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
• 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
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

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

Data reproducibility?

**MS-based proteomics**

The obtained MS/MS spectra were automatically searched against a database 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 context information was used for scoring. 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-

---


- Raw and processed data accessibility?
- Analytical reproducibility?
Back to 2014: Big Data in MS-based proteomics

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

usegalaxyp.org
(in development)

A (free) supermarket for ‘omics software?

- Any command-line software can be deployed
- Amenable to Windows software (LWR)
- Multiple-file compatibility
Capturing complete MS-based proteomic workflows

Galaxy workflow

Galaxy history
Exporting complete and reproducible workflows

**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

Workflow:
- Healthy
- Matched OSCC
- OPML
- OSCC

Lesion (e.g. OPML or OSCC)

Rovers

Protein identities and relative quantification

X!Tandem + iQuant (Galaxy)

LC-MS/MS

iTRAQ 4-plex
Can it be reproduced?
Submission to repository: ProteomeExchange
Project: PXD001044

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 were 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 maxXml using misconvert (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 scripts and developed in-house cell quant. A complete Galaxy-based history for data analysis here: https://galaxyp.msl.umn.edu/u/pj Jagtap/h/itraq-search-yang-2-tandem-scaffold A complete Galaxy-based workflow associated with this history: https://galaxyp.msl.umn.edu/u/pj Jagtap/h/itraq-search-yang-2- tandem-scaffold

Species
Homo sapiens (Human)

Cell Type
epithelial cell

Tissue
oral epithelium

Disease
oral squamous cell carcinoma

Instrument
LTQ Orbitrap Velos

Software
Sequest, rev. 12

Modification
Carbamidomethyl Oxidation

Quantification
iTRAQ

Experiment Type
Shotgun proteomics

Assay count
1

Contact
Tim Griffin, University of Minnesota
Re-analysis of data: importing workflow

Create new workflow

Upload or import workflow

Running workflow "Workflow for 4-plex iTRAQ XItandem Search Scaffold Processing (Imported from uploaded file)"

Step 1: Protein Database Downloader (version 0.2.0)
Step 2: Input detail
Step 3: Create Decay Database [reverse] (version 0.1.0)
Step 4: mconvert1.1, raw (version 0.2.0)
Step 5: XItandem 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
More complicated workflows: metaproteomics, protegenomics

- metagenomic database
- host protein database
- microbial protein database
- target sequence db
- target-decoy sequence db
- data processing
- spectrum matching
- BLAST output
- MEGAN
- UniPept
- BLAST-P analysis
- filter peptides
- NCB BLAST + remote blast
- fractions of mass spectra
- long peptides (>30 ass)
- short peptides (<30 ass)
Automating submission through Galaxy

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)
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

<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Location</th>
<th>Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>MP 033: Community-based Development and Evaluation of Biological Mass Spectrometry Software via the Galaxy Tool Shed</strong>&lt;br&gt;<strong>Bart Gottschalk, Minnesota Supercomputing Institute</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>MP 049: Characterizing molecular mechanisms of mammalian hibernation via non-model organism quantitative proteogenomics</strong>&lt;br&gt;<strong>Katie Vermillion, University of Minnesota-Duluth, Duluth, MN</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>MP 129: Large-Scale Quantitative Proteomic/Metaproteomic Platform Discovers Target Pathways and Promising Biomarkers of COPD-associated Lung Cancer</strong>&lt;br&gt;<strong>Brian Sandri, University of Minnesota, Minneapolis, MN</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Public sharing of complex MS-based qualitative and quantitative proteomic data analysis workflows: adding value to big data repositories</strong>&lt;br&gt;<strong>Tim Griffin, University of Minnesota</strong></td>
</tr>
<tr>
<td>Monday</td>
<td>10:30am - 1:00pm</td>
<td>Exhibit Hall C-G</td>
<td><strong>TP 077: Identifying Novel Peptide Sequence Variants from High Throughput RNA-Seq Data Via Flexible Proteomic Database Generation using the Galaxy Framework</strong>&lt;br&gt;<strong>James Johnson, Minnesota Supercomputing Institute</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>TP 078: Towards a Novel Unprecedentedly Comprehensive Protein Identification Strategy, Mass Spectrometry and Ribosome Profiling: The Perfect Match</strong>&lt;br&gt;<strong>Gerben Menschaert, Ghent University, Ghent, Belgium</strong></td>
</tr>
<tr>
<td></td>
<td>5:45pm - 7:00pm</td>
<td>Room 339-340</td>
<td><strong>Workshop 7: The Galaxy Framework for Biological MS Informatics: Practical Tips for Software Developers and Users</strong>&lt;br&gt;<strong>Tim Griffin (presiding), University of Minnesota, Minneapolis, MN</strong>&lt;br&gt;See below for more information</td>
</tr>
<tr>
<td></td>
<td>12:00pm - 2:30pm</td>
<td>Exhibit Hall C-G</td>
<td><strong>Metaproteomic Analysis using the Galaxy-P Platform</strong>&lt;br&gt;<strong>Pratik Jagtap, Center for Mass Spectrometry, St. Paul, MN</strong></td>
</tr>
</tbody>
</table>