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

Praveen Kumar, Priyabrata Panigrahi, James Johnson, Caleb Easterly, Subina Mehta, Andrew Rajczewski, Ray Sajula, Mohammad Heydarian, Krishnanpal Anamika, Timothy Griffin, Pratik Jagtap

Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, USA; Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, USA; 1Labs, Persistent Systems, Pune, India; 2Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA; 3Department of Biology, Johns Hopkins University, Baltimore, USA

Introduction
- 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-omic approaches.
- Several studies have used differential transcriptomic analysis to catalog gene expression in perturbed conditions, but there are other post-transcriptional regulatory mechanisms that 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 have developed accessible and user-friendly Galaxy tools and workflows.
- Using the mass-spectrometry-based proteomic data and RNASeq data, the workflow calculates the differential expression of proteins and transcripts respectively.
- The workflow also enables correlation study between expression of proteins and transcripts along with visualizations that will help users in interpretation of the data.

Discussion and Plans for Future Versions
- We generated a comprehensive workflow that generates an HTML output with the visualizations (Figure 3) using quantitative RNASeq data (FASTQ files) and proteomic data (MS/MS files) as input.
- The correlation tool explores the association of protein and transcript abundances at multiple levels. After correlating both the datasets, we can also look for outliers which can help in investigating the mechanistic cause behind the disparity in protein and mRNA expression.
- Clustering techniques can reveal the set of genes that show a similar pattern of mRNA versus protein expression, facilitating further functional富足 analyzation on genes that exhibit specific behaviors.
- Recently we have also developed an R-package for protein abundance which might further assist in finding proteins whose abundance or vice versa could reveal a class of genes that are being regulated differently.
- Currently, our tool enables performing correlation studies only on a single sample. We are working on including other biological samples which can enable users to perform a similar analysis on multiple samples with replicates, including multiple timepoints data.
- We are also exploring the compatibility of interactive visualization in the Galaxy which will provide a competent means to visualize and infer the disproportionate protein expression with transcript abundances.

Figure 1: Overview of the workflow

Figure 2: Factors influencing the correlation between transcriptome and proteome expression

Figure 3: Glimpse of outputs from the protein transcript correlation tool

Figure 4: The workflow and the tool implementation in Galaxy

Figure 5: The workflow and the tool implementation in Galaxy

Dataset Used for Testing
- For testing, we have used a published dataset from mouse developmental E10.5 samples (GEO:GEO:GSE103250). RNASeq data and protein data were derived from the pre-pro-B cells and pro-B cell development stages of B-cell.
- We have used protein iTRAQ labeled quantitation ratios from the ProteinPilot results and FPKM ratios from cuffdiff (after using hisat2, stringtie) to correlate the protein and mRNA expression.

Acknowledgements
This project is supported by National Science Foundation (NSF) grant 1602804 and National Institutes of Health (NIH) grant 1R01HG008917-01. J. Fischer and T. Doak from Indiana University for assistance on Jekyll Stream Galaxy instance.

Minnesota Supercomputing Institute