A resource for complete proteogenomic informatics in cancer research: Built on ITCR tools

Timothy J. Griffin
University of Minnesota

ITCR annual meeting
Park City, Utah
May 29, 2019

(Pictures and tweets welcome)
Acknowledgements

**Communities**

**Biochemistry, Molecular Biology & Biophysics**

Dr. Pratik Jagtap (Co-leader, Galaxy-P)
Praveen Kumar
Subina Mehta
Caleb Easterly
Ray Sajulga
Andrew Rajczewski
Dr. Shane Hubler
Mark Esler
Dr. Art Eschenlauer
Dr. Candace Guerrero
Matt Chambers

**Community-Based Software Development**

Harald Barsnes and Marc Vaudel
*University of Bergen, Bergen, Norway*
Bjoern Gruening (Galaxy community...)
*University of Freiburg, Freiburg, Germany*
Lennart Martens
*VIB Department of Medical Protein Research, UGent, Belgium*
Lloyd Smith/Michael Shortreed
*University of Wisconsin-Madison*

**ITCR Groups**

Rachel Karchin/Michael Ryan
*Johns Hopkins University/In-Silico Solutions*

**Collaborators**

Adrian Hegeman
Mo Heydarian/Karen Reddy
Brian Crooker/Wanda Weber

**NIH award U24CA199347**

Tom Doake/Jeremy Fischer
*Indiana University*

**Jetstream**

**NIH**
*National Cancer Institute
Informatics Technology for Cancer Research*
Outline

• ITCR project: A Galaxy-Based Multi-Omic Informatics Hub for Cancer Researchers
  • A focus on mass spectrometry-based multi-omics

• Proteogenomics: A multi-omics approach for cancer research

• Proteogenomics: Informatics challenges and solutions
  • Galaxy-P
  • Enhancement with ITCR tools

• Access and next steps
Overview of ITCR project

Proteogenomics

Multi-omic informatics hub for cancer researchers

Metabolomics (Adrian Hegeman)

Metaproteomics

Pratik Jagtap, Project co-leader
Poster #39

Learn more at galaxyp.org
z.umn.edu/itcrgalaxyvideo
Proteogenomics: A primer

Peptide fractionation coupled to tandem mass spectrometry (MS/MS)
Matching amino acid sequences to MS/MS data

Raw MS/MS spectrum

Protein sequence and/or DNA sequence database search

Direct identification of 1000s proteins from complex mixtures

Peptide sequence match

Protein identification
Detecting protein variants via proteogenomics

Comprehensive Database (Sample-specific, all possible sequences)

UCGAUCAGGGCAAAU
RNA sequences (e.g. RNA-seq) (3-frame translation)

TCGATCAGGGCAAT
DNA sequences (6-frame translation)

In-silico translation

GalaxyP: galaxypr.org
Proteogenomic outcomes

- Confirms translation of variants
- Direct evidence of potential functional variants
- Applications in neoantigen discovery (immuno-oncology)
Bringing proteogenomics to the masses: informatics challenges

- Many software tools, integration, automation....

- RNA-Seq assembly and analysis
- Customized protein dB generation
- Matching sequences to MS/MS data
- Filtering and QC!
- Interpretation! Beyond a list....
A web-based, community developed bioinformatics workbench for integrating disparate software -- flexible

Geared towards use by bench scientists; many training resources available

Already home to genomic/transcriptomic tools

Provenance tracking, sharing and reproducibility

Amenable to other ‘omic tools (e.g. Galaxy for proteomics project, Galaxy-P)

**Working philosophy:**
Integrative data processing: RNA-Seq + proteomics
What’s next? Beyond a big list....
Multi-Omics Visualization Platform: Characterizing the nature of detected variants

- HTML-based Galaxy plugin
- Interactive reading of mzsqlite dB
Viewing and mapping variants of interest

Reference protein sequence

Identified peptide sequence

Sequence variant

Link to genomic coordinates
Mapping protein-level variants to the genome
Assessing potential impact of variants: CRAVAT-P

Cancer-Related Analysis of Variants Toolkit
(Rachel Karchin/Michael Ryan)
Assessing potential impact of protein-level variants: CRAVAT-P

- Intersecting of transcript variants and confirmed protein variants
Viewing results in CRAVAT-P

- HTML-based Galaxy plugin
- Interactive viewer

https://jraysajulga.github.io/cravatp-galaxy-docker/
Unleashing the power of CRAVAT on proteogenomic results

Quantitative proteotranscriptomics

Praveen Kumar

Accessing the tools

- **Cloud-based accessibility**
  - Jetstream:
    - z.umn.edu/proteogenomicsgateway
    - 350+ users
  - usegalaxy.eu

- **“Containerized”**
  - CRAVAT-P Galaxy Docker
    - A Docker image containing a fully-operational Galaxy instance with pre-installed demonstration material for CRAVAT-P.

- **Training**
  - Galaxy Training Network
Next steps

• Scalability

• Further customization and integration with other tools
  • Trinity
  • Knowledgebase tools for further analysis of protein-level variants, clinical significance
  • Single-cell genomics
  • Galaxy core extensions
  • NCI Cancer Research Data Commons

• CPTAC tools: PepQuery