INTRODUCTION

Genome sequence variants that affect protein coding sequences introduce a significant challenge in the analysis of proteomics with many of the sequence not present in the reference proteome databases. A solution to this issue is the development of customized proteomic databases informed from companion RNA-Seq transcriptome data. Examples of advantage of such approach include:

• Single Amino acid Polymorphisms (SAPs) and novel splice junctions detected from RNA-Seq data enable the identification of proteins not contained in the reference proteome database.
• Removing proteins from the reference database that are not expressed in the sample increases the number and quality of peptide identifications.

Creation of these RNA-Seq informed reference proteome databases can be complex. We used the Galaxy platform to develop user-friendly workflows for the creation of customized proteomic databases. These workflows can be shared and further customized, contributing to a process of transparency and reproducibility.

METHODS

Galaxy Workflows for RNA-Seq-based protein database construction were created from available Galaxy tools along with four new tools developed for this project.

The workflows were applied to a deep-coverage human Jurkat cell dataset—80 million paired-end RNA-Seq reads from an Illumina HiSeq2000 and ~500,000 mass dataset—80 million paired-end RNA deep coverage human Jurkat cell. The workflows were applied to a...

The work was done with the support of the Minnesota Supercomputing Institute, Minneapolis, MN; the University of Wisconsin, Madison, WI; and the University of Minnesota, Minneapolis, MN.

RESULTS

The reduced database was derived from Ensembl (v73, 104,310 entries). Transcripts were quantified using "RNA-Seq by Expectation-Maximization" (RSEM) and all protein entries for which a transcript was below 1 transcript per million (TPM) were eliminated, leaving 82,903 entries. Peptide identifications increased by 0.4x (original: 77,840, reduced: 78,168), increasing detection sensitivity.

The SAP database was derived from SNPs called against the human reference genome, resulting in 20,599 SAP-containing peptide entries. 533 SAP peptides that mapped up to 522 unique SNP sites on the genome were identified, a 79% overlap with results from Shemyakin et al. [JPR, 2013, 13(1)].

The splice database was made from novel splice junction sequences not present in the Ensembl gene models and consisted of 125,256 candidate novel junctions (215,989 total entries due to multiple translation frames). 67 novel splice-junction peptides were identified at a 1% local FDR, a 57% overlap with results from Shemyakin et al. (MCP, 2013, 12(8)).

CONCLUSION

Identification of peptides from Mass Spectrometry can be improved by generating custom search databases from RNA sequencing of the sample specimen. Galaxy provides a comprehensive and flexible framework in which to combine the analysis of both proteomic and genomic data. The analysis workflows created in galaxy can be easily shared to verify and replicate results, and can be easily reused and modified for future projects.

The Galaxy workflows and tools are available from:
• Galaxy Toolshed: http://toolshed.g2.bx.psu.edu/
• Galaxy V: public sever: http://usgalaxy.org/