Identifying novel peptide sequence variants from high throughput RNA-seq data via flexible proteomic database generation using the Galaxy framework

James Johnson[1]; Gloria Sheynkman[2]; Pratik Jagtap[3]; Michael Shortreed[2]; Getiria Onsongo[1]; Lloyd Smith[2]; Tim Griffin[3]

1. Minnesota Supercomputing Institute, Minneapolis, MN; 2. University of Wisconsin, Madison, WI; 3. University of Minnesota, Minneapolis, MN

INTRODUCTION

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

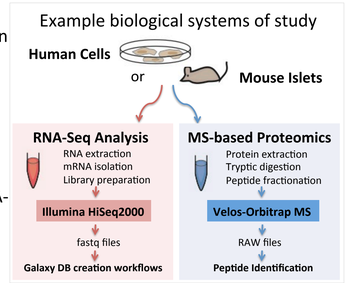
- Single Amino acid Polymorphisms (SAPs) and novel splice junctions detected from RNA-Seq data enable the identification of proteins not contained in the reference proteome database.
- Removing proteins from the reference database that are not expressed in the sample increases the number and quality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



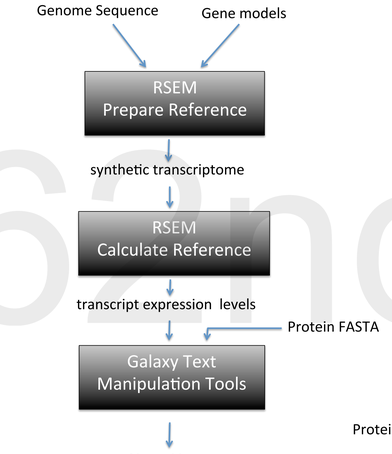
Reduced, SAP, and splice protein databases, all based on comparison to Ensembl gene models (GTF, v73), were constructed from the RNA-Seq data. These customized databases and Ensembl reference protein database were each searched against the MS data using SEQUEST/Percolator to identify novel peptide variants (1% FDR). The peptides were compared to previous published studies that used different gene models (RefSeq instead of Ensembl).

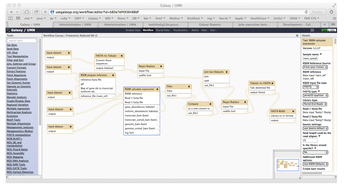
We further demonstrated the utility of the database workflows by applying them for analysis of islet cell samples from mouse strains: B6 and CAST.

Galaxy Database Creation Workflows

Reduced Database

RSEM determines the RNA-Seq transcripts expressed at detectable levels. Proteins from transcripts that are not expressed are filtered out.



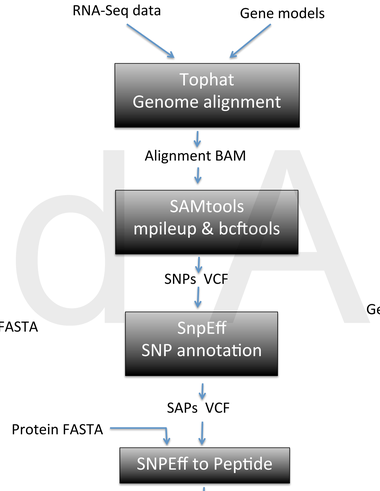


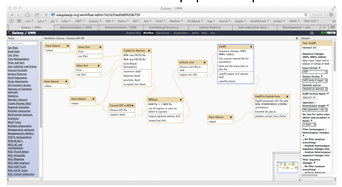
http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_reduced_db_workflow

http://z.umn.edu/reduceddb

SAP Database

RNA-Seq reads are aligned to the reference genome with tophat. SAMtools identifies variant DNA bases. SNPEff annotates the variants with effects to genes and proteins.



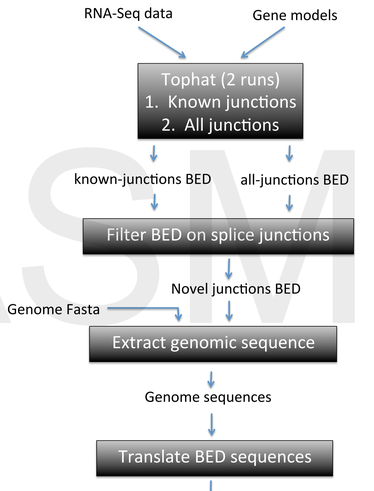


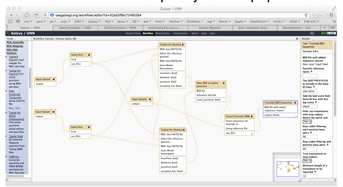
http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_sap_db_workflow

http://z.umn.edu/sapdb

Splice Database

Tophat alignments are used to find evidence of novel splice variant transcripts. The novel splice junctions are translated into a protein database.





http://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_rnaseq_splice_db_workflow

http://z.umn.edu/slicejndb

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 0.4% (original:77,840, reduced:78,168), increasing detection sensitivity.

			reduced database workflow results				
			original DB		reduced DB		
Sample	Rna-Seq reads	Mass spectra	# entries	peptide IDs	# entries	peptide IDs	% increase
Jurkat human cells	80M	500K	104,310	73,123	82,101	73,436	0.4
B6 mouse islets	94M	250K	52,165	30,137	18,052	30,220	0.3
CAST mouse islets	126M	250K	52,165	28,756	16,940	28,823	0.2

peptide passing a 1% FDR were identified

The SAP database was derived from SNPs called against the human reference genome, resulting in 20,595 SAP-containing peptide entries. 553 SAP peptides that mapped up to 522 unique SNP sites on the genome were identified, a 79% overlap with results from Sheynkman et al. (JPR,2013,13(1)).

	SAP database workflow results			
	sap DB			
Sample	# SAPs	# SNP sites	peptide IDs	SNPs ID'd
Jurkat human cells	9,168	6,924	522	491
B6 mouse islets	1	1	N/A	N/A
CAST mouse islets	476	249	22	19

peptide passing a 1% FDR were identified

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 (219,989 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 Sheynkman et al. (MCP, 2013,12(8)).

	splice database workflow results		
	splice database		
Sample	size	min. depth	peptide IDs
Jurkat human cells	33,372	6	67
B6 mouse islets	57,587	4	64
CAST mouse islets	43,244	4	66

peptide passing a 1% local FDR were identified

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/
- GalaxyP public sever: http://usegalaxyp.org/

Acknowledgements:

- The Galaxy Project: http://galaxyproject.org/
- Anne Lamblin for initiating the Galaxy Projects at the Minnesota Supercomputing Institute
- John Chilton for Galaxy-P development that adapted Galaxy for proteomics
- Donnie Stapleton, Mark Keller, and Alan Attie for supplying the mouse islet samples

- National Institute of Diabetes and Digestive Kidney Diseases grants 58037 and 66369
- NIH Genomic Sciences Training Program ST32HG002760
- Galaxy-P is supported through the National Science Foundation; DBI-ABI award 1147079

"Galaxy-P: A new community-based informatics paradigm for MS-based proteomics"