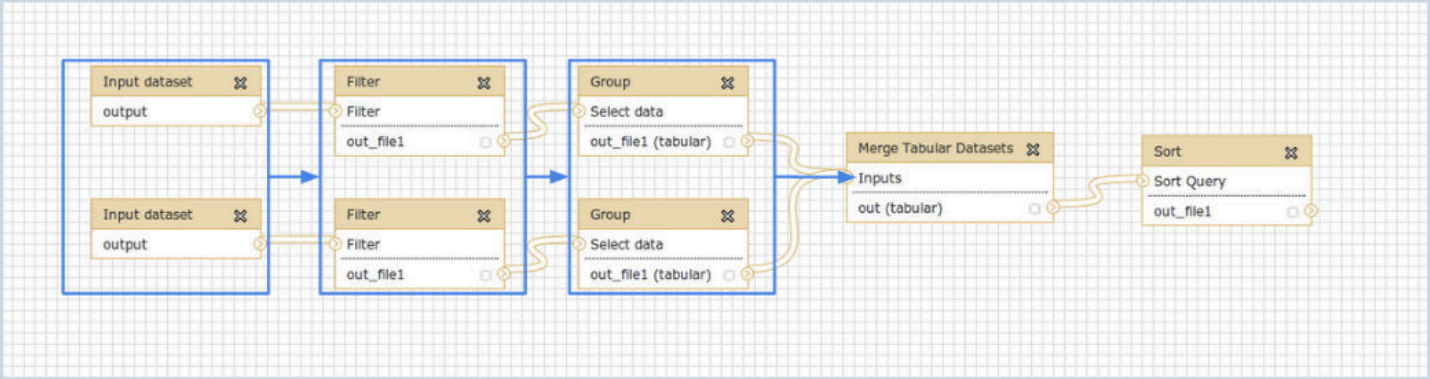
Tools with many-to-many outputs...

Merging or many-to-one tools...



Galaxy-P can!

Multiple File Datasets



...group multiple files into a single dataset to "flatten" workflow.

Not just for toy workflows...

Jagtap's "Shamelessly Seamless" Proteogenomics workflow

- 150+ steps - Multiple identifications, BLAST, custom tools for spectral validation and genome mapping
- Takes 3 days to run on a Fractionated 52 RAW file sample
- Multiple LWR steps.
- Tools and datatypes are NOT adapted for multiple file datasets

Not just for proteomics...

nothing proteomics related

bit.ly/galaxy-extras

Track or merge multiple file datasets into your Galaxy.

Other Applications:

Imaging:

"...it is needed for the community. I don't think we have other options for our requirements. [The] multiple file datasets implementation was a real saver for us."

Alex Khassapov, CSIRO

Genomics:

I am using it with success for chip-seq and rna-seq analysis.

Hagai Cohen, Hebrew University

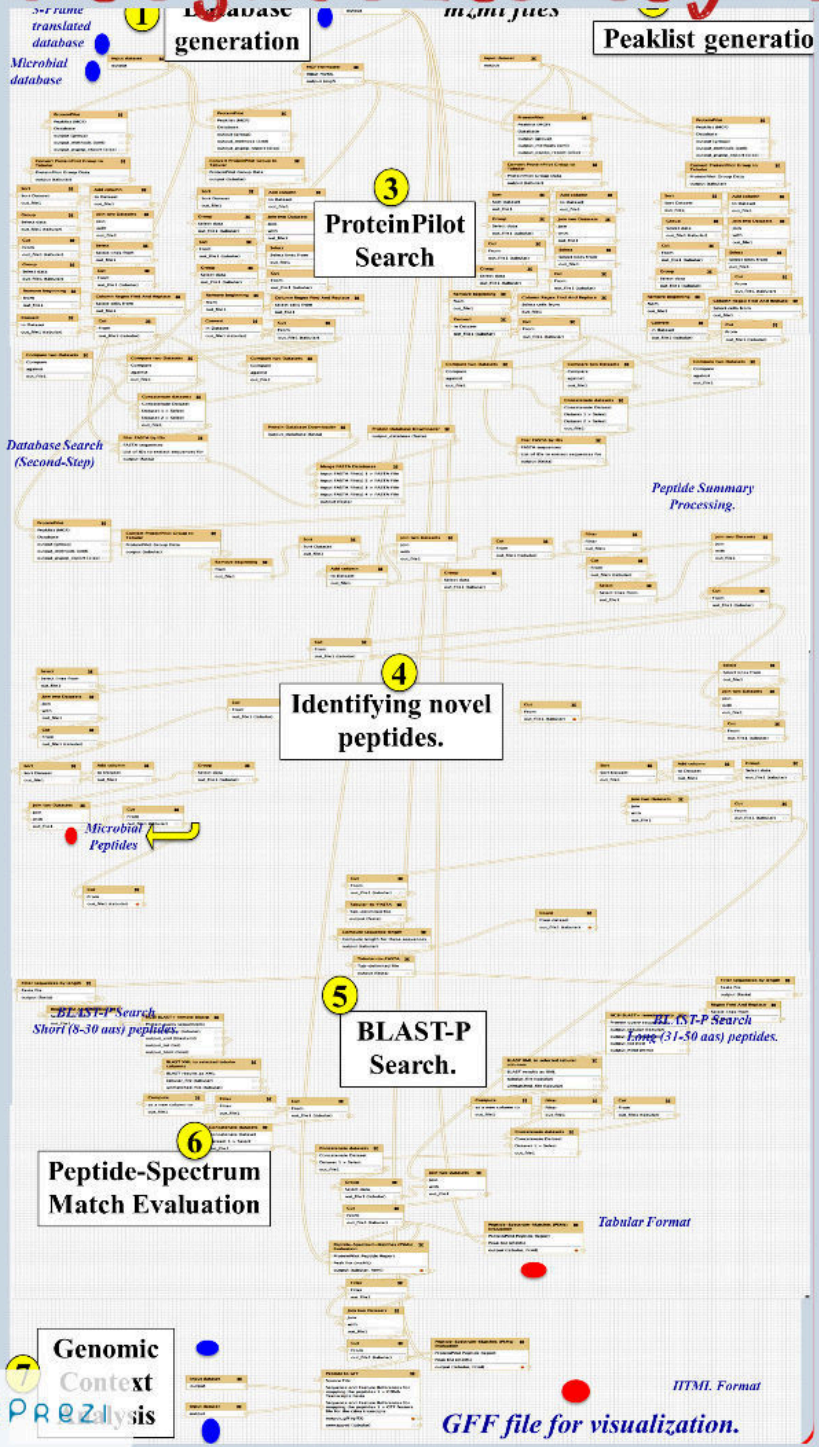
Not just improved grouping...

Addresses additional limitations of Galaxy.

Sample tracking throughout complex workflows...

Tools with many-to-many outputs...

Not just for toy workflows...



Jagtap's "Shamelessly Seamless" Proteogenomics workflow

- 150+ steps - Multiple identifications, BLAST, custom tools for spectral validation and genome mapping.
- Takes 3 days to run on a fractionated 52 RAW file sample
- Multiple LWR steps.
- Tools and datatypes are NOT adapted for multiple file datasets

Not just for proteomics...

nothing proteomics related

bit.ly/galaxy-extras

Track or merge multiple file datasets into your Galaxy.

Other Applications:

Imaging:

"... it is needed for the community. I don't think we have other options for our requirements, [the] multiple file datasets implementation was a real savior for us."

Alex Khassapov, CSIRO

Genomics:

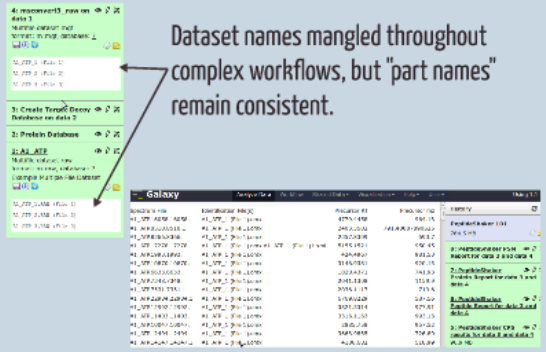
I am using it with success for chip-seq and rna-seq analysis.

Hagai Cohen, Hebrew University

Not just improved grouping...

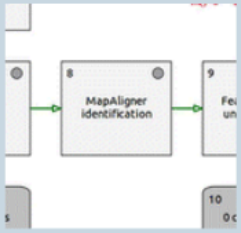
Addresses additional limitations of Galaxy.

Sample tracking throughout complex workflows...



Dataset names mangled throughout complex workflows, but "part names" remain consistent.

Tools with many-to-many outputs...



Aligns feature space of each input against all others...

With current implementation, the tool does "need to know" it produces multiple file dataset.

Sample tracking th

4: msconvert3_raw on data 1
Multifile dataset mgf
format: m:mgf, database: ?

A1_ATP_1 (File 1)
A1_ATP_2 (File 2)
A1_ATP_3 (File 3)

3: Create Target-Decoy Database on data 2

2: Protein Database

1: A1_ATP
Multifile dataset raw
format: m:raw, database: ?
Example Multiple File Dataset

A1_ATP_1.RAW (File 1)
A1_ATP_2.RAW (File 2)
A1_ATP_3.RAW (File 3)

Dataset names mangled throughout complex workflows, but "part names" remain consistent.

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 1.1

Spectrum File	Identification File(s)	Precursor RT	Precursor mz
A1_ATP.16058.16058.	A1_ATP_1 (File 1).omx	4779.4458	964.13
A1_ATP.8510.8510.1	A1_ATP_1 (File 1).omx	2403.5502	791.40087890625
A1_ATP.8366.8366.	A1_ATP_1 (File 1).omx	2357.8309	693.7
A1_ATP.17278.17278.	A1_ATP_1 (File 1).omx A1_ATP_1 (File 1).t.xml	5155.1521	950.45
A1_ATP.1993.1993.	A1_ATP_1 (File 1).omx	424.4967	981.53
A1_ATP.10870.10870.	A1_ATP_1 (File 1).omx	3146.9381	920.13
A1_ATP.6633.6633.	A1_ATP_1 (File 1).omx	1820.4371	741.93
A1_ATP.7348.7348.	A1_ATP_1 (File 1).omx	2041.1308	1159.9
A1_ATP.7331.7331.1	A1_ATP_1 (File 1).omx	2036.1117	719.6
A1_ATP.22984.22984.1	A1_ATP_1 (File 1).omx	6769.8229	587.56
A1_ATP.12992.12992.	A1_ATP_1 (File 1).omx	3821.3314	977.91
A1_ATP.11403.11403.	A1_ATP_1 (File 1).omx	3316.1163	993.13
A1_ATP.10047.10047.	A1_ATP_1 (File 1).omx	2885.758	957.52
A1_ATP.12494.12494.	A1_ATP_1 (File 1).omx	3663.9835	726.89
A1_ATP.14147.14147.1	A1_ATP_1 (File 1).omx	4186.682	518.99

History

PeptideShaker 101
764.5 MB

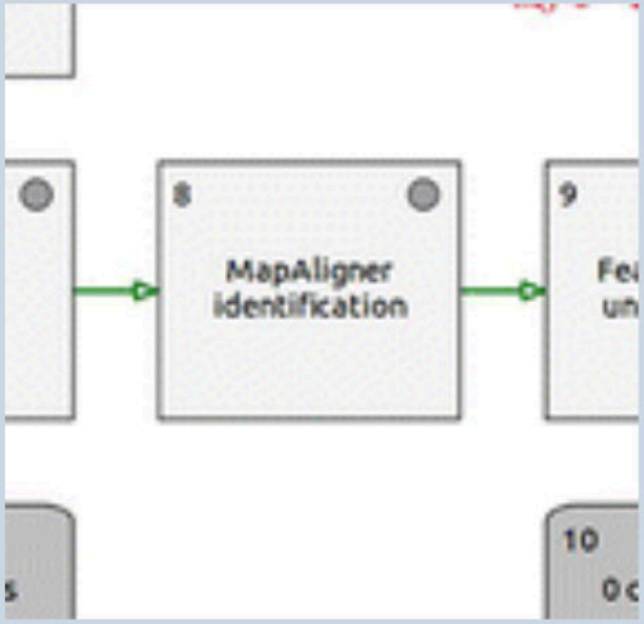
8: PeptideShaker PSM Report for data 3 and data 4

7: PeptideShaker Protein Report for data 3 and data 4

6: PeptideShaker Peptide Report for data 3 and data 4

5: PeptideShaker CPS results for data 3 and data 4
96.9 MB

many-to-many outputs...



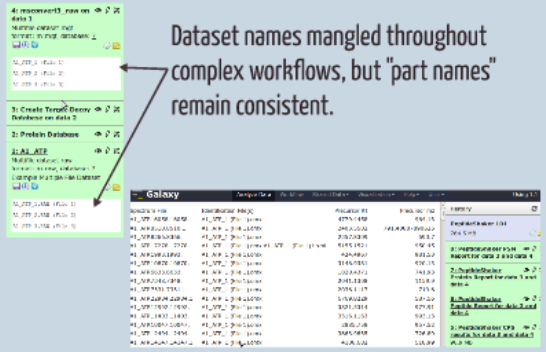
Aligns feature space of each input against all others...

with current implementation, the tool does "need to know" it produces multiple file dataset.

Not just improved grouping...

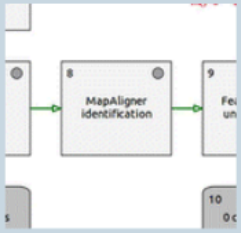
Addresses additional limitations of Galaxy.

Sample tracking throughout complex workflows...



Dataset names mangled throughout complex workflows, but "part names" remain consistent.

Tools with many-to-many outputs...



Aligns feature space of each input against all others...

With current implementation, the tool does "need to know" it produces multiple file dataset.

Special Thanks...

Jorrit Boekel

Implemented an early version of multiple file datasets at tool and datatype level.

Galaxy-P and Galaxy Teams @ MSI

w/Special, Special thanks to Anne-Francoise Lamblin & Benjamin Lynch

Dannon, Nate, Enis, Brad, Jeremy, Ira & Dave Clements

Hardest working man in Galaxy

People in Galaxy community who helped with various parts of this

Have mass spec data?

Galaxy-P!

Three ways to Galaxy-P...

Public **usegalaxyp.org**

Local **getgalaxyp.org**

Cloud **biocloudcentral.msi.umn.edu**

or maybe 4...

Twitter **@usegalaxyp**

Why fork?

Because the Galaxy team would like a more flexible, deeply integrated concept of dataset grouping.

I do too!

Galaxy team doesn't want to support this.

Make me a committer on Bitbucket, problem solved!

Encumber the code base with special cases. (Technical debt.)

Pull Requests:

#86: Datatype Tracking Enhancements

#169: Support config files with task parallelism.

Galaxy team doesn't want to support this.

Make me a committer on Bitbucket, problem solved!

Encumber the code base with special cases. (Technical debt.)

Pull Requests:

#86: Datatype Tracking Enhancements

#169: Support config files with task parallelism.

#87: Framework for per job task parallelism.

#:83 More flexible task merging.

#122: Eliminate hard coding of upload1 tool hacks.

#133: Fix file_ext in image datatypes.

#123: Refactor huge functions in library_common.

#:142 Simplify task splitting input specification.

#156: Fix 'from_work_dir' with task splitting.