Galaxy-P: Beyond Proteomics
Galaxy-P: Beyond Proteomics

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

This presentation...
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
Proteomics category has 3rd most repositories on tool shed.

Jagtap's Minnesota Two Step

Dual-label workflow

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: usegalaxy.org
- Local: getgalaxy.org
- Staff: biocloudcentral.mri.umn.edu

Outreach

Tools, Workflows

This presentation...
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

- Cloud
- Windows
- Large Numbers of Files
- Binary Data Types
- Tool Macros
- TexEx Filters
Windows LWR

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

- No need for shared file system
- Cross-platform
- 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. \textit{then any Galaxy instance can run these tools on your resource}

• Easily share access to specialized hardware, large datasets, etc...
• One click access...

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 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 proteomics...

Not just for proteomics...

Not just for proteomics...

Not just improved grouping...

Addresses additional limitations of Galaxy.

Other applications:
- Other applications
- Other applications
- Other applications
Merging or many-to-one tools...

<table>
<thead>
<tr>
<th>Protein Database:</th>
<th>8: MTB H37Rv Target-..oy Database ▼</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Peak Lists (mgf):</td>
<td>4: A1_ATP_1.mgf</td>
</tr>
<tr>
<td></td>
<td>5: A1_ATP_2.mgf</td>
</tr>
<tr>
<td></td>
<td>6: A1_ATP_3.mgf</td>
</tr>
<tr>
<td>Precursor Ion Tolerance Units:</td>
<td>Parts per million (ppm) ▼</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein Database:</th>
<th>8: MTB H37Rv Target-..oy Database ▼</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Peak Lists (mgf):</td>
<td>9: A1_ATP ▼</td>
</tr>
<tr>
<td></td>
<td>select files directly</td>
</tr>
<tr>
<td>Precursor Ion Tolerance Units:</td>
<td>Parts per million (ppm) ▼</td>
</tr>
</tbody>
</table>
Galaxy-P can!

Multiple File Datasets

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

Not just for proteomics...

- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.

Not just for proteomics...

- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.

Not just for proteomics...

- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.
- Addresses additional limitations of Galaxy.

Not just for proteomics...

- Addresses additional limitations of Galaxy.
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...

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...

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

Tools with many-to-many outputs...

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

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

With current implementation, the tool does "need to know" if 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.
#87: Framework for use job 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.