An automated, accessible proteogenomic pipeline for high confidence detection and rigorous validation of novel peptide sequence variants in Galaxy-P

Andrew T. Rajczewski[1]; Bo Wen[2]; James E. Johnson[1]; Ray Sajulga[1]; Subina Mehta[1]; Qiyuan Han[1]; Praveen Kumar[1]; Pratik D. Jagtap[1]; Bing Zhang[2]; Timothy J. Griffin[1]; Natalia Y. Tretyakova[1]

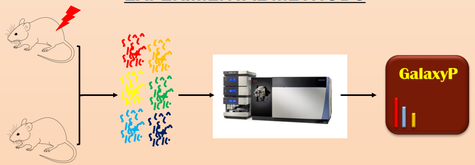
[1] *Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN*

[2] *Baylor College of Medicine, Houston, TX*

INTRODUCTION

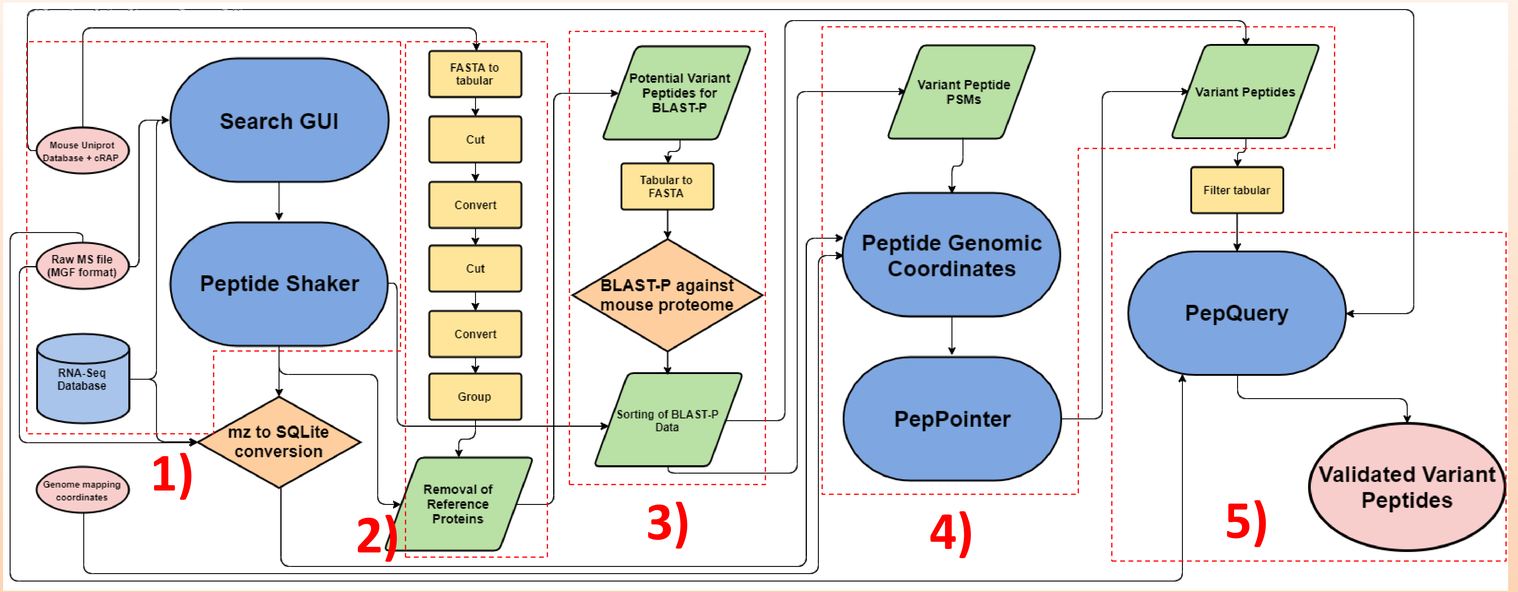
- When bottom-up proteomics experiments utilize FASTA files derived from genomic data, information on proteoforms unique to the samples under analysis is missed (indels, intron retention, alternative splicing, etc.)
- Proteogenomics workflows utilize custom-built FASTA libraries based on transcriptomic data to inform attempts to detect these protein variants
- Many proteogenomics workflows neglect to validate the variants they detect, introducing the potential for false positives
- The PepQuery search engine validates variant peptides by searching MS/MS data against specific variant sequences of interest
- We brought PepQuery into the Galaxy for Proteomics (Galaxy-P) suite and integrated it into an automated workflow to validate newly discovered variant peptides from proteogenomics data
- This workflow was tested using previously created proteogenomics data, and validated variant peptides were verified using targeted mass spectrometry

EXPERIMENTAL METHODS

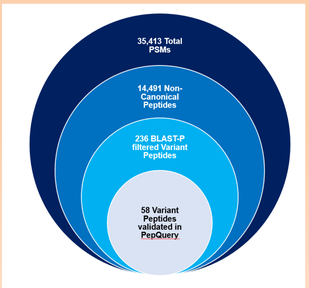


- Samples of control, inflamed (n=3 each) mice were obtained as a part of a study of inflammatory bowel disease
- Proteins were extracted from proximal colon samples and prepped for MS analysis with digestion, TMT-6plex labeling, and concatenation
- Concatenated, labeled peptides were fractionated using high pH reverse-phase fractionation before LC-MS analysis
- Raw MS runs were searched against a FASTA database constructed from RNA-Seq data from comparable samples to determine total peptide spectral matches (PSMs) (Section 1 in workflow)
- Total PSMs are filtered to remove PSMs corresponding to conventional mouse peptides as well as common contaminant peptides (Section 2)
- BLAST-P searching the remaining PSMs against the mouse proteome was used to remove "variant" peptides which have close matches to convention peptides (Section 3)
- Genomic coordinates were assigned to variant peptides (Section 4)
- PepQuery used to validate the identified variant peptides, with the results filtered to remove variants with *any* other potential matches (Section 5)
- Parallel reaction monitoring (PRM) experiments performed to independently validate variant peptides

VARIANT PEPTIDE DETECTION AND VALIDATION WORKFLOW

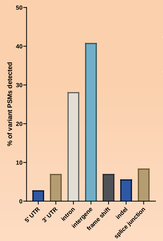


VARIANT PEPTIDES ARE IDENTIFIED AND SUCCESSFULLY VALIDATED VIA OUR WORKFLOW



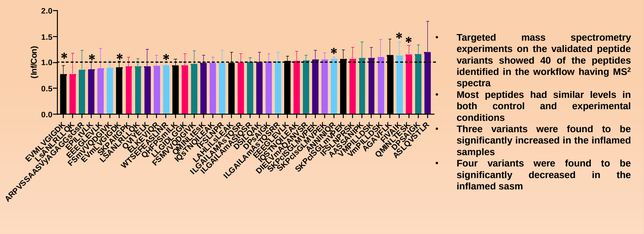
- Initial searches of the eight mouse peptide fractions against an RNA-Seq-supplemented FASTA database revealed 35,413 PSMs
- Comparison against canonical proteins and common MS contaminants leaves half of those PSMs as potential non-canonical peptide variants
- BLAST-P analysis narrows the variant peptides further
- PepQuery analysis shows that 58 variant peptide spectra have no better match in the data, and can be considered valid

Variant Types Detected in MS/MS data



- The variant peptides validated in the workflow correspond to several distinct types of variation
- Most validated variant peptides correspond to peptides from intergenic regions not known to translate into proteins (40% of validated peptides) as well as peptides stemming from retained introns in the proteins (28% of validated peptides)

Ratio of Variant Peptide Abundances in Inflamed and Control Samples



- Targeted mass spectrometry experiments on the validated peptide variants showed 40 of the peptides identified in the workflow having MS² spectra
- Most peptides had similar levels in both control and experimental conditions
- Three variants were found to be significantly increased in the inflamed samples
- Four variants were found to be significantly decreased in the inflamed sasm

SUMMARY

- The PepQuery search engine was brought into the Galaxy-P suite and incorporated into a workflow designed to identify variant peptides in proteogenomic data
- With this new workflow, we were able to identify and validate 58 variant peptides in proximal colon proteogenomics data from an earlier study into inflammatory bowel disease
- Most of the validated 58 variant peptides corresponded to intergenic regions as well as retained introns
- Subsequent validation via targeted mass spectrometry experiments showed direct evidence of 40 of the 58 variant peptides

FUTURE DIRECTIONS

- This workflow will be further tested on open-source proteogenomics datasets to ascertain its ability to detect and validate variant peptides
- An add-on to this workflow is currently in development to automatically generate an inclusion list for the method development of targeted mass spectrometry assays
- A version of this workflow is in development for the label-free quantitation of validated variant peptides in mass spectrometry data

REFERENCES

1. Alfaro, J. A., Sinha, A., Kislinger, T., & Boutros, P. C. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. Nature methods, 11(11), 1107.
2. Johansson, H.J., Socciarelli, F., Vacanti, N.M. et al. Breast cancer quantitative proteome and proteogenomic landscape. Nat Commun 10, 1600 (2019).
3. Payne, S. H. (2015). The utility of protein and mRNA correlation. Trends in biochemical sciences, 40(1), 1-3.
4. Wen, B., Wang, X., & Zhang, B. (2019). PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. Genome research. https://doi:10.1101/gr.235028.118

ACKNOWLEDGEMENTS

This research was supported in part by the National Cancer Institute. Andrew Rajczewski was supported by an NIH biotechnology training grant T32GM008347 from the NIH National Institute of General Medical Sciences. The Galaxy-P platform was supported through NCI grant U24CA199347. PepQuery was created by Bo Wen of the Zhang lab at the Baylor College of Medicine. RNASeq sample preparation was performed by Qiyuan Han. Generation of FASTA database was performed using the Galaxy-P platform. Samples were obtained from the Tannenbaum lab at the Massachusetts Institute of Technology.

National Institutes of Health
National Institute of General Medical Sciences
BASIC DISCOVERIES FOR BETTER HEALTH

University of Minnesota
Driven to Discover℠