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

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℠

UNIVERSITY OF MINNESOTA
SUPERCOMPUTING INSTITUTE

proteome CHANGE

Biochemistry, Molecular Biology and Biophysics
Dr. Julie Yang
Dr. Ebbing de Jong
Dr. Joel Kooren

Dr. Yue Chen

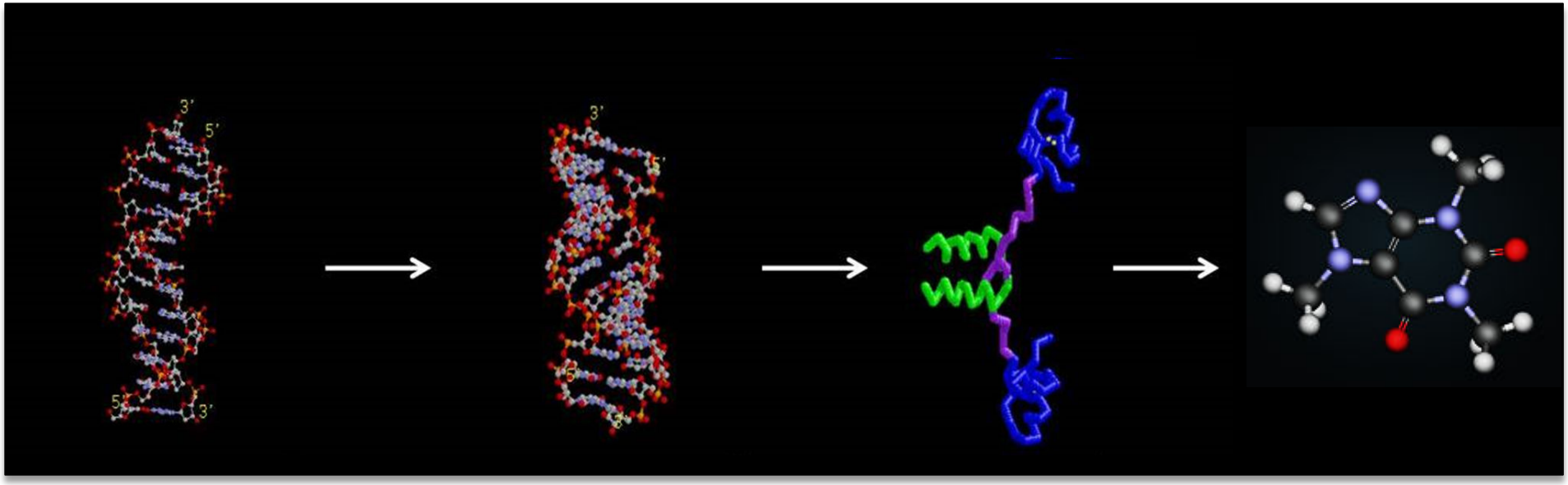
Center for Mass Spectrometry and Proteomics
Dr. Pratik Jagtap
Dr. LeeAnn Higgins

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

The era of "Big Data" in the biological sciences



DNA
Genome

RNA
Transcriptome

Protein
Proteome

Metabolite
Metabolome

High-throughput sequencing

High resolution mass spectrometry

The era of "Big Data" in the biological sciences

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

Nature 2013 **498**:255-60

…..and opportunities:

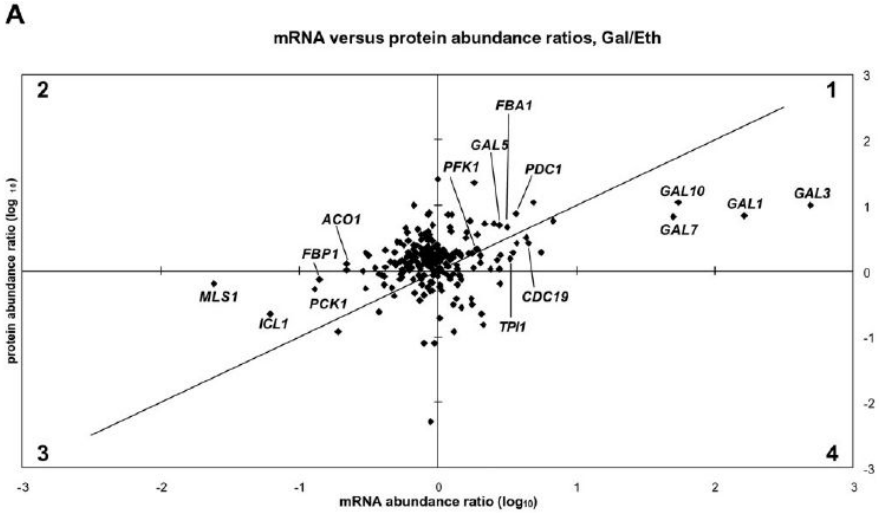
- Promotes reproducibility
- Data mining for new discoveries
- Creation of data resources (spectral libraries, etc)
- Re-analysis using new tools
  - evaluation and testing of new software
  - new results

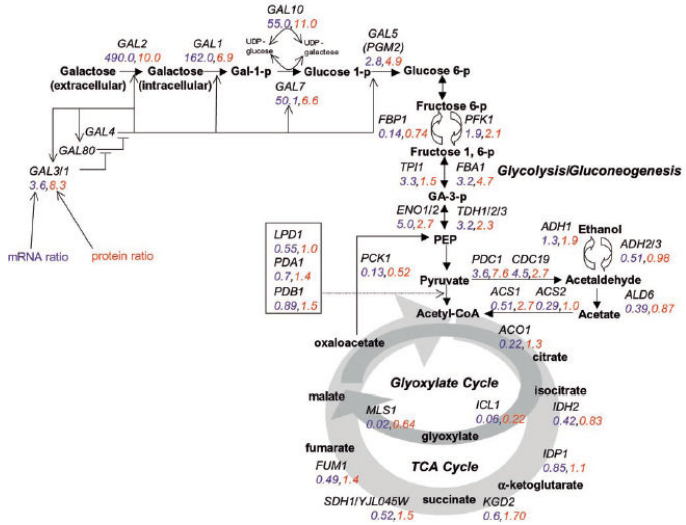
A historical anecdote in quantitative proteomics

(or a confession of past sins)

Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in *Saccharomyces cerevisiae**

Timothy J. Griffin‡, Steven P. Gygi§, Trey Ideker¶, Beate Rist‖, Jimmy Eng, Leroy Hood, and Ruedi Aebersold**





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

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

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?

Back to 2014: Big Data in MS-based proteomics

proteome CHANGE

MassIVE

Mass Spectrometry
Interactive **V**irtual **E**nvironment

Chorus

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

Enhancing Big Data Repositories: sharing the whole story

Galaxy

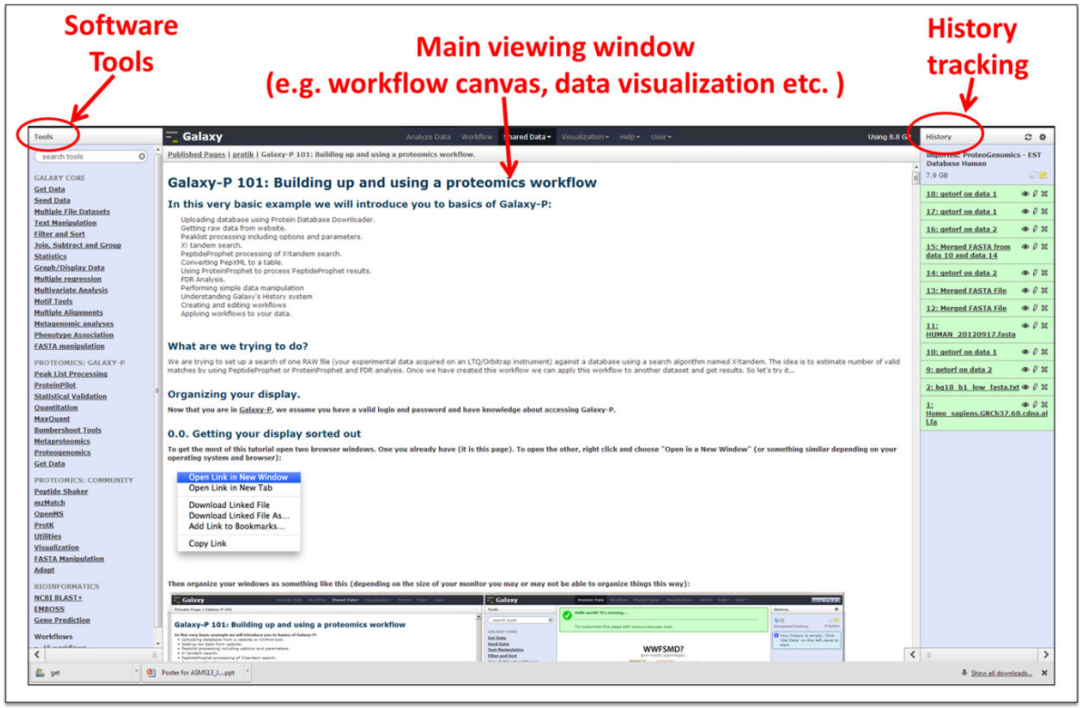
- A web-based, community developed bioinformatics framework/platform/workbench

- 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

usegalaxyp.org
(*in development*)

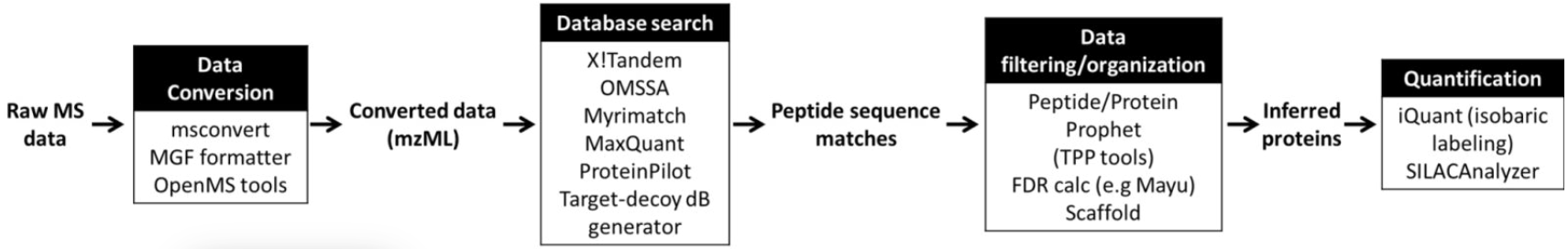
Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. ***Genome Biol.*** 2010, **11**: R86.

A (free) supermarket for ‘omics software?

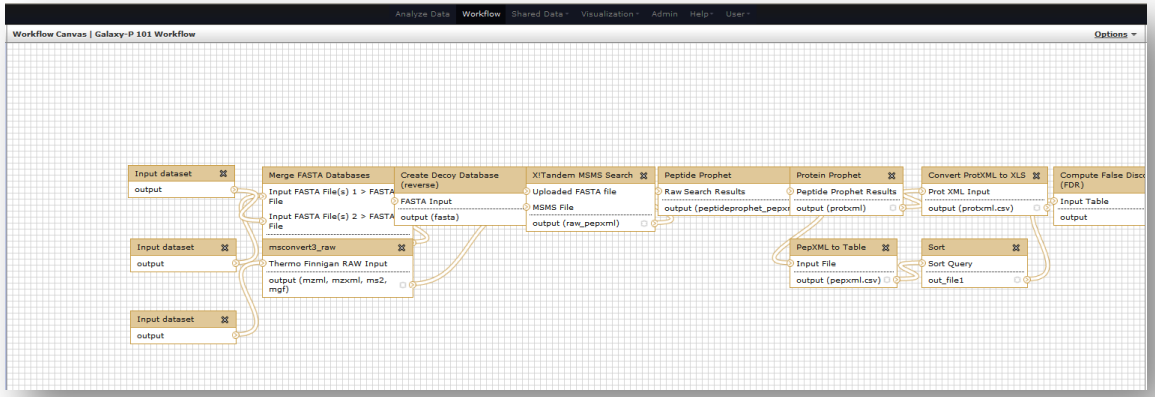


- Any command-line software can be deployed
- Amenable to Windows software (LWR)
- Multiple-file compatability

Capturing complete MS-based proteomic workflows

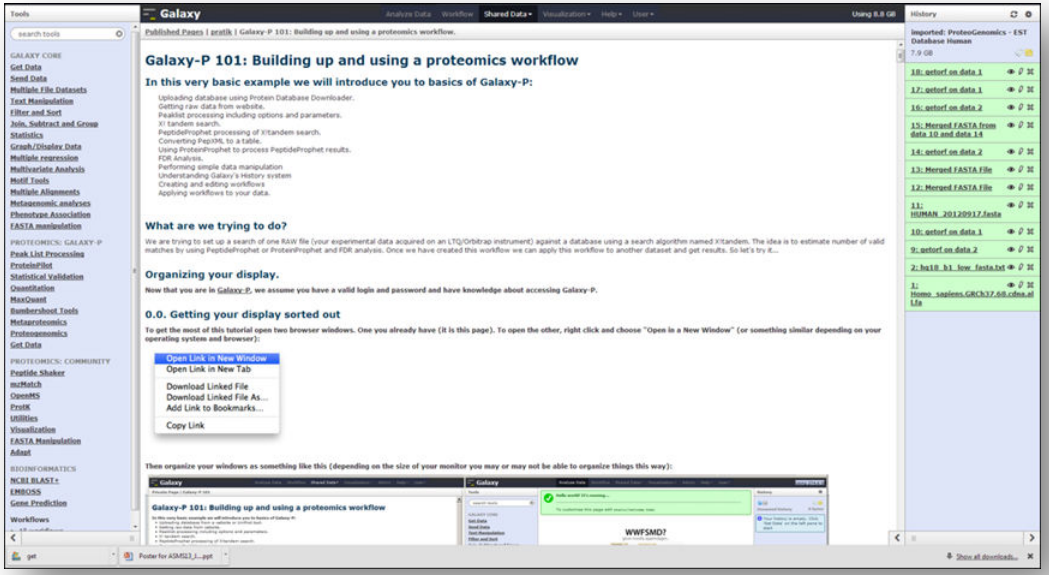


Galaxy history



Galaxy workflow

Exporting complete and reproducible workflows



URL export

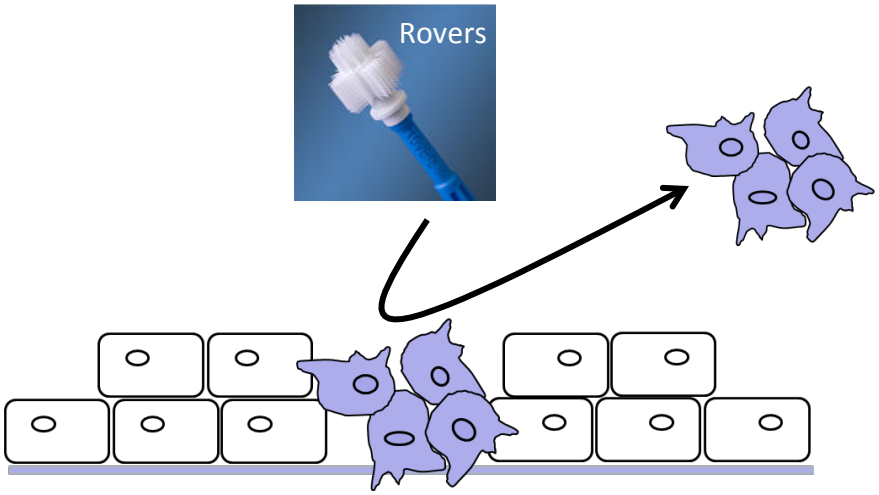
File export (.ga)

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

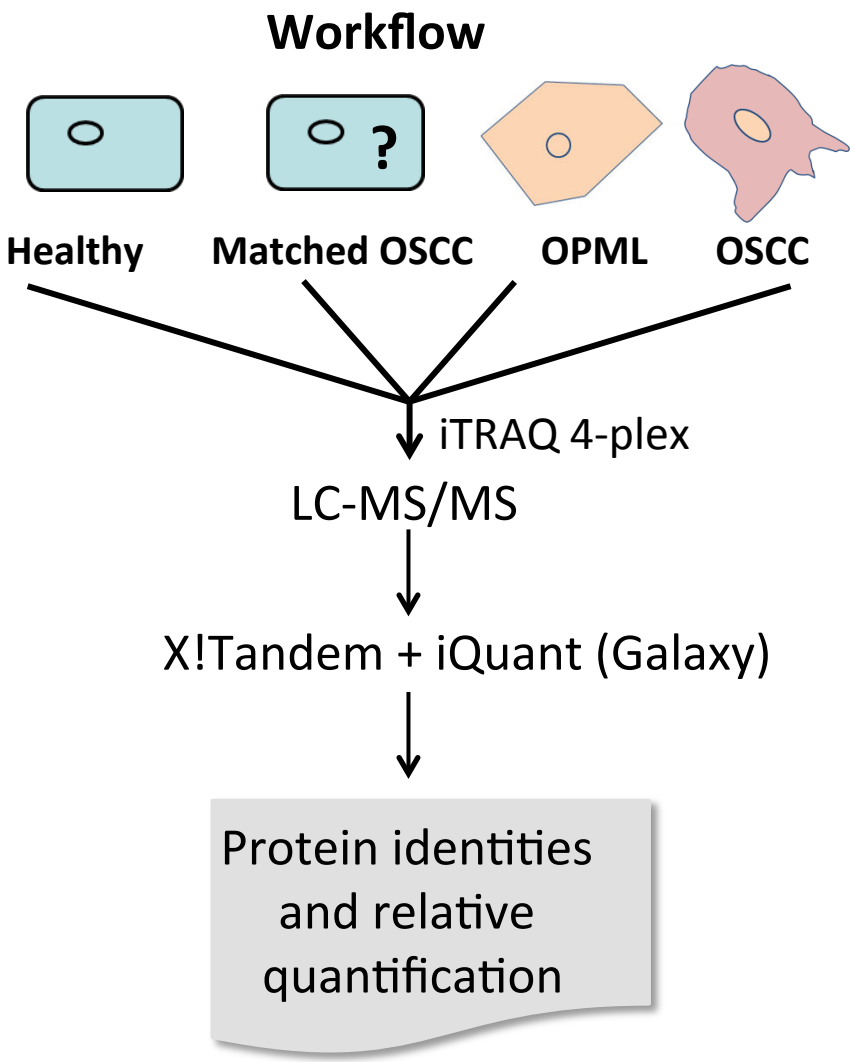
WORKFLOW: https://galaxyp.msi.umn.edu/u/pjagtap/w/workflow-for-4-plex-itraq-xtandem-search-scaffold-processing

Galaxy-Workflow-Workflow_for_4-plex_iTRAQ_X_tandem_Search_Scaffold_Processing.**ga**

Example: quantitative proteomic analysis of oral cancer



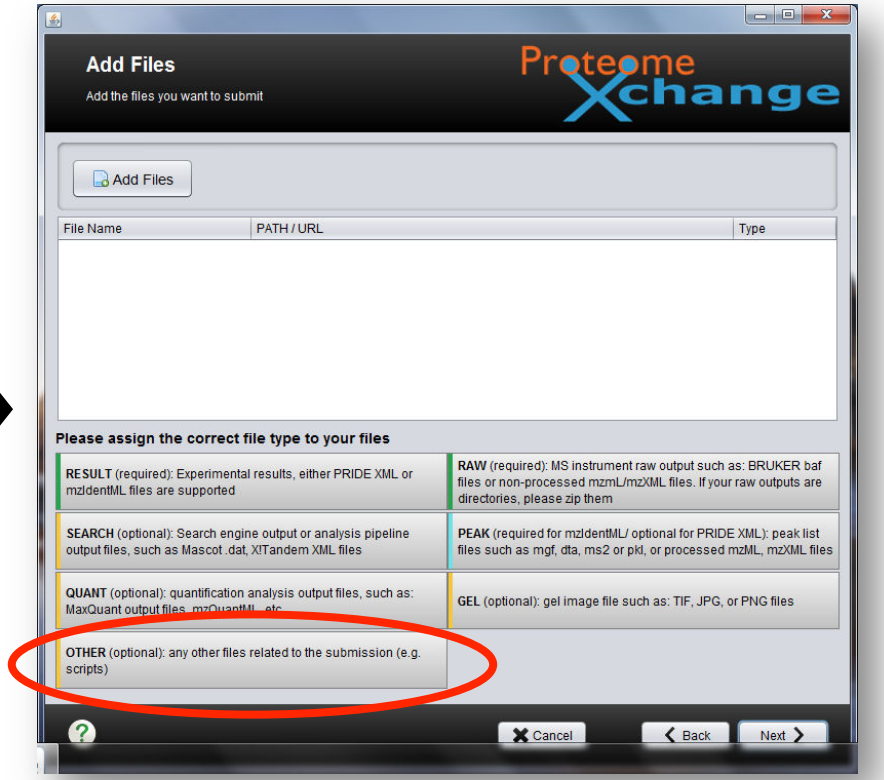
Lesion
(e.g. OPML or OSCC)



Dr. Julie Yang

Can it be reproduced?

Submission to repository: ProteomeExchange



ProteomeExchange Submission: Enhanced-Value

Project : PXD001044

View in PRIDE Inspector

Download Project Files

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

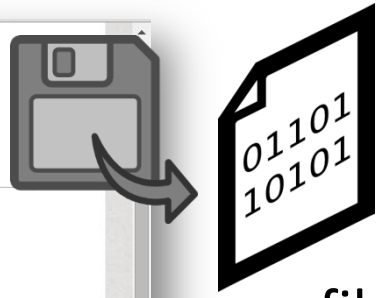
Raw files were converted to mzXml using msconvert (distributed as part of ProteoWizard 1.6.1260). MS/MS spectra were searched against the Uniprot human database including scrambled sequences and common contaminant proteins (a total of 136,002 entries) using X!Tandem (CYCLONE release, 2013.2.01). Search 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

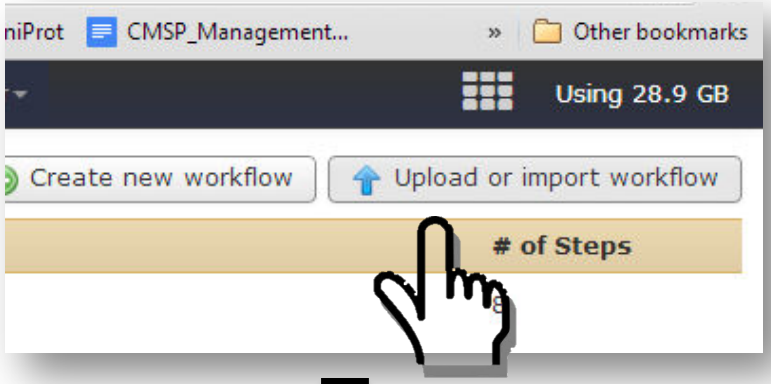
Species	Tissue
Homo sapiens (Human)	oral epithelium
Cell Type	**Disease**
epithelial cell	oral squamous cell carcinoma
Instrument	**Software**
LTQ Orbitrap Velos	Sequest 27, rev. 12 Scaffold Scaffold_4.3.2
Modification	**Quantification**
Carbamidomethyl Oxidation	iTRAQ
Experiment Type	**Assay count**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

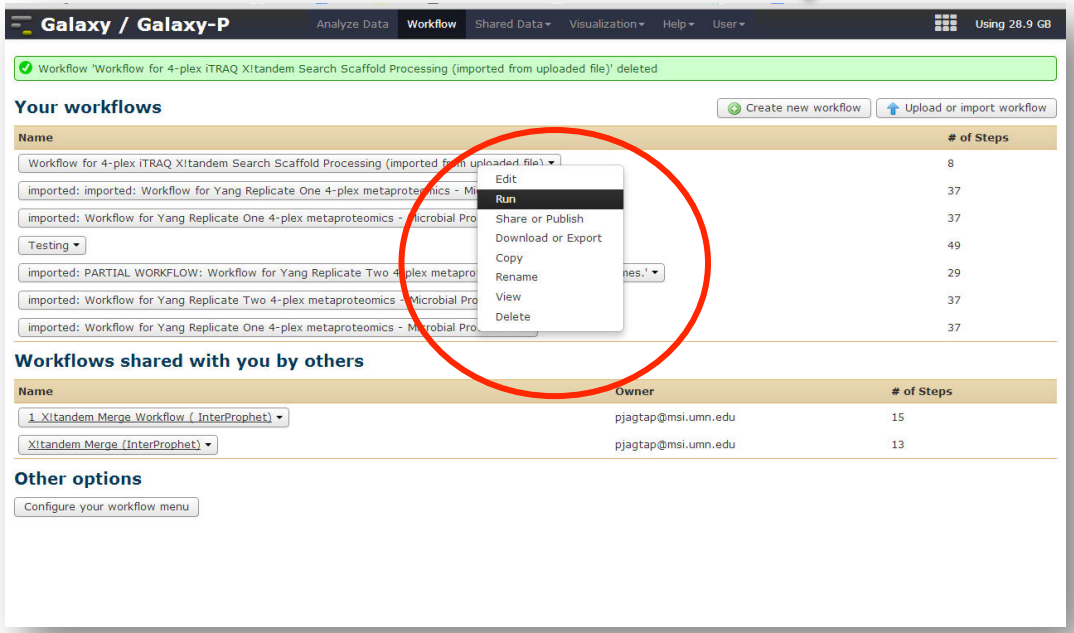


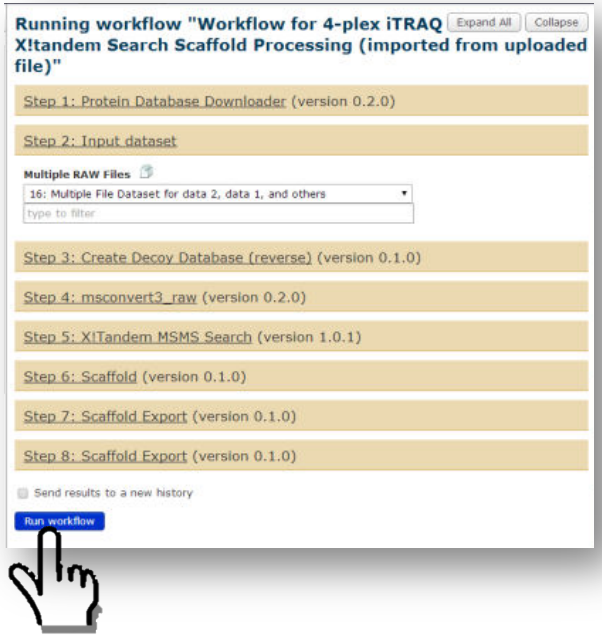
.ga file

Raw MS data

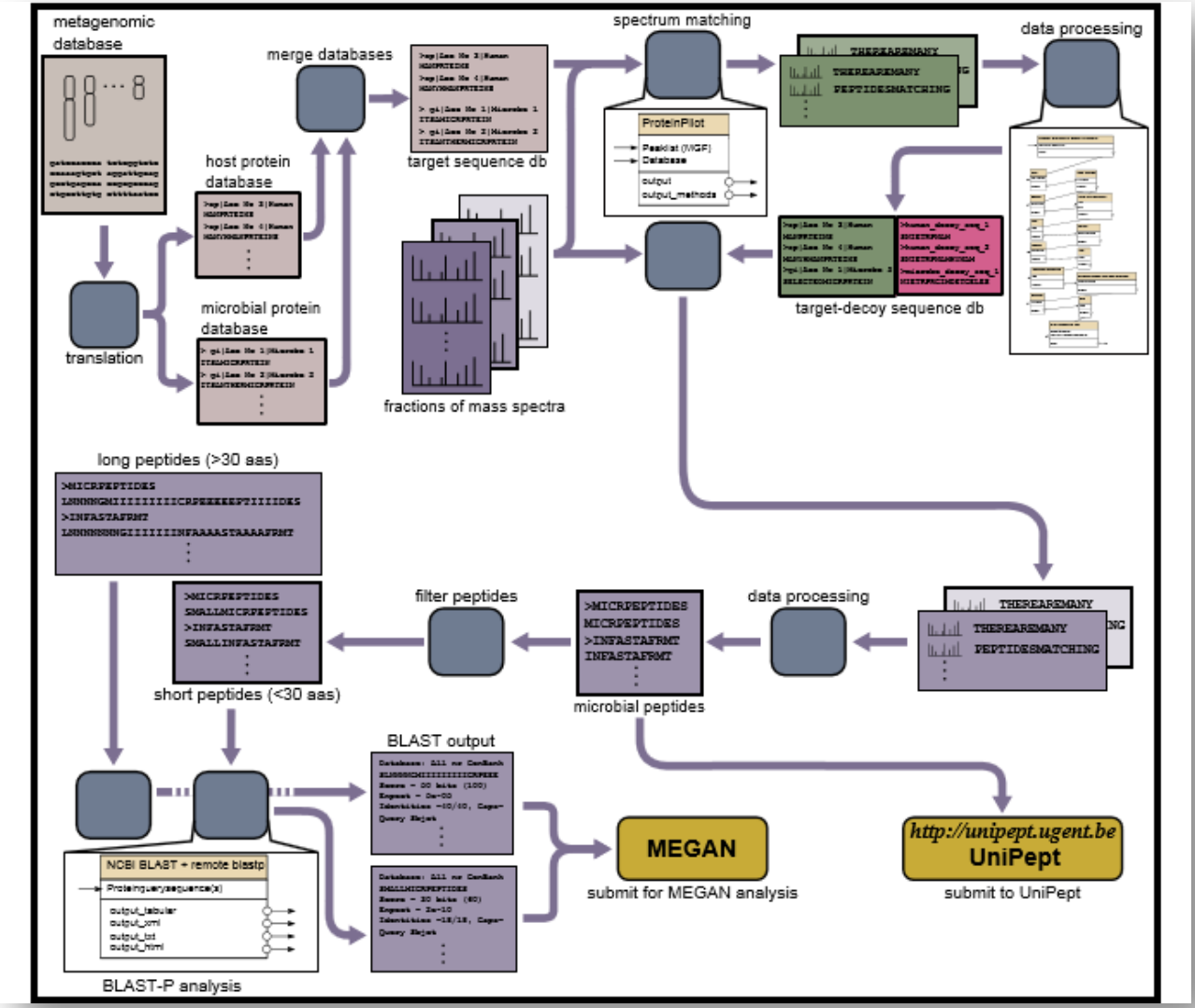
Re-analysis of data: importing workflow



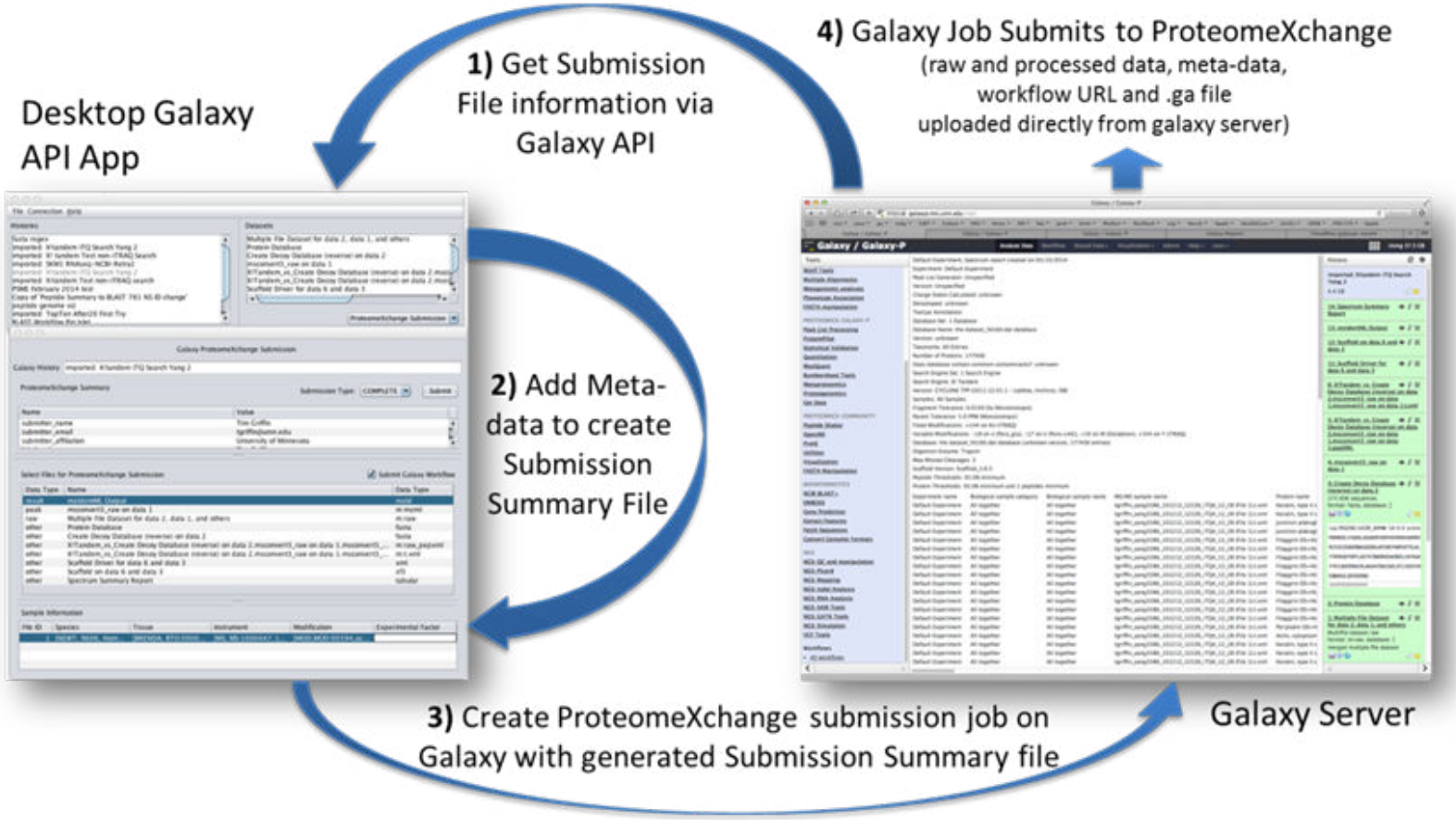




More complicated workflows: metaproteomics, protegenomics



Automating submission through Galaxy



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