Tackling Analytical challenges in Cancer proteogenomics using Galaxy framework

December 12, 2018

Pratik Jagtap
Galaxy-P Team
University of Minnesota
PROTEOGENOMICS & ITS CHALLENGES

Challenges

• Large search database sizes
• False-positive sources and their elimination.
• Validation of novel peptide identification.
• PSM Quality Evaluation
• Targeted proteomics of identified peptides.
• Genomic localization.
• Disparate tools and numerous processing steps.

Ruggles et al. Mol Cell Proteomics 2017;16:959-981

© 2017 by The American Society for Biochemistry and Molecular Biology, Inc.
Galaxy Platform


- A web-based bioinformatics data analysis platform.
- **Software accessibility** and usability.
- **Share-ability** of tools, workflows and histories.
- **Reproducibility** and ability to test and compare results after using multiple parameters.
- Ability to assimilate disparate software into integrated workflows.
Software tools can be used in a sequential manner to generate analytical workflows that can be reused, shared and creatively modified.

For example, Protein Database Downloader downloads UniProt protein FASTA databases of various organisms.
An accessible proteogenomics informatics resource for cancer researchers

Instructions for accessing workflows and training material

This page provides links to instructional materials describing how to access and use the proteogenomic workflows described in the manuscript text.

Please access the links indicated below for the following materials:

Training on usegalaxy.org

- How to access and use the usegalaxy.org training site: http://z.umn.edu/accessusegalaxyorg
- Mouse Proteogenomics tutorial: http://z.umn.edu/pginnov18

Interactive Galaxy Tours

- Accessing an interactive Galaxy Tour describing the RNA-seq-to-protein-database workflow: http://z.umn.edu/rnaseqtoproteinodbtour
Proteogenomics Workflows In Galaxy

Workflow #1
RNA-Seq to Variant
FASTA database

Workflow #2
Database Searching
Using MS/MS data

Workflow #3
Identifying Novel Variants
And Visualization
WORKFLOW 1: RNA-SEQ TO FASTA DATABASE CREATION

- **Genome**
- **RNA-Seq FASTQ**
- **GTF**

**Assembly Workflow**

- STRINGTIE: RNA-Seq to transcripts
- GFF COMPARE: Evaluates the assembly with annotated transcripts
- Translate transcripts

**SAV / In-Del Variants**

- FREEBAYES: Variant Calling
- CustomPro DB: Variant annotation, Genome mapping
- Mapping Files, Sequence FASTA

**WORKFLOW 2: MAPPING TO FASTA DATABASE CREATION**

- **Genome**
- **RNA-Seq FASTQ**
- **GTF**

**Assembly Workflow**

- STRINGTIE: RNA-Seq to transcripts
- GFF COMPARE: Evaluates the assembly with annotated transcripts
- Translate transcripts

**SAV / In-Del Variants**

- FREEBAYES: Variant Calling
- CustomPro DB: Variant annotation, Genome mapping
- Mapping Files, Sequence FASTA
**WORKFLOW 1: OUTPUTS**

### FASTA Sequence File

```
>generic|ENSMUSP00000107433|Erp29|ER protein 29
MAAAAGVSAASLLPPLLVLGGLLLLFAPHHGGSLHTKGALPLDTVTFYKSRLLGP

>generic|ENSMUSP00000120715|Rps2|ribosomal protein S2
MADDAGAAGGPGPGLGLGGFRGGFGSGLRGRGRRGRRGRRGRRGARGGKAEDKEWIPVTKLGRLVMDKIKSLEEYLFSLPIKESEIDFFLGASLKDEVKLIMPVQKQTRAGQR
```

### Genomic Mapping File

<table>
<thead>
<tr>
<th>Ensembl</th>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Direction</th>
<th>Gap</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSMUSP00000107433</td>
<td>chr5</td>
<td>121452190</td>
<td>121452340</td>
<td>-</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td>ENSMUSP00000107433</td>
<td>chr5</td>
<td>121449139</td>
<td>121449163</td>
<td>-</td>
<td>150</td>
<td>174</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720275</td>
<td>24720452</td>
<td>+</td>
<td>0</td>
<td>177</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720533</td>
<td>24720731</td>
<td>+</td>
<td>177</td>
<td>375</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24720968</td>
<td>24721302</td>
<td>+</td>
<td>375</td>
<td>709</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721622</td>
<td>24721727</td>
<td>+</td>
<td>709</td>
<td>814</td>
</tr>
<tr>
<td>ENSMUSP00000120715</td>
<td>chr17</td>
<td>24721802</td>
<td>24721897</td>
<td>+</td>
<td>814</td>
<td>909</td>
</tr>
</tbody>
</table>
WORKFLOW 2: Database Searching Using MS/MS data

- **Protein FASTA:** reference proteins + potential variants
- **Peaklist of MS/MS data**
- **Multiple algorithms for matching MS/MS to peptides**
- **Organization and scoring of peptide spectral matches (PSMs)**
- **Generation of an SQLite database for downstream data visualization and filtering**
- **Putative variant peptide sequences for further verification and analysis**

**RNA-Seq database** → **SearchGUI**

**SearchGUI**

Proteomics. 11:996-9

**PEPTIDE SHAKER**

Nat Biotechnol. 33:22-4

**mgIdentml**

**Mz to sqlite**

**Peptides for BLASTP**

GalaxyP
WORKFLOW 3: Identifying Novel Variants & Visualization

Workshop Documentation: z.umn.edu/galaxypinmumbai

5.2 BlastP analysis 32
5.3 Novel proteoform analysis 33
5.4 Using Multi-omics Visualization Platform for visualizing novel proteoforms 35
MULTI-OMICS VISUALIZATION PLATFORM FOR VISUALIZING NOVEL PROTEOFORMS

SPECTRAL QUALITY VISUALIZATION (Lorikeet Viewer)

GENOMIC LOCALIZATION (Integrated Genomics Viewer)
An accessible proteogenomics informatics resource for cancer researchers

Instructions for accessing workflows and training material

This page provides links to instructional materials describing how to access and use the proteogenomic workflows described in the manuscript text.

Please access the links indicated below for the following materials:

Training on usegalaxy.org
- How to access and use the usegalaxy.org training site: http://z.umn.edu/accessusegalaxyorg
- Mouse Proteogenomics tutorial: http://z.umn.edu/pginnov18

Interactive Galaxy Tours
- Accessing an interactive Galaxy Tour describing the RNA-seq-to-protein-database workflow: http://z.umn.edu/rnaseqtoproteindbtour

Proteogenomics Gateway: z.umn.edu/proteogenomicsgateway
Step-by-step instructions for Galaxy instance usage: z.umn.edu/pginnov18
SELECT INPUT FILES & RUN WORKFLOWS
ACKNOWLEDGMENTS

Biologists / collaborators
Laurie Parker
Joel Rudney
Maneesh Bhargava
Amy Skubitz
Chris Wendt
Brian Croker
Steven Friedenberg
Kevin Viken
Kristin Boylan
Marnie Peterson
Somiah Afiuni
Brian Sandri
Alexa Pragman
Wanda Weber
Amy Treeful

Producers
Minnesota Supercomputing Institute
James Johnson
Thomas McGowan
Lee Parsons
Michael Milligan

GalaxyP (Co-I)
Bart Messure
Robbert G Singh
VIB, UGhent, Belgium

Lennart Martens (Co-I)

Carolin Kolmeder
University of Helsinki, Finland

Thilo Muth
Bernhard Renard
Robert Koch Institute

Alessandro Tanca
Porto Conte Ricerche, Italy

Anamika Krishanpal
Priyabrata Panigrahi
Persistent Systems Limited

Bjoern Gruening
Bérénice Batut
University of Freiburg, Freiburg, Germany

Ira Cooke
Melbourne, Australia

University of Minnesota
Timothy Griffin
Pratik Jagtap
Praveen Kumar
Candace Guerrero
Subina Mehta
Adrian Hegeman (Co-I)
Art Eschenlauer
Shane Hubler
Ray Sajulga
Caleb Easterly
Andrew Rajczewski

University of Bergen, Norway
Harald Barsnes
Marc Vaudel

University of Bergen, Norway

University of Freiburg, Freiburg, Germany

VIB, UGhent, Belgium

Judson Hervey
Naval Research Institute
Washington, D.C.

Karen Reddy
Mo Heydarian
Johns Hopkins University

Thomas Doak
Jeremy Fisher
Indiana University

Matt Chambers
Nashville, TN

Stephan Kang
Intero Life Sciences

Josh Elias
Stanford University

Josh Elias
Michael Shortreed
UW-Madison

Lloyd Smith (Co-I)

Brook Nunn
U of Washington

Funding

NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

Funding

Anamika Krishanpal
Priyabrata Panigrahi
Persistent Systems Limited

Judy Nunn
U of Washington

GalaxyP

Harald Barsnes
Marc Vaudel

University of Bergen, Norway

University of Bergen, Norway

University of Freiburg, Freiburg, Germany

VIB, UGhent, Belgium

Judson Hervey
Naval Research Institute
Washington, D.C.

Karen Reddy
Mo Heydarian
Johns Hopkins University

Thomas Doak
Jeremy Fisher
Indiana University

Matt Chambers
Nashville, TN

Stephan Kang
Intero Life Sciences

Josh Elias
Stanford University

Josh Elias
Michael Shortreed
UW-Madison

Lloyd Smith (Co-I)
QUESTIONS?

Visit: Poster P-085 at PSI 2018
http://galaxyp.org/contact

Proteogenomics Gateway: z.umn.edu/proteogenomicsgateway
Step-by-step instructions for Galaxy instance usage: z.umn.edu/pginnov18