

# **Public sharing of complex MS-based qualitative and quantitative proteomic data analysis workflows: adding value to big data repositories**

ASMS annual conference  
June 16, 2014

Tim Griffin  
tgriffin@umn.edu



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**

# Outline

- Sharing “big data” in proteomics
- Historical perspective: sharing results in MS-based proteomics
- A way forward: The Galaxy framework
- A strategy for data sharing via public repositories using Galaxy
- Concluding thoughts



# Acknowledgements



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

UNIVERSITY OF MINNESOTA

SUPERCOMPUTING  
INSTITUTE



## Biochemistry, Molecular Biology and Biophysics

Dr. Julie Yang

Dr. Ebbing de Jong

Dr. Joel Kooren

Dr. Yue Chen

## Center for Mass Spectrometry and Proteomics

Dr. Pratik Jagtap

Dr. LeeAnn Higgins

James Johnson

John Chilton (Penn State)

Trevor Wennblom

Getiria Onsongo

Bart Gottschalk

Anne Lamblin

Ben Lynch

Attila Csordas

Henning Hermjakob

Juan Antonio Vizcaíno

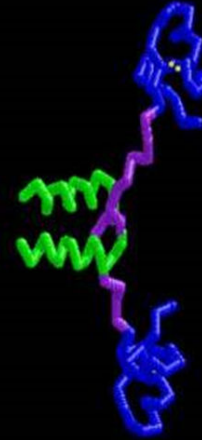
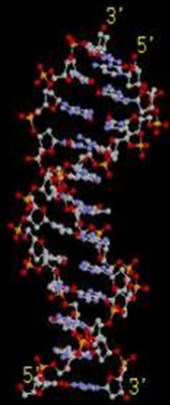


Funding  
NSF, NIH



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# The era of “Big Data” in the biological sciences



DNA  
*Genome*

RNA  
*Transcriptome*

Protein  
*Proteome*

Metabolite  
*Metabolome*

*High-throughput sequencing*

*High resolution  
mass spectrometry*



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**

## TECHNOLOGY FEATURE

# THE BIG CHALLENGES OF BIG DATA

*Nature* 2013 **498**:255-60

### .....and opportunities:

- Promotes reproducibility
- Data mining for new discoveries
- Creation of data resources (spectral libraries, etc)
- Re-analysis using new tools
  - evaluation and testing of new software
  - new results

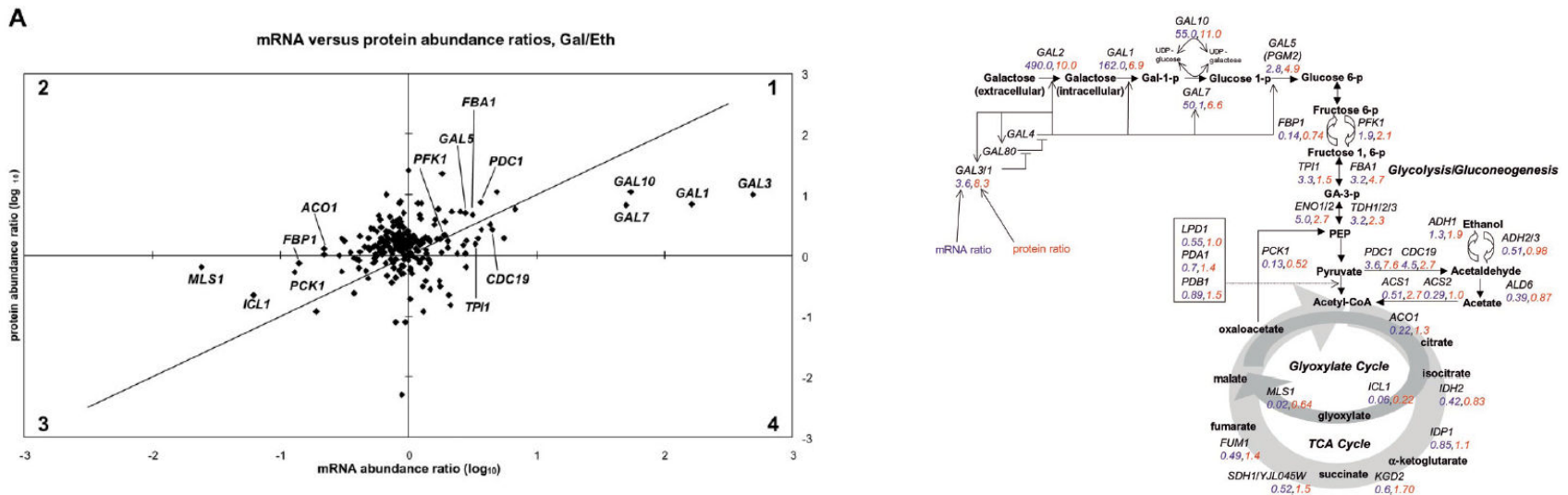


# A historical anecdote in quantitative proteomics

(or a confession of past sins)

## Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in *Saccharomyces cerevisiae*\*<sup>§</sup>

Timothy J. Griffin<sup>‡</sup>, Steven P. Gygi<sup>§</sup>, Trey Ideker<sup>¶</sup>, Beate Rist<sup>||</sup>, Jimmy Eng, Leroy Hood, and Ruedi Aebersold<sup>\*\*</sup>



- ICAT labeling for quantitative proteomics
- LCQ mass spectrometer
- DNA microarray containing ~6200 yeast ORFs

*Molecular & Cellular Proteomics* 1:323–333, 2002.



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# Data reproducibility?

## *MS-based proteomics*

The obtained MS/MS spectra were automatically searched against a data base of predicted proteins derived from the ~6100 open reading frames in the *S. cerevisiae* genome using the SEQUEST algorithm (30). The cleavage specificity for the protease used was not specified for the search, and oxidized methionines and ICAT reagent-labeled cysteines (both the d(0) and d(8) forms) were specified as static modifications in the search parameters. No sequence con-

...

sequence matches. Quantification of each identified protein was done by reconstructing the ion-chromatographic trace for the d(0) and d(8) form of each peptide and comparing the peak area for corresponding peptide pairs using XPRESS, a novel quantification software routine that enables visual inspection of reconstructed ion chromatograms for identified peptides (31). The criteria used in determining the ac-

???



*Molecular & Cellular Proteomics 1:323–333, 2002.*

- Raw and processed data accessibility?
- Analytical reproducibility?



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**

# Back to 2014: Big Data in MS-based proteomics



Mass Spectrometry  
Interactive Virtual Environment



- Raw and processed data archiving
- Tools for analysis and visualization
- Public availability for re-analysis



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**



# Enhancing Big Data Repositories: sharing the whole story



- A web-based, community developed bioinformatics framework/platform/workbench
- Originally designed to address issues in *genomic* informatics including:
  - Software accessibility and usability (disparate software integration)
  - Analytical transparency
  - Reproducibility
  - Scalability
  - **Share-ability: complete sharing of even complex workflows**

**usegalaxy.org**  
(*in development*)

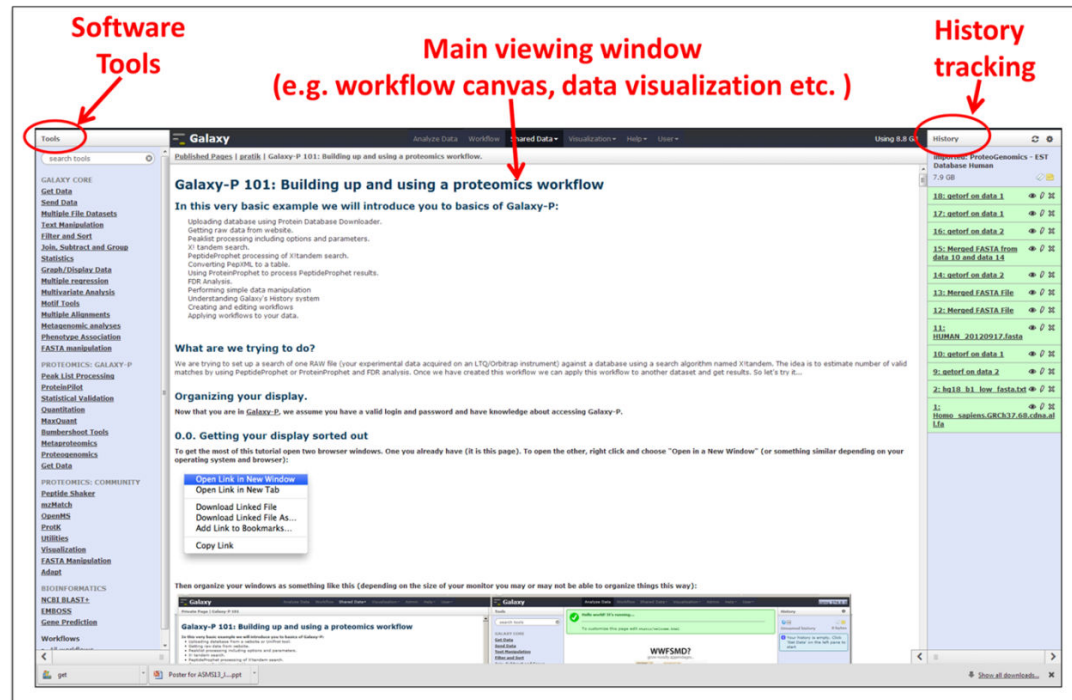


Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. **Genome Biol.** 2010, **11**: R86.



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**

# A (free) supermarket for 'omics software?

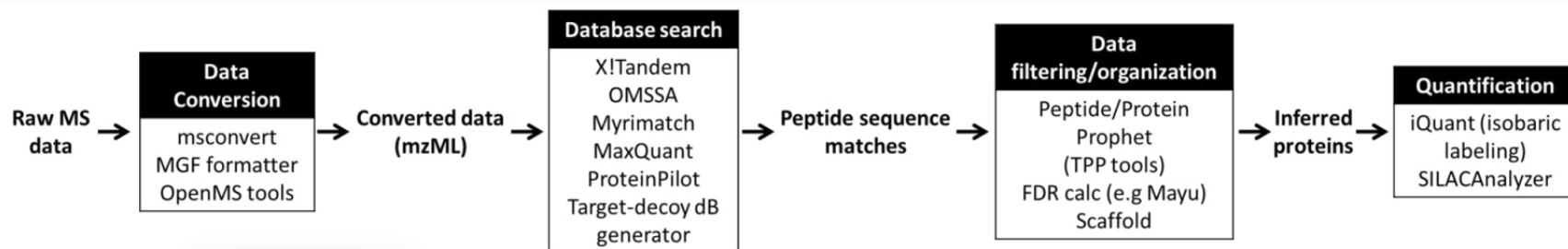


- Any command-line software can be deployed
- Amenable to Windows software (LWR)
- Multiple-file compatability

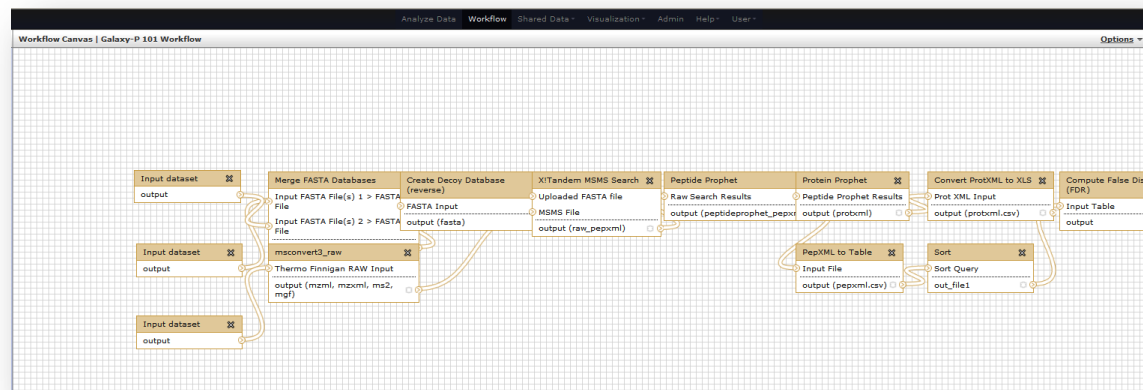


UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# Capturing complete MS-based proteomic workflows



History		
Galaxy-P 101 1.0 GB		
15: Cut on data 14	🔍	🗑️
14: Add column on data 13	🔍	🗑️
13: Sort on data 11 with FDR	🔍	🗑️
12: Sort on data 11	🔍	🗑️
11: Table peptide_prophet Raw101.pep.xml.csv	🔍	🗑️
10: Convert ProtXML to Tabular on data 9	🔍	🗑️
9: protein_prophet.peptide_prophet.X!Tandem vs Target Decoy Human Contaminants.Peaklist Raw101.Peaklist Raw101.pepXML.pep.xml.protXML	🔍	🗑️
8: peptide_prophet Raw101.pep.xml	🔍	🗑️
7: X!Tandem vs Target Decoy Human Contaminants.Peaklist Raw101.Peaklist Raw101.pepXML	🔍	🗑️
6: Peaklist Raw101	🔍	🗑️
5: Raw101.RAW	🔍	🗑️
4: Target_Decoy_Human_Contaminants on data 3	🔍	🗑️
3: Merged Human UniProt cRAP	🔍	🗑️
2: CRAP	🔍	🗑️
1: Human UniProt	🔍	🗑️

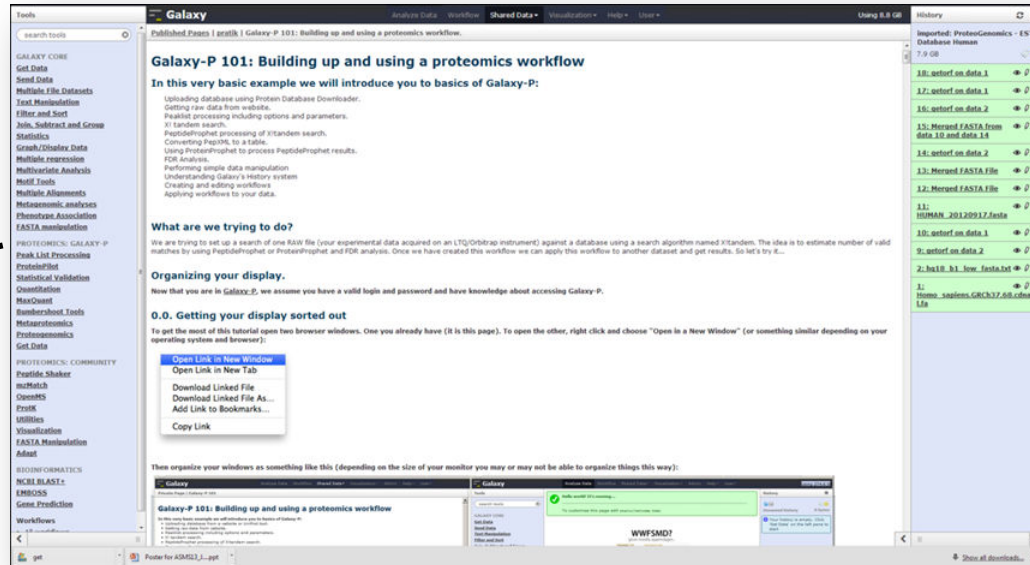


Galaxy workflow

Galaxy history



# Exporting complete and reproducible workflows



URL  
export

File export  
(.ga)

**HISTORY:** <https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold>

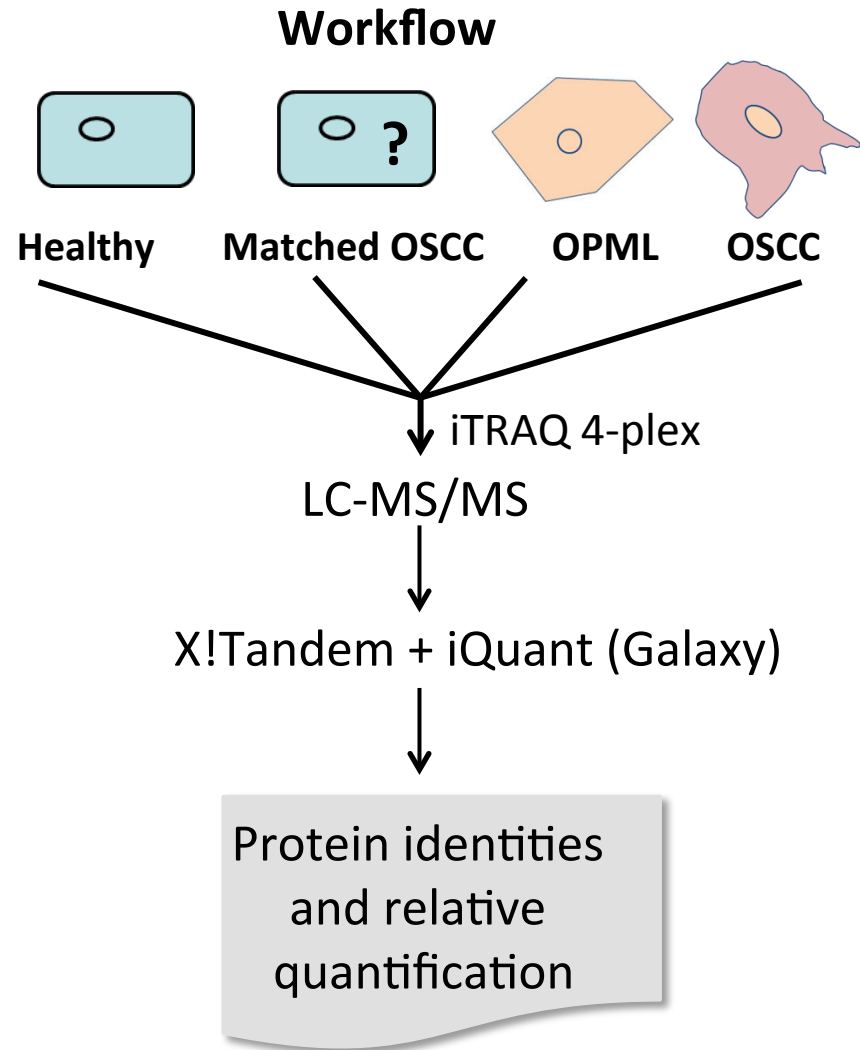
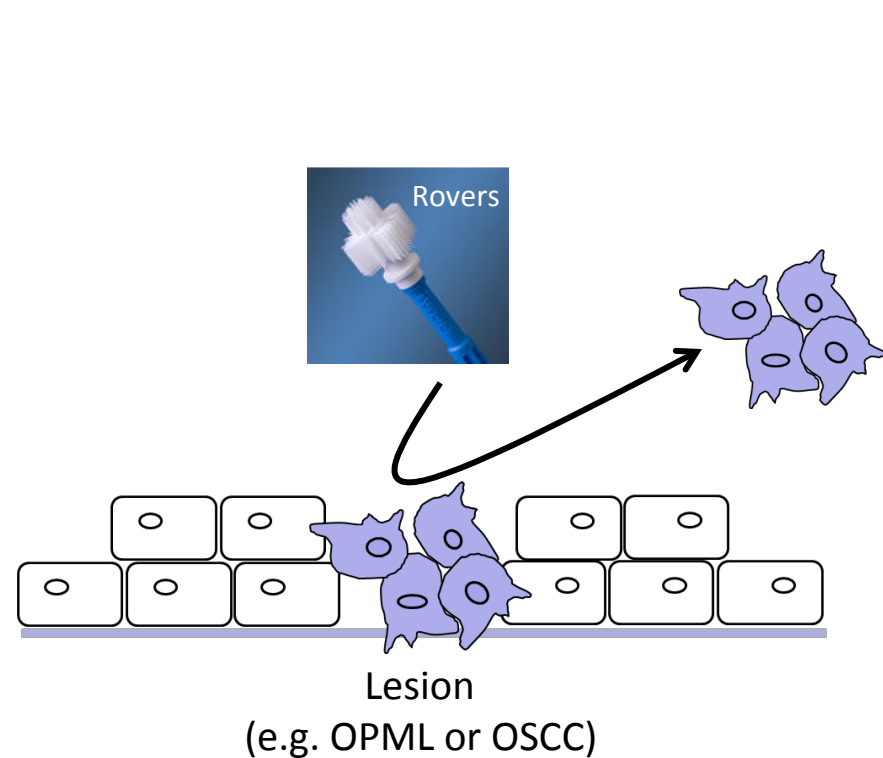
**WORKFLOW:** <https://galaxyp.msi.umn.edu/u/pjagtap/w/workflow-for-4-plex-itraq-xtandem-search-scaffold-processing>

Galaxy-Workflow-Workflow\_for\_4-  
plex\_ITRAQ\_X\_tandem\_Search\_Scaffold\_Processing.ga



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# Example: quantitative proteomic analysis of oral cancer

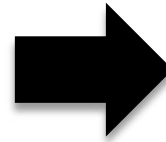


Dr. Julie Yang



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**

Can it be reproduced?



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**



# Submission to repository: ProteomeExchange



Welcome  
ProteomeXchange Submission Tool (version 2.0.1)

Choose submission option below

**Complete Submission**

✓ ✓  
✓ ✓

**Partial Submission**

✓ ✓  
○

[Resubmission](#) [Bulk submission](#) [Submission guidelines](#) [More about ProteomeXchange](#)

**You need to provide**

**Result Files**  
PRIDE XML or mzIdentML(+ spectra)



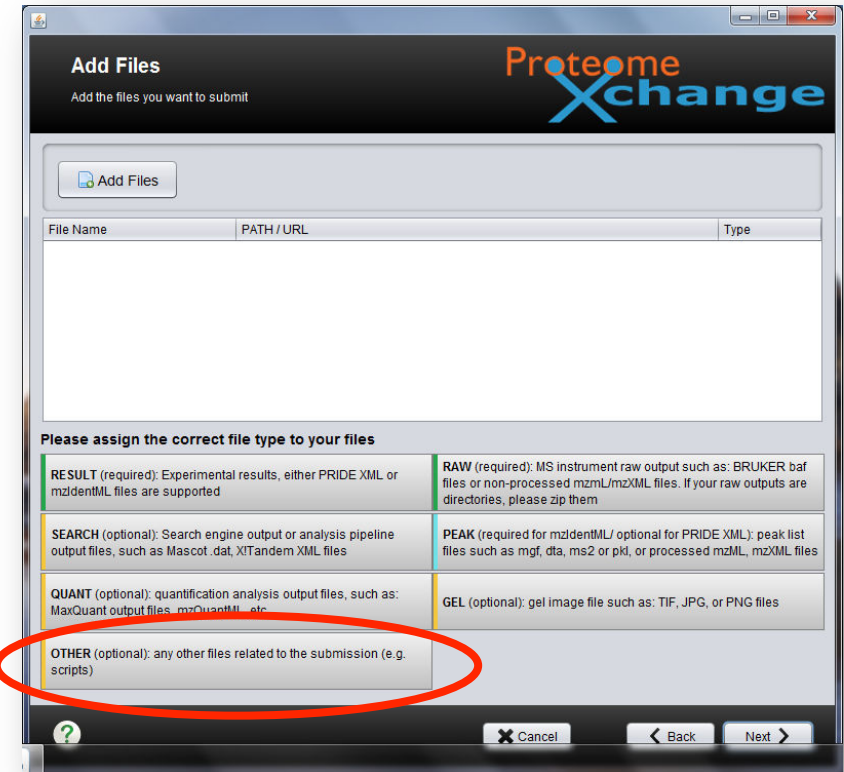
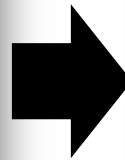
**Raw Data**  
MS instrument raw output



**PRIDE Login**  
PRIDE user credentials [Register](#)



? Cancel < Back Next >



Add Files  
Add the files you want to submit

Add Files

File Name	PATH / URL	Type
-----------	------------	------

**Please assign the correct file type to your files**

<b>RESULT</b> (required): Experimental results, either PRIDE XML or mzIdentML files are supported	<b>RAW</b> (required): MS instrument raw output such as: BRUKER baf files or non-processed mzML/mzXML files. If your raw outputs are directories, please zip them
<b>SEARCH</b> (optional): Search engine output or analysis pipeline output files, such as Mascot .dat, X!Tandem XML files	<b>PEAK</b> (required for mzIdentML/ optional for PRIDE XML): peak list files such as mgf, dta, ms2 or pkl, or processed mzML, mzXML files
<b>QUANT</b> (optional): quantification analysis output files, such as: MaxQuant output files, mzQuantML, etc.	<b>GEL</b> (optional): gel image file such as: TIF, JPG, or PNG files
<b>OTHER</b> (optional): any other files related to the submission (e.g. scripts)	

? Cancel < Back Next >



# ProteomeExchange Submission: Enhanced-Value

Project : PXD001044

PRIDE ASSIGNED TAGS: [Biomedical Dataset](#)

## Summary

**Title**  
Human oral cancer brush biopsy Galaxy-iTRAQ analysis

**Description**  
iTRAQ-based comparison of proteins derived from oral cells collected by brush biopsy. Protein abundance levels compared between oral pre-malignant cells, oral cancer cells and healthy normal cells, all collected from human patients. Two separate iTRAQ labeled biological replicate analyses were conducted. Analysis was achieved via a reproducible Galaxy-based workflow.

**Sample Processing Protocol**  
Cells were lysed, proteins digested with trypsin and iTRAQ labeled. Combined peptide mixtures were fractionated by high pH HPLC offline, and combined fractions were analyzed via LC-MS/MS on an Orbitrap Velos using HCD fragmentation.

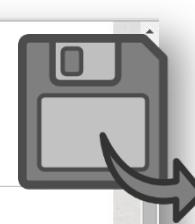
**Data Processing Protocol**  
Raw files were converted to mzXml using msconvert (distributed as part of ProteoWizard 1.6.1260). MS/MS spectra were searched against the Uniprot human database including scrambled sequences and common contaminant proteins (a total of 136,002 entries) using X!Tandem (CYCLONE release, 2013.2.01). Search parameters included a 1.6 amu (atomic mass units) precursor and 0.8 amu fragment mass tolerance, 2 missed cleavages, partial trypsin specificity, fixed modifications of carbamidomethylated cysteine, iTRAQ reagent modification at lysines and N-termini, and variable modification of methionine oxidation. Search results were filtered to 99% protein probability and 95% peptide probability in Scaffold (v3.3.1, Proteome Software), producing a false discovery rate of 1%. Proteins were quantified using customized software developed in-house call iQuant. A complete Galaxy-based history for data analysis here: <https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold> A complete Galaxy-based workflow associated with this history: <https://galaxyp.msi.umn.edu/u/pjagtap/h/itraq-search-yang-2-xtandem-scaffold>

[Close](#)

**Contact**  
Tim Griffin, University of Minnesota

[View in PRIDE Inspector](#)  
[Download Project Files](#)

<b>Species</b> <a href="#">Homo sapiens (Human)</a>	<b>Tissue</b> <a href="#">oral epithelium</a>
<b>Cell Type</b> <a href="#">epithelial cell</a>	<b>Disease</b> <a href="#">oral squamous cell carcinoma</a>
<b>Instrument</b> <a href="#">LTQ Orbitrap Velos</a>	<b>Software</b> <a href="#">Sequest 27, rev. 12</a> <a href="#">Scaffold</a> <a href="#">Scaffold_4.3.2</a>
<b>Modification</b> <a href="#">Carbamidomethyl</a> <a href="#">Oxidation</a>	<b>Quantification</b> <a href="#">iTRAQ</a>
<b>Experiment Type</b> <a href="#">Shotgun proteomics</a>	<b>Assay count</b> <a href="#">1</a>



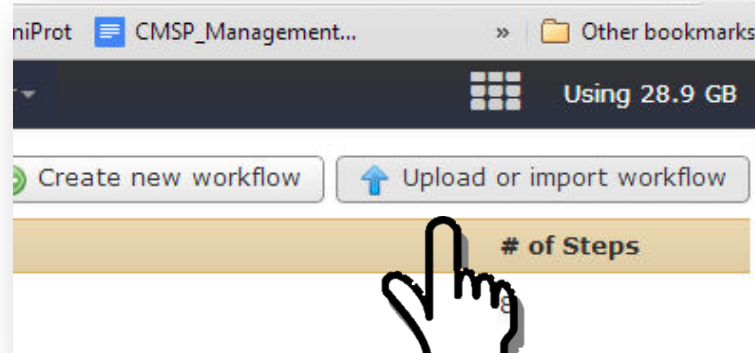
.ga file  
Raw MS data



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>



# Re-analysis of data: importing workflow



Galaxy / Galaxy-P

Analyze Data Workflow Shared Data Visualization Help User Using 28.9 GB

Workflow 'Workflow for 4-plex iTRAQ X!tandem Search Scaffold Processing (imported from uploaded file)' deleted

### Your workflows

Name	# of Steps
Workflow for 4-plex iTRAQ X!tandem Search Scaffold Processing (imported from uploaded file)	8
imported: imported: Workflow for Yang Replicate One 4-plex metaproteomics - Microbial Pro	37
imported: Workflow for Yang Replicate One 4-plex metaproteomics - Microbial Pro	37
Testing	49
imported: PARTIAL WORKFLOW: Workflow for Yang Replicate Two 4-plex metaproteomics - Microbial Pro	29
imported: Workflow for Yang Replicate Two 4-plex metaproteomics - Microbial Pro	37
imported: Workflow for Yang Replicate One 4-plex metaproteomics - Microbial Pro	37

Workflow for 4-plex iTRAQ X!tandem Search Scaffold Processing (imported from uploaded file) context menu:

- Edit
- Run
- Share or Publish
- Download or Export
- Copy
- Rename
- View
- Delete

### Workflows shared with you by others

Name	Owner	# of Steps
1 X!tandem Merge Workflow (InterProphet)	pjagtap@msi.umn.edu	15
X!tandem Merge (InterProphet)	pjagtap@msi.umn.edu	13

### Other options

Configure your workflow menu



### Running workflow "Workflow for 4-plex iTRAQ X!tandem Search Scaffold Processing (imported from uploaded file)"

Step 1: Protein Database Downloader (version 0.2.0)

Step 2: Input dataset

Multiple RAW Files

16: Multiple File Dataset for data 2, data 1, and others

type to filter

Step 3: Create Decoy Database (reverse) (version 0.1.0)

Step 4: msconvert3\_raw (version 0.2.0)

Step 5: X!Tandem MSMS Search (version 1.0.1)

Step 6: Scaffold (version 0.1.0)

Step 7: Scaffold Export (version 0.1.0)

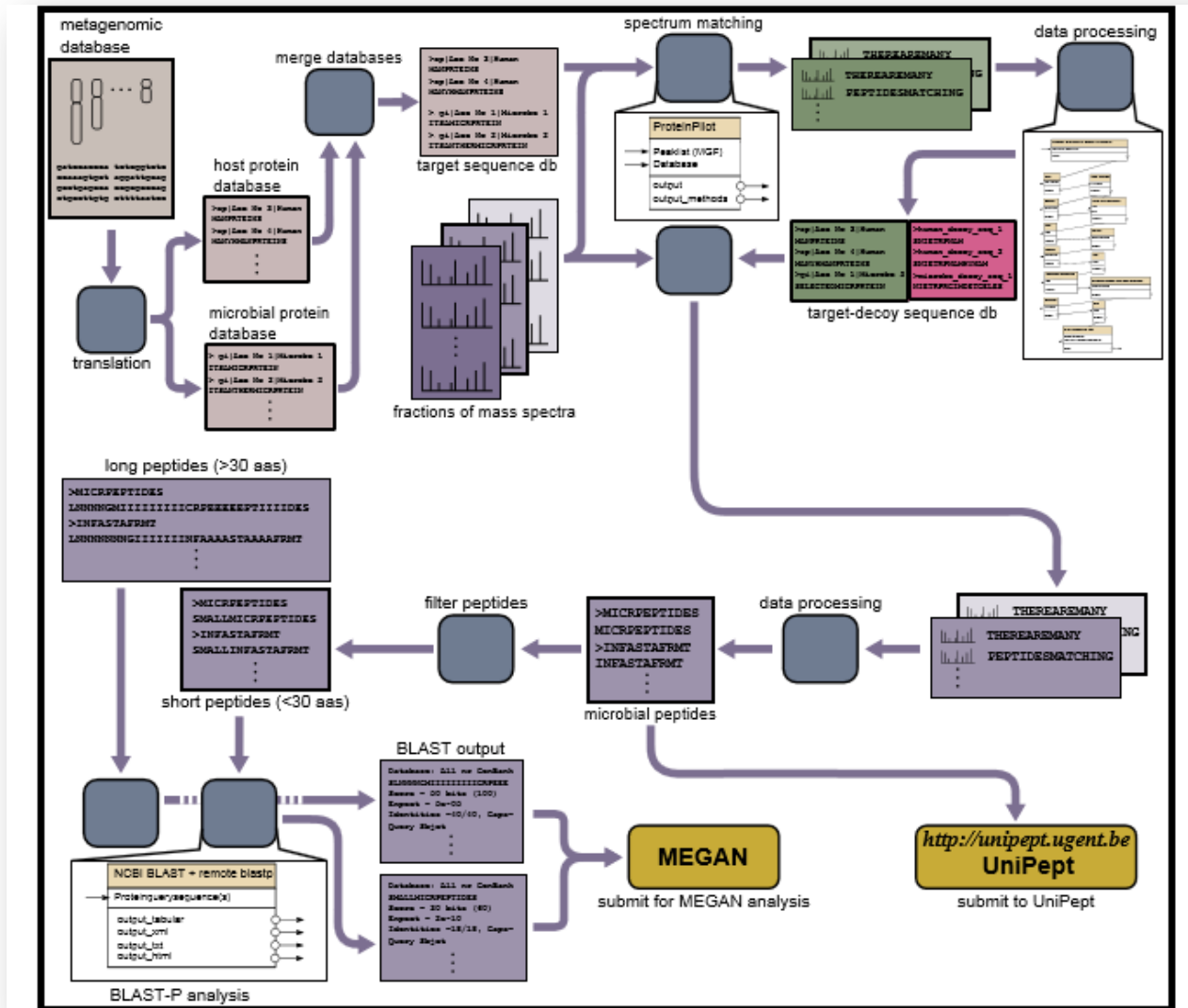
Step 8: Scaffold Export (version 0.1.0)

☐ Send results to a new history

Run workflow



# More complicated workflows: metaproteomics, proteogenomics



# Automating submission through Galaxy

Desktop Galaxy  
API App

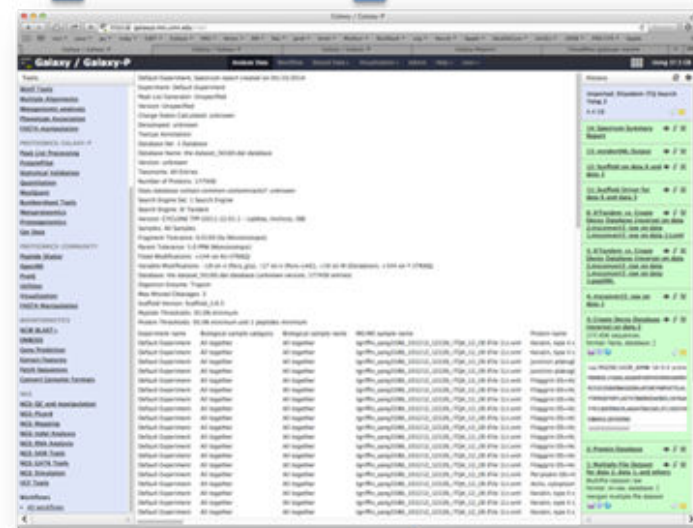
1) Get Submission  
File information via  
Galaxy API

2) Add Meta-  
data to create  
Submission  
Summary File

3) Create ProteomeXchange submission job on  
Galaxy with generated Submission Summary file

4) Galaxy Job Submits to ProteomeXchange  
(raw and processed data, meta-data,  
workflow URL and .ga file  
uploaded directly from galaxy server)

Galaxy Server



UNIVERSITY OF MINNESOTA  
Driven to Discover<sup>SM</sup>

# Future work and possibilities

- Modification of ProteomeExchange to communicate with Galaxy API
- Deployment of existing tools in Galaxy for ProteomeExchange submission (e.g. PeptideShaker tools)
- Automated data retrieval – re-analysis and mining of public data for new discoveries
- Bring the tools to the data: Galaxy cloud instance residing in the repository (e.g. Chorus)



# Galaxy at ASMS

Day	Time	Location	Presentation
Monday	9:10am - 9:30am	Ballroom III	<b>Novel Galaxy Workflows Combining RNA-seq and Proteomic MS/MS Reveal New Insights into Non-Model Organisms</b> <i>Conrad Bessant, Queen Mary University of London</i>
	10:30am - 1:00pm	Exhibit Hall C-G	<b>MP 033: Community-based Development and Evaluation of Biological Mass Spectrometry Software via the Galaxy Tool Shed</b> <i>Bart Gottschalk, Minnesota Supercomputing Institute</i>
			<b>MP 049: Characterizing molecular mechanisms of mammalian hibernation via non-model organism quantitative proteogenomics</b> <i>Katie Vermillion, University of Minnesota-Duluth, Duluth, MN</i>
			<b>MP 429: Large-Scale Quantitative Proteomic/Metaproteomic Platform Discovers Target Pathways and Promising Biomarkers of COPD-associated Lung Cancer</b> <i>Brian Sandri, University of Minnesota, Minneapolis, MN</i>
	4:10pm - 4:30pm	Ballroom III	<b>Public sharing of complex MS-based qualitative and quantitative proteomic data analysis workflows: adding value to big data repositories</b> <i>Tim Griffin, University of Minnesota</i>
Tuesday	10:30am - 1:00pm	Exhibit Hall C-G	<b>TP 077: Identifying Novel Peptide Sequence Variants from High Throughput RNA-Seq Data Via Flexible Proteomic Database Generation using the Galaxy Framework</b> <i>James Johnson, Minnesota Supercomputing Institute</i>
	12:00pm - 2:30pm		<b>TP 078: Towards a Novel Unprecedentedly Comprehensive Protein Identification Strategy, Mass Spectrometry and Ribosome Profiling: The Perfect Match</b> <i>Gerben Menschaert, Ghent University, Ghent, Belgium</i>
Wednesday	5:45pm - 7:00pm	Room 339-340	<b>Workshop 7: The Galaxy Framework for Biological MS Informatics: Practical Tips for Software Developers and Users</b> <i>Tim Griffin (presiding), University of Minnesota, Minneapolis, MN</i> <i>See below for more information</i>
Thursday	12:00pm - 2:30pm	Exhibit Hall C-G	<del><b>TP 044: Flexible, Accessible and Reproducible Workflows for Tandem Proteogenomic and Metaproteomic Analysis using the Galaxy-P Platform</b></del> <i>Pratik Jagtap, Center for Mass Spectrometry, St. Paul, MN</i>



UNIVERSITY OF MINNESOTA  
**Driven to Discover<sup>SM</sup>**