Title: Integrative proteo-transcriptomics workflows within the Galaxy framework to explore the correlation between the expression of RNA and proteins

Authors: Praveen Kumar¹²*, Priyabrata Panigrahi³*, James Johnson⁴, Caleb Easterly², Subina Mehta², Andrew Rajczewski², Ray Sajulga², Mohammad Heydarian⁵, Krishanpal Anamika³, Timothy Griffin², Pratik Jagtap²

Affiliations: ¹Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, USA
²Biochemistry Molecular Biology and Biophysics, University of Minnesota, Minneapolis, USA
³LABS, Persistent Systems, Pingala-Aryabhata, Erandwane, Pune 411004, India
⁴Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA
⁵Department of Biology, Johns Hopkins University, Baltimore, USA

Email address of the presenting author: kumar207@umn.edu
URL for the overall project web site: galaxyp.org
URL for accessing the code: https://github.com/galaxyproteomics/tools-galaxyp/tree/master/tools

Main Text of Abstract: Technological advancement in the area of RNA/DNA sequencing techniques, protein/peptide identification techniques along with advances in bioinformatics have led researchers to explore multi-omics approaches. Several studies have used differential transcriptomic analysis to catalog gene expression in perturbed conditions (DOI:10.1593/neo.07112; DOI:10.1093/nar/gkv1282). However, several post-transcriptional regulatory mechanisms may lead to discordant mRNA and protein expression levels. Given that proteins are the cell's functional molecules, there has been a considerable interest in comparing protein expression with the cognate mRNA expression. In order to facilitate systems-biology analyses, we sought to develop accessible and user-friendly Galaxy tools and workflows.

For testing, we have used a published dataset from mouse developmental B-cell samples (DOI:10.4172/jpb.1000302). The dataset has both proteomic (with iTRAQ-labeling for quantitation) and transcriptomic (strand-specific single-end RNA sequencing) data. In order to correlate the transcript and protein abundances, we have developed a Galaxy workflow, using existing tools (Trimmomatic, HISAT2, StringTie, Cuffdiff, SearchGUI, PeptideShaker, ProteinPilot, etc.) and are developing new tools (moFF, REPORTER, etc.) that enable integration of these two omics data by calculating quantitative values from both transcriptomics and proteomics data.

We have used protein iTRAQ ratios from the ProteinPilot results and FPKM ratios from Cuffdiff to correlate the protein and mRNA expression. We observed 3590 genes common between transcriptome and proteome data, out of which 201 were significantly differentially expressed. These 201 genes showed a significant correlation between their mRNA and protein expression level (160 of the 201 genes [79.6%] showed similar expression pattern). We are exploring data mining methods to detect outliers (such as distance- and density-based anomaly detection methods) to identify candidate genes that show an altered protein expression as compared to mRNA expression. We identified eight influential genes (by using Cook’s distance method to identify outliers), which showed altered RNA expression as compared to protein expression. We will investigate these genes/proteins further for their biological significance. We are also exploring additional methods to convert RNA- and protein-level expression measurements into a uniform score (e.g. Z-score, T-score). This will enable us to compare both the molecular RNA and protein expression values on the same scale.

The workflow also enables generation of diagrams (static or interactive) which will help users in biological interpretation of the data. We will also explore the use of functional interaction network to understand mechanistic details involved in post-transcriptional regulation.

The Galaxy-P team has already developed workflows to a) generate proteomic databases from RNA-Seq data, b) search mass spectrometry data, and c) visualize the spectral quality and genomic localization of identified peptides. We believe that the development of the aforementioned RNA-protein correlation workflow will be a great addition to these workflows and will enable deeper insight in multiple biological studies.