THE GALAXY FRAMEWORK AS A UNIFYING BIOINFORMATICS SOLUTION FOR ‘OMICS’ CORE FACILITIES.

Pratik Jagtap¹; James Johnson²; Bart Gottchalk²; Getiria Onsongo²; Sricharan Bandhakavi³; Joel Kooren⁴; Brian Sandri⁴; Ebbing de Jong¹; Todd Markowski¹; LeeAnn Higgins¹; Chris Wendt⁴; Joel Rudney⁴ and Timothy Griffin⁴

Majority of proteomics research is focused on matching spectral data with proteins from the organism under study.

However, the dataset might have a lot more information apart from proteins from the organism.

Identification of proteins from other organisms (e.g. bacteria) – Metaproteomics.

Identification of peptides from the unannotated genomic region of the organism – Proteogenomics.
CHALLENGES IN IDENTIFYING ADDITIONAL PROTEINS IN THE DATASET

• Appropriate databases – bacterial protein database for Metaproteomics and genomic DNA / cDNA / RNASeq-derived database for Proteogenomics.

• Accessible domain-specific software for data processing and analysis.
OUR APPROACH

- Convert these genomic databases into protein databases by using appropriate software.

- Use Galaxy – a web-based bioinformatics data analysis platform – popular in genomics field. We developed, tested and deployed proteomics tools into Galaxy platform (Galaxy-P) to address this multi-omic analysis.
BENEFITS OF USING GALAXY

**Galaxy Tools**

Galaxy-P has multiple software tools from the genomics Galaxy framework – to which we added proteomics-specific tools.

**Galaxy Workflow**

Tools can be used in a sequential manner to generate workflows that can be reused, shared and creatively modified for multiple studies.

- **Software accessibility and usability.**
- **Share-ability of tools, workflows and histories.**
- **Reproducibility and ability to test and compare results after using multiple parameters.**
- **Analytical transparency.**
- **Scalability of data.**
METHODS AND DATASETS

- RAW files from Orbitrap Velos instrument.
- Processed peak lists were searched using ProteinPilot™ version 4.5 (AB Sciex) within Galaxy-P.
- After optimization and testing, multiple workflows were used in a sequential manner.

METAPROTEOMICS

- **Severe Early Childhood Caries (SECC)** dataset was acquired for clinical comparison of oral microcosm biofilms grown from plaque either in presence or absence of sucrose.
- Salivary supernatant that was 3D-fractionated with or without ProteoMiner treatment (Bandhakavi et al 2009) was used.
- Both the datasets were searched against the human oral microbiome database (HOMD) using the two-step method.

PROTEOGENOMICS

- Salivary supernatant (same as above).
- Oral pre-malignant lesion (OPML) dataset was collected as brush biopsy sample from six individuals with pre-malignant lesions and a matched control sample from adjacent oral cavity (Kooren et al unpublished).
- Both the datasets were searched against 3-frame translated Ensembl cDNA database by using two-step method.
OVERVIEW OF MODULES AND ANALYTICAL WORKFLOWS FOR METAPROTEOMIC ANALYSIS.

A. Peak list processing
- Mass spectra
- mconvert
- Peak list (MGF or mzml)
- Peak list processing

B. Database generation
- Metagenomic Database
- Translation
- Microbial protein db
- Merge FASTA
- Database generation

C. Two-step Database Search method
- Peptide Summary
- Target-Decoy database
- Target Database

D. Identifying peptides from microbial db
- Peptides in FASTA format
- Microbial Peptides
- Data Processing

E. BLAST-P Analysis
- BLAST output
- NCBi BLAST + remote blastp
- Proteinquerysequence(s)
- Output tabelular
- Output xml
- Output html
- BLAST-P Analysis

Submit for MEGAN Analysis

Submit to UniPept for analysis

http://unipept.ugent.be
SHAREABLE WORKFLOWS → RESULTS

Shareable workflows:
A. z.umn.edu/peaklistconversion
B. z.umn.edu/dbgenmp
C. z.umn.edu/mn2stp
D. z.umn.edu/pepfrommicrobialdb
E. z.umn.edu/blastp

All together: z.umn.edu/mp65

Results Summary

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total spectra</th>
<th>Distinct peptides of microbial origin</th>
<th>Phyla*</th>
<th>Genera*</th>
<th>Species*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole human salivary supernatant</td>
<td>988,974</td>
<td>1926</td>
<td>12</td>
<td>65</td>
<td>123</td>
</tr>
<tr>
<td>SECC without sucrose</td>
<td>153,019</td>
<td>28,126</td>
<td>5</td>
<td>33</td>
<td>56</td>
</tr>
<tr>
<td>SECC with sucrose</td>
<td>139,759</td>
<td>23,029</td>
<td>5</td>
<td>13</td>
<td>33</td>
</tr>
</tbody>
</table>

*Analysis using MEGAN.

- 20 KEGG pathways.
- Most prevalent pathway: Carbohydrate metabolism.
- ‘Best-populated’ pathway: Glycolysis (Carbohydrate metabolism).
- Protein with highest number of reads: Glyceraldehyde-3-phosphate.
OVERVIEW OF MODULES AND ANALYTICAL WORKFLOWS FOR PROTEOGENOMIC ANALYSIS.
SHAREABLE WORKFLOWS → RESULTS

Shareable workflows: A z.umn.edu/peaklistconversion B z.umn.edu/dbgen C z.umn.edu/mn2stp D z.umn.edu/peptidesfromcdnadb E z.umn.edu/blastp F z.umn.edu/psme G z.umn.edu/pep2gtf All together: z.umn.edu/pg140

Results Summary

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of spectra</th>
<th>Novel proteoform peptides</th>
<th>Novel proteoform peptides filtered after PSM evaluation</th>
<th>Number of distinct peptides after visualization and for genome localization.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salivary supernatant</td>
<td>988,974</td>
<td>254</td>
<td>105</td>
<td>52</td>
</tr>
<tr>
<td>OPML Control</td>
<td>156,405</td>
<td>904</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>OPML Lesion</td>
<td>157,299</td>
<td>887</td>
<td>29</td>
<td>21</td>
</tr>
</tbody>
</table>

Representation of genomic organization of identified novel proteoform-specific peptides from PRB1 and PRB2 genes on chromosome 12.
**QUANTITATIVE PROTEOMICS**

### RNASeq Data

- Single amino-acid polymorphism database
- Spliced junction database
- Reduced database (based on number of reads per million)

*In collaboration with Shenykman and Smith (University of Wisconsin)*

### Proteomics Data

- Galaxy-P Workflows*
  - iTRAQ quantitation dataset
  - Correlate RNASeq and proteomic quantitative datasets.

* Galaxy-P Workflows
**IMMEDIATE PLANS:**

- Improving on current metaproteomic and proteogenomic workflows.
- Installation and testing of open-source tools. The installation and testing is being carried out through and international collaboration between developers and users.
- Working along with UMN Genomics Center and research community and to offer as a combined service for these emerging areas in proteomics research.

**Tutorial:** [z.umn.edu/ppingp](http://z.umn.edu/ppingp)

**Acknowledgements:**

- Funded by NSF grant 1147079.
- Thanks to Gloria Shenykman and Lloyd Smith (UW) for their work on RNASeq db workflows.
- Many thanks to John Chilton (Penn State) for his work on this project.