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

Praveen Kumar[1,2*], Priyabrata Panigrahi[3*], James Johnson[4], Caleb Easterly[2], Subina Mehta[2], Andrew Rajczewski[2], Ray Sajulga[2], Mohammad Heydarian[5], Krishanpal Anamika[3], Timothy Griffin[2], Pratik Jagtap[2]

[1]Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, USA; [2]Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, USA;
[3]LABS, Persistent Systems, Pingala-Aryabhata, Pune, India; [4]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, USA; [5]Department 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-omics approaches.
- Several studies have used differential transcriptomic analysis to catalog gene expression in perturbed conditions, but there are other 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 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.

Figure 1: Overview of the workflow

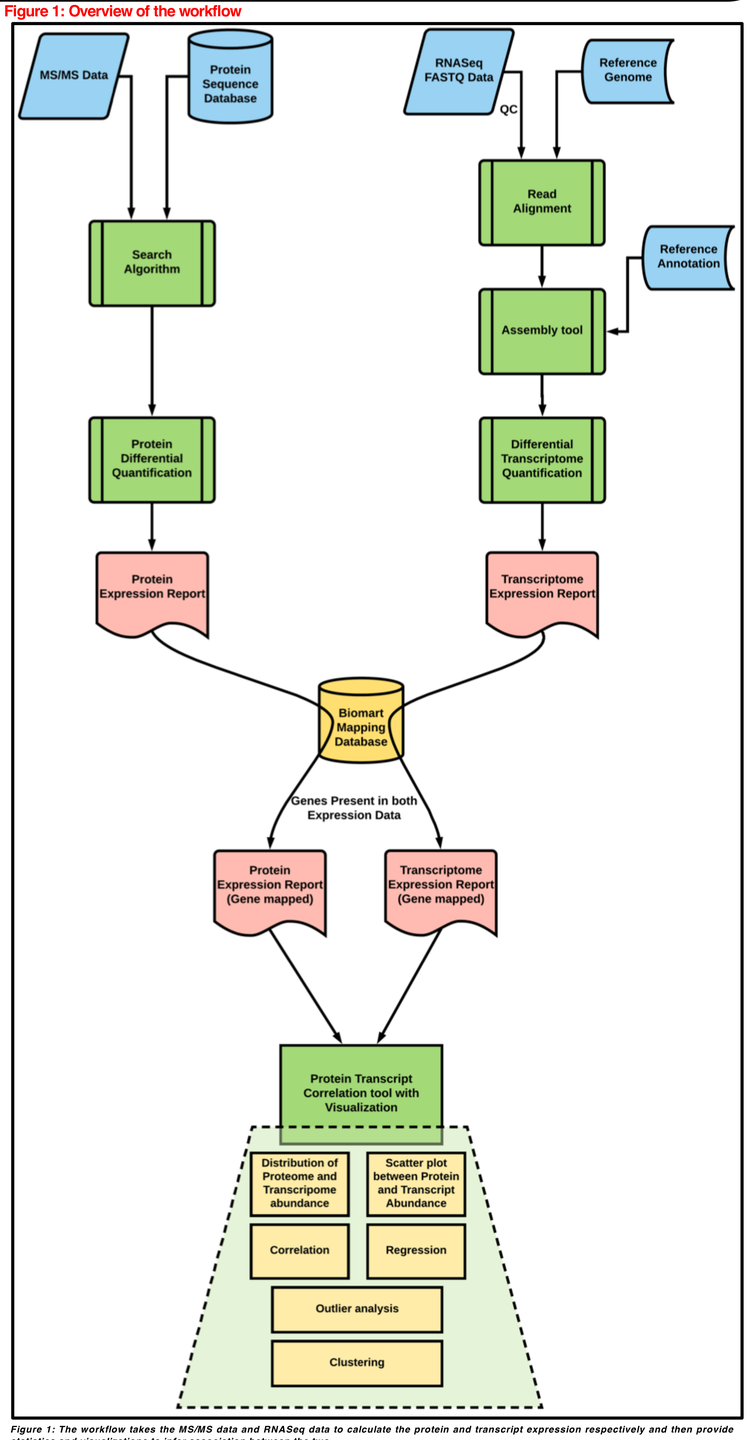


Figure 1: The workflow takes the MS/MS data and RNASeq data to calculate the protein and transcript expression respectively and then provide statistics and visualizations to infer association between the two.

Dataset Used for Testing

- For testing, we have used a published dataset from mouse developmental B-cell samples (DOI:10.4172/jpb.1000302). RNASeq data and protein data were derived from the pre-pro-B cells and pro-B cells (two developmental 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.

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

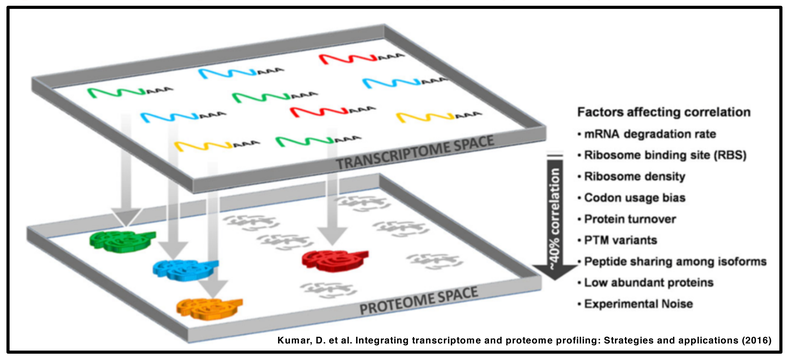
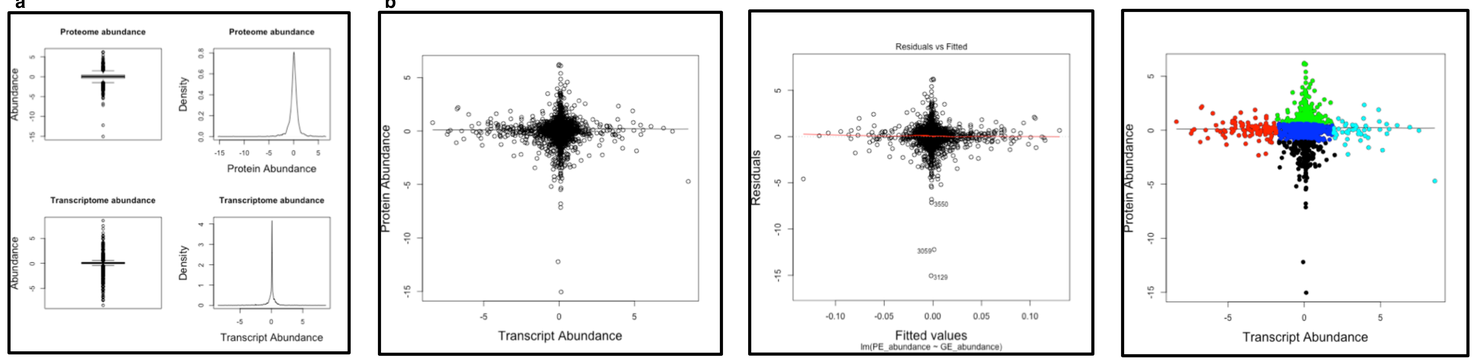


Figure 2: Even though proteins are translated from the mRNA, the correlation between transcriptome and proteome expressions are often low. Apart from some experimental limitations, there are other biological factors that can lead in observing discordant behavior between protein and transcript expression.

Discussion and Plans for Future Versions

- We generated a comprehensive workflow that generates a HTML output with the visualizations (Figure 3) using quantitative RNASeq data (FASTQ files) and proteomic data (MGF files) as inputs.
- The correlation tool explores the association of protein and transcript abundances at multiple levels. After correlating between both and exploring regression analysis, it also looks for outliers which can be vital 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. Performing a further functional enrichment analysis on genes that either show high mRNA abundance with low protein 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 options that 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 disproportion of protein expression with transcript abundance.

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



c

Parameter	Method 1	Method 2	Method 3
Correlation method used	Pearson's product-moment correlation	Spearman's rank correlation rho	Kendall's rank correlation tau
Correlation	-0.01553363	0.01800726	0.0123737
Pvalue	0.3932348	0.3223001	0.3255128

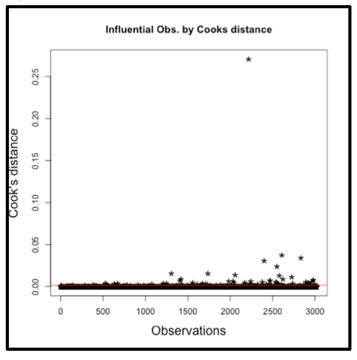
*Note that correlation is sensitive to outliers in the data. So it is important to analyze outliers/influential observations in the data. Below we use cook's distance based approach to identify such influential observations.

d

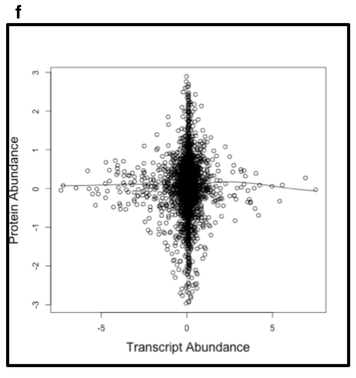
Top 10 influential observations (cook's distance > 4 * mean cook's distance)

PE_ID	PE_abundance	GE_ID	GE_abundance	cooksd
ENSMUSP00000103007.3	-3.043138	ENSMUST00000000783	0.09744768	0.00154561
ENSMUSP00000151536.1	2.542879	ENSMUST00000001256	-0.6392783	0.001496248
ENSMUSP00000089336.4	1.291007	ENSMUST00000002391	-2.493962	0.001881985
ENSMUSP00000101587.1	-3.583475	ENSMUST00000003156	-0.3868691	0.002451763
ENSMUSP00000030805.7	-0.951691	ENSMUST00000003469	3.68848	0.001951605
ENSMUSP00000011492.8	-0.3073879	ENSMUST00000004326	-7.421478	0.001720689
ENSMUSP00000029658.7	4.856793	ENSMUST00000004473	0.09744768	0.003943291
ENSMUSP00000099904.4	-3.268196	ENSMUST00000004673	0.9833685	0.003447334
ENSMUSP00000026274.7	-3.223033	ENSMUST00000004715	-0.08455913	0.00173285
ENSMUSP00000034539.5	0.5582226	ENSMUST00000005365	5.399083	0.002098078

e



f



g

Correlation with removal of outliers / influential observations

Parameter	Method 1	Method 2	Method 3
Correlation method used	Pearson's product-moment correlation	Spearman's rank correlation rho	Kendall's rank correlation tau
Correlation	0.008689192	0.02049365	0.01405192
Pvalue	0.6389949	0.2685151	0.2730888

Figure 3: The protein transcript correlation tool generates HTML output with visualizations and statistics. Some of the output figures are (a) Distribution of protein and transcript abundances (b) Scatter plot between protein and transcript abundances (c) Table showing correlation statistics (d) Top 10 outliers or influential points based on the cook's distance (e) Plot showing the cook's distance distribution of each observations (f) Scatter plot after removing the outliers/influential points (g) Correlation statistics after removing the influential points (h) Regression plot checking linear relationship (i) Clustering based on the K-means clustering algorithm.

Figure 4: The workflow and the tool implementation in Galaxy

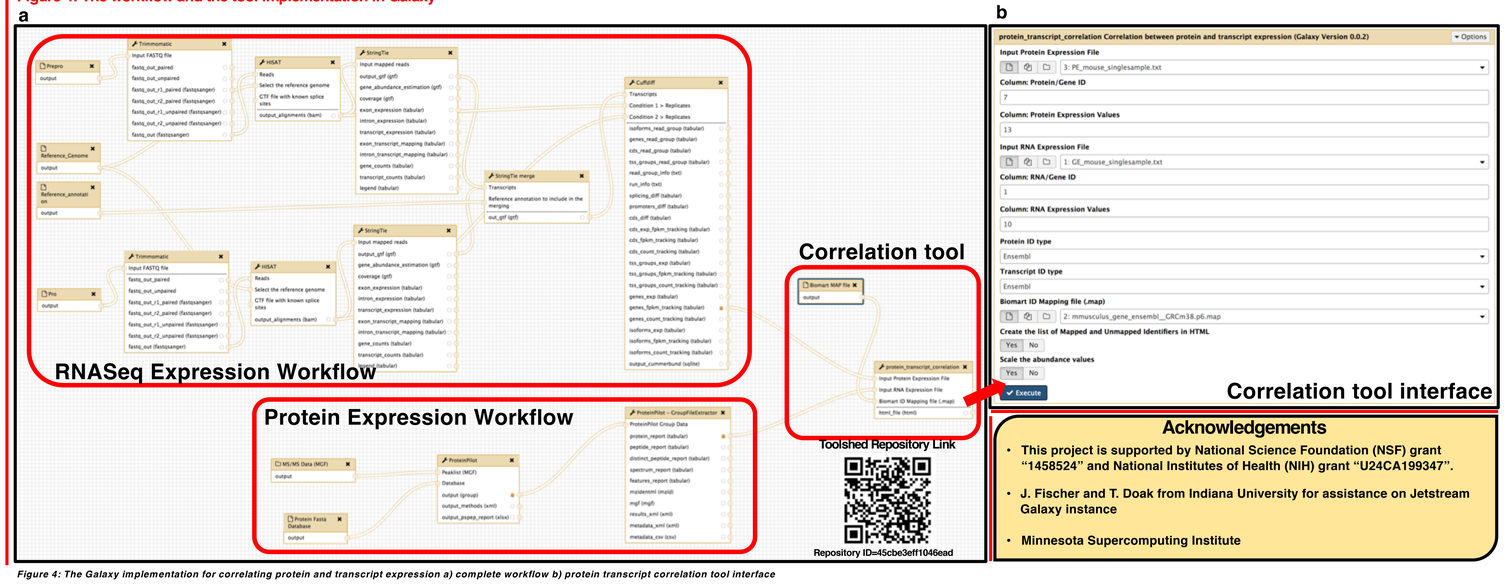


Figure 4: The Galaxy implementation for correlating protein and transcript expression a) complete workflow b) protein transcript correlation tool interface

Acknowledgements

- This project is supported by National Science Foundation (NSF) grant "1458524" and National Institutes of Health (NIH) grant "U24CA199347".
- J. Fischer and T. Doak from Indiana University for assistance on Jetstream Galaxy instance
- Minnesota Supercomputing Institute