ProteoGenomics in Galaxy:
Identifying novel ‘constellations’ of proteoforms using transcriptomic and proteomic data.

Galaxy Community Conference (GCC 2016)
Bloomington, IN
28th June 2016

PRATIK JAGTAP
Center for Mass Spectrometry and Proteomics, St. Paul, MN
Department of Biochemistry, Molecular Biology and Biophysics, Minneapolis, MN
Acknowledgements

**Biochemistry, Molecular Biology & Biophysics**
- TIM GRIFFIN
  - Candace Guerrero
  - Kevin Murray

**Biology & Biophysics**
- Harald Barsnes & Marc Vaudel
  - University of Bergen, Bergen, Norway

**Biologists / collaborators**
- Katie Vermillion
- Kyle Anderson
- Matt Andrews
  - Brett Noel, PhD
  - Tzu-Yi Yang, PhD
- Laurie Parker

**Minnesota Supercomputing Institute**
- Ben Lynch
- Trevor Wennblom
- Tom McGowan
- Getiria Onsongo
- James Johnson

**Center for Mass Spectrometry & Proteomics**
- LeeAnn Higgins
- Todd Markowski

**GalaxyP**
- John Chilton
  - Galaxy Team
  - Penn State University

**Funding**
- Ira Cooke
  - La Trobe University, Melbourne, Australia

**University of Minnesota**
ProteoGenomics in Galaxy

• PROTEOMICS / PROTEOFORMS / PROTEOGENOMICS

• GALAXY WORKFLOWS AS A SOLUTION

• EXAMPLE PROTEOGENOMICS PROJECTS

• WHAT’S NEXT?
PROTEOFORM: all of the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications. (Smith and Kelleher 2013 - Nat Methods. doi:10.1038/nmeth.2369)
“Our complex workflow (approximately 140 steps) can be easily shared using built-in Galaxy functions, enabling their use and customization by others. Our results provide a blueprint for the establishment of the Galaxy framework as an ideal solution for the emerging field of proteogenomics.”

doi: 10.1021/pr500812t
Galaxy Workflows

**Database generation**
- RNASeq → Protein database generation (Novel junctions, Single amino acid mutations, Reduced databases).

**Raw Data conversion**
- msconvert and MGF Formatter to convert RAW (Thermo) files.
- Sciex Converter and msconvert to convert wiff (Sciex) files.
- Jagtap et al 2014; doi: 10.1021/pr500812t

**Database Search**
- ProteinPilot; SearchGUI / PeptideShaker; Morpheus.
- Search strategies (two-step method, multi-step method, etc.)
  - Boekel et al 2015; doi: 10.1038/nbt.3134
  - Shortreed et al 2015; doi: 10.1021/acs.jproteome.5b00599

**Genome Visualization**
- Generate BED files so that they can be visualized using Integrated Genomics Viewer
  - Vermillion et al 2015; doi: 10.1021/acs.jproteome.5b00575
  - Anderson et al 2016; doi: 10.1021/acs.jproteome.5b01138

**Data Processing**
- BLAST-P Workflow

**PSM Evaluation**
- mzSQLite
  - PSM Evaluation Tool
  - doi: 10.1021/pr500812t
“Our customized Galaxy-based software includes automated, batch-mode BLASTP searching and a Peptide Sequence Match Evaluator tool, both useful for evaluating the veracity of putative novel peptide identifications.”


doi: 10.1021/pr500812t
MultiOmics Visualization Platform

Data Exploration and Validation
Starting from your Galaxy History

Filter on Galaxy Data Sets
- Access tabular data sets from history
- Select single or multiple sequences for data search

MVP Viewer is invoked from the existing Galaxy visualization framework

The MVP Viewer has API access back to the originating Galaxy history.
- User can send data back to their history
- User can access other datasets from their Galaxy history
- User will be able to save the entire MVP, including state, back to Galaxy
- User will be able to share a saved MVP with other Galaxy users

Artwork by Gavin Susantio

University of Minnesota
Figure 4. Representation of organization of identified peptides corresponding to a novel proteoform from PRB1 and PRB2 genes on chromosome 12. View is a zoomed-in screenshot of chromosome 12, which shows the orientation of expression, amino acid sequences within three frames of translation, reference files in the tracks, and amino acid sequence of the identified peptide corresponding to a novel proteoform. The red arrows indicate the direction and amino acid sequence (from amino-terminal to carboxy-terminal) of the identified peptides. A red asterisk indicates a stop codon in the normal coding frame. Block arrows in red indicate multiple distinct peptides identified during the proteogenomic analysis.
Tracing of core body temperature ($T_b$, black line) from a single animal measured by a surgically implanted transmitter, along with the controlled ambient temperature (blue line) over the course of the hibernation season. *TOR (Torpor), J-IBA (January IBA), M-IBA (March IBA)*

*J. Proteome Res., 2015, 14 (11), pp 4792–4804*
DOI: 10.1021/acs.jproteome.5b00575
**Proteogenomics Projects**

**Hibernation (Heart and Skeletal)**

**HEART MUSCLE:**

Novel peptide sequence analysis. (A) Venn diagram depicting the number of BLAST-P identified novel peptide sequences from each set of proteomic samples, including peptides that were found in more than one data set. There were 162 novel peptide sequences identified in all three replicates.

Vermillion *et al.* *J. Proteome Res.* 2015, 14, 4792-4804. DOI: 10.1021/acs.jproteome.5b00575

**SKELETAL MUSCLE:**

The transition out of hibernation to an active state (April) exhibits most notable changes in the skeletal muscle proteome.

- Relative to hibernation phase, proteins related to *glucose metabolism* increase sharply, while *fatty acid metabolism* proteins decline at the active state phase.
- Relative to hibernation phase, there is an increase in *fast-twitch associated isoform myosin* (which uses a more glycolytic mechanism) and a reduction in *slow-twitch isoforms*. Anderson *et al.* *J. Proteome Res.* 2016, 15, 1253-1261. DOI: 10.1021/acs.jproteome.5b01138
iTRAQ-labeled dataset with pre-pro-B cells and pro-B cells were analyzed by searching against RNASeq data derived protein databases.

a) 3-frame translated novel lncRNA db
b) 3-frame translated annotated lncRNA db.
c) 3-frame translated antisense lncRNA db.
d) 3-frame translated repetitive lncRNA db.
e) Novel junction database
f) 3-frame translated EnSEMBL cDNA db.

Heydarian et al 2014; doi: 10.4172/jpb.1000302
Ongoing Projects
MultiDrug Resistance (Parker Lab)

Human CML cell line (inhibitor sensitive)
TKI resistant lines: Grow in [drug] at levels that kill the sensitive line ~3 months

TKI sensitive line: untreated, maintained in culture for same amount of time

University of Minnesota
What is Next?

**Database generation**
- RNAsSeq → Protein database generation (Novel junctions, Single amino acid mutations, Reduced databases).

**Raw Data conversion**
- msconvert and MGF Formatter to convert RAW (Thermo) files.
- Sciex Converter and msconvert to convert wiff (Sciex) files.

**RNASeq Quantitation Output**
- RNAsSeq or EnSEMBL database generation for genome visualization.
- Replace HiSAT with TopHAT

**Database Search**
- ProteinPilot; SearchGUI / PeptideShaker; Morpheus.
- Search strategies (two-step method, multi-step method, etc.)

**Genome Visualization**
- Generate BED files so that they can be visualized using Integrated Genomics Viewer

**Genome-Proteome Quant Integration**
- Integration at quantification, localization, pathway analysis levels.
- Ability to parse out interesting peptide sequences for MRM analysis and submit to Skyline?
- Visualization of genome-proteome data. Ability to parse out interesting RNA sequences for RT-PCR analysis (Quant?)
Acknowledgements

Biochemistry, Molecular Biology & Biophysics
TIM GRIFFIN
Candace Guerrero
Kevin Murray

Biologists / collaborators
Katie Vermillion
Kyle Anderson
Matt Andrews
Brett Noel, PhD
Tzu-Yi Yang, PhD
Laurie Parker

Minnesota Supercomputing Institute
Ben Lynch
Trevor Wennblom
Tom McGowan
Getiria Onsongo
James Johnson

GalaxyP

Harald Barsnes & Marc Vaudel
University of Bergen,
Bergen, Norway

Lennart Martens
VIB, Ugent,
Belgium

John Chilton
Galaxy Team
Penn State University

Bjoern Gruening
University of Freiburg,
Freiburg, Germany

Mo Heydarian
Karen Reddy

INTERESTED IN GALAXY-P TEAM?
POST-DOC POSITION AVAILABLE

z.umn.edu/galaxypreferences

Funding

Ira Cooke
La Trobe University,
Melbourne, Australia

University of Minnesota