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

Andrew T. Rajczewski1; Bo Wen2; James E. Johnson1; Ray Rajput1; Subina Mehta1; Qiyan Han1; Praveen Kumar1; Pratik D. Jagtap1; Bing Zhang2; Timothy J. Griffin1; Natalya Y. Treyakov1

1 Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN
2 College of Medicine, Houston, TX

INTRODUCTION

- When bottom-up proteomics experiments utilize FASTA files derived from genomic data, information on post-translational changes 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-P 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 (3x3 each) mice were obtained as part of a study of inflammatory bowel disease
- Proteins were extracted from proximal colon samples and prepped for MS analysis with digestion, TFE/Glycidel labeling, and concentration
- Consensus-based, labeled peptides were fractionated using high pH strong cation exchange chromatography
- Raw MS runs were searched against a FASTA database constructed from RNA-seq data from compared normal control to determine total peptide spectra matches (PSMs) (Section 1 in workflow)
- Total PSMs are filtered to remove PSMs corresponding to conventional mouse protein sequences as well as common contaminant peptides (Section 2)
- BlastP against the remaining PSMs against the mouse proteome was used to remove “variant” peptides which have close matches to conventional mouse sequences (Section 3)
- Genomic coordinates were assigned to validated peptides (Section 4)
- PepQuery was 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 PEPTIDES ARE IDENTIFIED AND SUCCESSFULLY VALIDATED VIA OUR WORKFLOW

SUMMARY

- The PepQuery search engine was brought into the Galaxy-P suite and incorporated into a workflow designed to validate variant peptides in proteogenomic data
- With this novel workflow, we were able to identify and validate 58 variant peptides in proximal colon proteogenomic 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 of validation of targeted mass spectrometry assays
- A version of this workflow is in development for the label-free quantification of validated variant peptides in mass spectrometry data

ACKNOWLEDGEMENTS

This research was supported in part by the National Cancer Institute. Andrew T. Rajczewski was supported by an NIH intramural training grant (T32 GM008804). This work was supported by grants from the NH/NIH National Institute of General Medical Sciences. The Galaxy-P software development was supported by National Institutes of Health (NIH) grant no. 5G12GM087330-03. The research work was performed by Qiyan Han. Generation of FASTA database was performed using PepShaker platform. Samples were obtained from the Tannenbaum lab at the Massachusetts Institute of Technology.

REFERENCES


SUMMARY

- The PepQuery search engine was brought into the Galaxy-P suite and incorporated into a workflow designed to validate variant peptides in proteogenomic data
- With this novel workflow, we were able to identify and validate 58 variant peptides in proximal colon proteogenomic 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 of validation of targeted mass spectrometry assays
- A version of this workflow is in development for the label-free quantification of validated variant peptides in mass spectrometry data

ACKNOWLEDGEMENTS

This research was supported in part by the National Cancer Institute. Andrew T. Rajczewski was supported by an NIH intramural training grant (T32 GM008804). This work was supported by grants from the NH/NIH National Institute of General Medical Sciences. The Galaxy-P software development was supported by National Institutes of Health (NIH) grant no. 5G12GM087330-03. The research work was performed by Qiyan Han. Generation of FASTA database was performed using PepShaker platform. Samples were obtained from the Tannenbaum lab at the Massachusetts Institute of Technology.

REFERENCES


SUMMARY

- The PepQuery search engine was brought into the Galaxy-P suite and incorporated into a workflow designed to validate variant peptides in proteogenomic data
- With this novel workflow, we were able to identify and validate 58 variant peptides in proximal colon proteogenomic 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 of validation of targeted mass spectrometry assays
- A version of this workflow is in development for the label-free quantification of validated variant peptides in mass spectrometry data

ACKNOWLEDGEMENTS

This research was supported in part by the National Cancer Institute. Andrew T. Rajczewski was supported by an NIH intramural training grant (T32 GM008804). This work was supported by grants from the NH/NIH National Institute of General Medical Sciences. The Galaxy-P software development was supported by National Institutes of Health (NIH) grant no. 5G12GM087330-03. The research work was performed by Qiyan Han. Generation of FASTA database was performed using PepShaker platform. Samples were obtained from the Tannenbaum lab at the Massachusetts Institute of Technology.

REFERENCES